CONTENTS

1. **OVERVIEW**
   1.1 Guided wave optical devices
   1.2 Rationale

2. **ELECTROMAGNETIC FIELDS AND PLANE WAVES**
   2.1 Maxwell's equations
   2.2 The differential form of Maxwell's Equations
   2.3 Harmonically-varying fields and the wave equation
   2.4 Plane waves
   2.5 Power flow
   2.6 The propagation of general time-varying signals

3. **MATERIAL EFFECTS**
   3.1 Introduction
   3.2 The origin of the dielectric constant
   3.3 Material dispersion
   3.4 The origin of electrical conductivity
   3.5 Anisotropic media
   3.6 Non-linear effects
   3.7 The electro-optic effect
   3.8 Stress-related effects

4. **THE OPTICS OF BEAMS**
   4.1 Scalar theory
   4.2 Spherical waves
   4.3 Lenses
   4.4 Diffraction theory
   4.5 Fraunhofer diffraction
   4.6 Fresnel diffraction and Gaussian beams
   4.7 Transformation of Gaussian beams by lenses
   4.8 Mirrors and resonators

5. **REFLECTION AND REFRACTION AT A SINGLE INTERFACE**
   5.1 The behaviour of light at a dielectric interface
   5.2 Superposition of fields
   5.3 Boundary matching
   5.4 Electromagnetic treatment of the interface problem
   5.5 Modal treatment of the dielectric interface problem
   5.6 Surface plasma waves

6. **THE SLAB WAVEGUIDE**
   6.1 Guided waves in a metal guide
   6.2 Guided waves in a slab dielectric waveguide
   6.3 Other types of mode
   6.4 The weak-guidance approximation
   6.5 The orthogonality of modes
   6.6 The power carried by a mode
6.7 The expansion of arbitrary fields in terms of modes
6.8 Application of the overlap principle

7. **PLANAR WAVEGUIDE INTEGRATED OPTICS**
   7.1 Overview of planar waveguide components
   7.2 Phase matching at a single interface
   7.3 The FTIR beamsplitter
   7.4 The prism coupler
   7.5 Phase matching for guided modes
   7.6 Refractive optical components
   7.7 Gratings
   7.8 Gratings in guided wave optics

8. **OPTICAL FIBRES AND FIBRE DEVICES**
   8.1 Optical fibre types
   8.2 Loss in silica and fluoride glass fibre
   8.3 Step-index optical fibres
   8.4 Parabolic-index optical fibres
   8.5 Signal dispersion
   8.6 Mode conversion in fibres
   8.7 Coupling to fibres
   8.8 Fibre interconnects
   8.9 Fibre-based components

9. **CHANNEL WAVEGUIDE INTEGRATED OPTICS**
   9.1 Channel waveguide types
   9.2 Input and output coupling
   9.3 Sources of propagation loss
   9.4 Polarizers
   9.5 Mirrors
   9.6 Tapers and Y-junctions
   9.7 Phase modulators
   9.8 Frequency shifting and high speed operation
   9.9 Interferometers

10. **COUPLED MODE DEVICES**
    10.1 Introduction
    10.2 The directional coupler - basic principles
    10.3 The directional coupler - theoretical analysis
    10.4 Solutions of the equations at synchronism
    10.5 Asynchronous solution
    10.6 Fibre directional couplers
    10.7 Coupling between dissimilar waveguides
    10.8 Sidelobe suppression using tapered coupling
    10.9 The reflection grating filter - basic principles
    10.10 The reflection grating filter - theoretical analysis
    10.11 Solution of the equations at synchronism
    10.12 Asynchronous solution
    10.13 Fibre gratings
CONTENTS

10.14 Other coupled mode interactions

11. OPTOELECTRONIC INTERACTIONS IN SEMICONDUCTORS
   11.1 Wave-particle duality
   11.2 Photons
   11.3 Electrons
   11.4 Band theory
   11.5 Effective mass
   11.6 Carrier statistics
   11.7 Intrinsic and extrinsic semiconductors
   11.8 Detailed balance
   11.9 Rate equations

12. OPTOELECTRONIC DEVICES
   12.1 The p-n junction diode
   12.2 Electro-optic semiconductor devices
   12.3 Photodiodes
   12.4 The light emitting diode
   12.5 The semiconductor laser: basic operation
   12.6 The semiconductor laser: steady state analysis
   12.7 The semiconductor laser: modulation
   12.8 DBR and DFB lasers
   12.9 Array lasers
   12.10 Surface-emitting lasers
   12.11 Travelling-wave laser amplifiers

13. OPTICAL DEVICE FABRICATION
   13.1 Overview
   13.2 Planar processing
   13.3 Substrate growth and preparation
   13.4 Deposition and growth of material
   13.5 Material modification
   13.6 Etching
   13.7 Lithography
   13.8 Optical fibre fabrication

14. SYSTEMS AND APPLICATIONS
   14.1 Introduction
   14.2 The planar integrated optic RF spectrum analyser
   14.3 The planar integrated optic disc read head
   14.4 Guided wave optical chip-to-chip interconnects
   14.5 The channel waveguide integrated optic A-to-D converter
   14.6 Optical fibre sensors
   14.7 The integrated optic fibre gyroscope
   14.8 High speed, guided wave optical communications
   14.9 Fibre lasers and amplifiers

ANSWERS TO SELECTED PROBLEMS
OVERVIEW

1.1 GUIDED WAVE OPTICAL DEVICES

There is now little doubt that guided wave optical devices will have an increasing impact on electrical engineering in the coming years. Primarily, this is due to the sustained development of two extremely successful components: optical fibre, and the semiconductor laser. Together, these are currently in the process of revolutionising the telecommunications industry, and making considerable advances in a variety of other new and exciting application areas.

Optical fibre acts as a propagation medium for guided optical waves. The most important fibre type is single-mode optical fibre, which can support guided waves for extremely long distances (tens or hundreds of kilometres) with low loss and low signal dispersion. Single-mode fibre is generally formed from silica glasses, which are arranged in a cylindrical geometry with a core region of high refractive index surrounded by a lower-index cladding layer (as shown in Figure 1.1-1).

![An optical fibre.](image1.png)

Figure 1.1-1  An optical fibre.

The overall diameter of a single-mode fibre is comparable to that of a human hair. However, the light actually travels almost entirely in the much smaller core region (which typically has a diameter of only a few microns) and is contained inside this region and guided along the fibre by a phenomenon known as total internal reflection. This can occur when light is incident on an interface between two lossless dielectric media with different refractive indices, provided the light is incident from the high-index medium at a shallow enough angle, as shown in Figure 1.1-2. In this case, there is no transmission of energy across the interface, which simply acts as a highly efficient reflector.

![Total internal reflection.](image2.png)

Figure 1.1-2  Total internal reflection.

In the last two decades, the optical fibre has finally allowed the practical use of light as a carrier of information, and this application has given rise to the entirely new field of
photonic engineering. Since an optical field oscillates at extremely high frequency ($\approx 10^{15}$ Hz), a light wave acts as a high-frequency carrier. In combination with the low dispersion provided by an optical fibre, this allows information to be transferred at very high data rates. In fact, the optical fibre offers an increase in channel capacity of a factor of $10^4 - 10^5$ over its nearest competitor, the microwave guide, coupled with a considerable reduction in size, weight and cost. Because of the low cost of the raw materials used (effectively, purified sand), optical fibre will replace copper cables in many applications involving the transfer of low-power signals.

A further type of optical fibre, known as multi-mode fibre, also exists. This has a similar cylindrical geometry, but a much larger core diameter (typically 50 - 100 $\mu$m). It can be formed from either glass or plastic. However, as the name implies, multimode fibre can support more than one characteristic propagation mode, and consequently suffers from much greater dispersion than single-mode fibre. It is also extremely lossy, and is therefore suitable only for medium-bandwidth, short-haul applications.

The semiconductor laser is a small, efficient light source. It is the natural successor of the light-emitting diode (or LED), a low-coherence, low-power source which may only be modulated at slow rates (up to tens of megahertz). However, because of its relative cheapness and simplicity, the LED has been widely used in the past in multimode optical fibre systems. Figure 1.1-3 shows a commercial near-infrared buried heterostructure semiconductor laser contained in a dual in-line package. The laser itself is no bigger than a grain of sand, and is primarily formed from layers of the crystalline material indium gallium arsenide phosphide, grown in differing material compositions by sophisticated epitaxial techniques. It is designed to emit light at around 1.3 $\mu$m wavelength.

The technology required to fabricate a semiconductor laser is now well developed. Its optical output can be very intense - beam powers of tens of milliwatts are now routinely available - and it can be directly modulated at very high speeds, typically in the gigahertz range. Since the semiconductor laser is also a guided wave device, and emits light from a small stripe window, which has dimensions comparable to those of the core of an optical fibre, it is immediately compatible with fibre systems. Indeed, the module in Figure 1.1-3 is already fitted with a single-mode fibre pigtail. This is butted up to the emitting facet of the
laser, deep inside the package, allowing very simple connection to a fibre system. Furthermore, the semiconductor laser can emit light of the wavelengths at which silica fibre shows minimum dispersion and minimum propagation loss, near $\lambda_0 = 1.3$ and $1.5 \, \mu m$ respectively. With careful design, the output can also consist of very nearly a single optical frequency. This is an important advantage for communications, since it reduces the effect of dispersion.

Tuneable lasers for use in wavelength-division multiplexed communications systems, surface-emitting lasers for optical processor applications and high-power laser arrays are also the subject of an intense development effort. The latter can generate quite staggering amounts of power - many watts, CW - from a tiny chip, and may well pose a challenge to large gas lasers in the future. Recent advances include the use of semiconductor lasers as in-line amplifiers and repeaters in optical fibre systems, and the development of semiconductor laser-pumped optical fibre lasers.

Together with photodiodes, which are light detecting devices fabricated in semiconductor materials, optical fibre and semiconductor lasers form the key elements of optoelectronics. In addition, a large number of other components have been developed, which allow a complete guided wave circuit to be assembled. These ancillary components are in a more rudimentary form at present, and it is hard to predict which of them will be the most successful in the long run. As with most things, the deciding factor will almost certainly be cost rather than absolute performance.

The subsidiary components have evolved in several different formats. For example, a number of passive devices are available in all-fibre form, and may readily be spliced into fibre systems. The most widely-used device is the fibre coupler, which can be used for beam-splitting and filtering operations. For example, Figure 1.1-4 shows a packaged fused tapered fibre coupler, which is available in a variety of fixed splitting ratios ranging from 50 : 50 to 1 : 99. More complicated power division functions - for example, a 1 x N split - are possible using similar technology. Similarly, all-fibre polarizers and polarization controllers provide ways of manipulating the polarization state of guided beams. However, although the devices are fabricated using relatively simple equipment, they must generally made individually, so the construction of a complicated optical circuit is time-consuming. Furthermore, since optical fibre is based on amorphous material, it has been found to be extremely difficult to fabricate more sophisticated active devices (such as high-speed modulators) in this way.

![Figure 1.1-4 A packaged fibre coupler.](image)

Another common component format, which offers a solution to some of these problems, is provided by integrated optics. As the name implies, this allows a number of guided wave devices to be combined on a common substrate. In contrast to optical fibre devices, integrated optic components are made using a mass-production technology akin to that required for very large scale integrated (VLSI) microelectronics. Complex optical circuits may therefore be constructed as easily as simple ones, the main limitation being the size of available substrates. There are two basic configurations. Historically, planar waveguide
integrated optics was the first system to be developed, being followed shortly by the more sophisticated and successful channel waveguide integrated optics system.

Planar waveguide integrated optics is concerned with the manipulation of sheet beams. These can propagate in any direction parallel to the surface of a high-index guiding layer, which provides optical confinement in a single direction. Figure 1.1-5 shows a three-layer planar guide, formed by depositing a thin layer of material of high refractive index on a thicker, low-index substrate. The third layer in the system can often be air itself, or an additional low-index cover layer can be used. Once again, guidance is provided by total internal reflection at the layer interfaces.

Because sheet beams allow many of the operations that are possible using free-space optics (for example, focusing by a lens, or beam deflection), planar integrated optic chips often contain circuits that are in effect miniaturised and ruggedised versions of bulk optic systems. Many of these are used for parallel signal processing operations, based on the Fourier transform properties of a lens.

In the second configuration, channel waveguide integrated optics, the beams propagate along high-index guiding channels. These may be formed as ridges on the surface, or as buried channels as shown in Figure 1.1-6.

The beams are therefore confined in two dimensions, and so can only follow predefined pathways round the chip. If the channel dimensions are chosen to correspond with those of an optical fibre core, this form of integrated optics is directly compatible with fibre optics, and fibre and channel guide components may be connected together quite simply. Integrated optic components can of course be formed on passive substrates such as glass or plastic. At the very least, these allow the construction of beamsplitting and combining devices. For example, Figure 1.1-7 shows a number of different 1 x 2 and 1 x 4 passive splitters formed by silver-sodium ion exchange in glass. The integrated optic components themselves are contained inside the packages, and are again fitted with fibre pigtails.
Figure 1.1-7 Passive channel waveguide integrated optic beamsplitting components formed by ion exchange in glass (photograph courtesy D.DeRose, NSG America Inc.).

If other materials are used, more exotic functions are possible. For example, using electro-optic substrates like LiNbO$_3$, GaAs and InP, modulation and switching can be performed, often at extremely high speeds (tens of gigahertz). For example, Figure 1.1-8 shows an electro-optic directional coupler, a generic device type which can be used to switch light between two adjacent parallel waveguides under electrical control, or modulate an information channel.

These operations can be used to place information on the carrier wave, and then route it around a network. Similarly, filtering operations allow the channel to be multiplexed and demultiplexed by wavelength-division. These functions can even be combined with the generation and detection of light, if the waveguide is fabricated in a semiconductor substrate with a suitable bandgap (such as GaAs or InP). Integration also offers the intriguing possibility of combining optical components with their controlling electronics, in the form of integrated optoelectronics, or of using light as a method of communicating between very high speed electronic circuits in fast computers.

Using integrated optics, complete optical circuits can be constructed on a common substrate, often with very small dimensions. One example might be a switch matrix, capable of routing light between N input and N output fibres in a communications system, thus forming an
compact electrically-controllable optical 'telephone exchange'. Figure 1.1-9 shows a packaged 4 x 4 lithium niobate switch array, constructed from a set of directional coupler switches. Here, each directional coupler acts as a switching node, and suitable configuration of the array can allow a pathway to be set up between any input fibre and any output.

![Figure 1.1-9 A packaged, pigtailed 4 x 4 switch array in titanium-diffused LiNbO₃ (photograph courtesy Dr P. Granestrand, Ericsson Telecom A B).](image)

A second application might involve the fabrication of an optical interferometer, using a coil of optical fibre (which provides the sensor element) and a single channel waveguide integrated optic chip (which carries a number of beamsplitting and signal processing components). Such devices can be arranged to sense variations in a wide range of physical parameters. For example, a Sagnac interferometer based on an optical fibre coil can act as a rotation sensor or gyroscope. In contrast to its mechanical equivalent, the optical fibre gyroscope requires no precision mechanical parts, little maintenance, and no 'run-up' time. Combined with linear accelerometers, such devices may well one day form the basis of solid-state inertial navigation systems, sensing rotational motion about three orthogonal axes.

### 1.2 RATIONALE

This book is intended to provide an introduction to the whole field of guided wave devices, which has proved to be one of the major technological growth areas of this decade. It is not a research monograph, and we make no claim that the topics covered are a description of the latest advances; in fact, the whole field is developing so rapidly that it is hard to identify the leading edge of technology in many cases. Instead, our aim is to describe a large range of devices and applications in a reasonably simple, self-contained and unified way, using language likely to be understood by third-year science undergraduates and M.Sc. students.

Most of the material has been taken from option courses given in the Department of Electrical Engineering at Imperial College London over the last decade. The contents of Chapter 2 and Chapters 5 - 9 (which cover electromagnetic theory, interface problems, waveguides, planar and channel guide integrated optics, and fibre optics) are currently being
used as a 20-lecture introductory course in guided-wave optical devices for electrical engineers. Chapter 3 adds further background on optical materials, and Chapter 4 illustrates the application of electromagnetic theory to the optics of beams and beam-forming components. None of this material requires a knowledge of solid-state theory. This is reviewed in Chapter 11, and applied to optoelectronic devices (mainly light-emitting devices and detectors) in Chapter 12. The final chapters (13 and 14) cover device fabrication, and the use of guided-wave components in typical optical systems.

A number of simple worked examples are contained within the body of each chapter. In the main, these are design exercises, intended to illustrate the application of particular formulae or the values of important parameters. However, each chapter is also provided with a set of example questions. These are more demanding; they are intended to stimulate further thought, and often cover material omitted from the text. Worked solutions to most of these problems are gathered together at the end of the book.
CHAPTER
TWO

ELECTROMAGNETIC FIELDS AND
PLANE WAVES

2.1 MAXWELL’S EQUATIONS

The understanding of any field of physics or electrical engineering requires a suitable theoretical basis. In optics, we are fortunate that two highly developed and accurate theories are available. In the older theory - often described as the ‘classical theory’ - the behaviour of light is described in terms of electromagnetic fields and waves. This is particularly appropriate for the analysis of passive devices, where the absorption and emission of radiation is unimportant, and consequently where the interaction of a wave with matter may represented in a somewhat phenomenological way.

In the newer theory - the quantum theory - a different model is employed. Light is considered to be composed of photons, which are the elementary units or quanta of radiation. The interaction of light and matter is then understood in terms of exchanges of energy between photons and electrons. For example, the generation of a photon may be identified with the transition of a single electron between two energy levels. Quantum theory is therefore directly applicable to active optical devices. The development of this alternative model, and of its later incarnation, quantum electrodynamics, occupied the first half of the century, and involved many of the world's foremost physicists.

Quantum theory may also describe situations that do not involve matter at all, but are still not accurately represented by classical theory. One example is provided by low light levels, where the photon flux may be extremely small and the arrival of radiation in discrete units is important. However, for a high enough photon flux, the two theories are equivalent. We shall encounter both of them in this book. Since our early discussions will centre on passive devices, we shall begin with a classical approach, turning only to the quantum theory at a much later stage (Chapter 11).

In classical theory, the laws of electricity and magnetism are described by Maxwell's equations. These represent the result of a synthesis of several existing theories and experimental observations by James Clerk Maxwell (1831-1879). In effect, Maxwell's equations are a set of relations linking the values of a number of quantities that describe electric and magnetic fields. These are the electric flux density \( \mathbf{D} \), the magnetic flux density \( \mathbf{B} \), the electric field strength \( \mathbf{E} \), the magnetic field strength \( \mathbf{H} \), and the current density \( \mathbf{J} \). All are vector quantities, and are functions not only of the three spatial coordinates \( x, y, z \) but also of time \( t \). Of these parameters, \( \mathbf{D}, \mathbf{B}, \mathbf{E} \) and \( \mathbf{H} \) are the most important to high frequency electromagnetic theory. In this regime, they are essentially 'inventions'; they are not directly observable, but are linked by a self-consistent set of equations, which correctly predict the magnitudes of other measurable quantities (like the flow and distribution of power).

Often, different representations are used for the fields; for example, it is common to separate their time- and space-variation. The electric field strength \( \mathbf{E} \) may therefore be written alternatively as:

\[
\mathbf{E}(x, y, z, t) = \mathbf{E}(x, y, z) \, \tilde{f}(t)
\]
Here $E$ accounts for the spatial variation of the field, and $f(t)$ for temporal changes. Note the use of a bold-faced type for the complete field $E$, and lighter type for the time-independent field $E$.

It is also common to refer to individual elements of the vectors concerned. In cartesian coordinates, we may write the time-dependent electric field $E$ as:

$$E(x, y, z, t) = [E_x(x, y, z, t), E_y(x, y, z, t), E_z(x, y, z, t)]$$

Alternatively, we could write for the time-independent field $E$:

$$E(x, y, z) = [E_x(x, y, z), E_y(x, y, z), E_z(x, y, z)]$$

Here $E_x$ refers to the $x$-component of $E$, and so on.

Because of the nature of the fields, Maxwell's equations are written in terms of vector calculus. Though this is hard for the novice to begin with, the notation rapidly becomes a useful (not to say essential) tool-of-the-trade in electromagnetic theory. Unfortunately, for reasons of space, we must assume here some familiarity with the basic techniques involved.

The equations can be written in two different forms. From experimental observations, by and large carried out in the previous century, the common integral version was obtained. This is generally presented in the form of a series of laws. The first pair of these (Gauss' law, and its magnetic equivalent) describe relations which are most important for static fields.

1. GAUSS' LAW

Consider a closed surface, surrounding a volume containing electric charges, as shown in Figure 2.1-1. **Gauss' law** (named after the German mathematician Karl Gauss, 1777-1855) states that the electric flux flowing out of the surface is equal to the charge enclosed inside the volume. This assertion can be written mathematically in terms of the vector fields involved in Maxwell's theory in the following way.

$$\nabla \cdot D = \frac{\rho}{\varepsilon}$$

**Figure 2.1-1** Geometry for illustration of Gauss' Law

Firstly, we define a vector $\text{da}$ as having magnitude equal to the area $\text{da}$ of a small element of the surface, and a direction normal to that element. Secondly, we introduce a new operation, the **scalar or dot product** between two vectors $F$ and $G$. This is written as $F \cdot G$, and is defined in cartesian coordinates as:

$$F \cdot G = F_x G_x + F_y G_y + F_z G_z$$

Simple trigonometry may then be used to show that Equation 2.1-4 can also be written as:
\[ \mathbf{E} \cdot \mathbf{G} = |\mathbf{E}| \cdot |\mathbf{G}| \cdot \cos(\theta) \]  

2.1-5

Where \( |\mathbf{E}| \) and \( |\mathbf{G}| \) are the moduli or lengths of the two vectors, and \( \theta \) is the angle between them.

Consequently, if \( \mathbf{D} \) is the electric flux density, the term \( \mathbf{D} \cdot \, da \) represents the product of the component of \( \mathbf{D} \) normal to the small surface element \( da \) and the area of the element. When integrated over the whole of the surface, this will give the net outward normal electric flux. Similarly, if we define the scalar term \( \rho \) as the charge density, the integral of \( \rho \) over the whole volume must give the charge enclosed. We may therefore write Gauss’ law for vector fields as:

\[ \iint \mathbf{D} \cdot \, da = \iiint \rho \, dv \]  

2.1-6

Here the left-hand integral is a surface integral, taken over the whole of the closed surface, while the right-hand one is a volume integral, over the volume enclosed.

2. MAGNETIC EQUIVALENT OF GAUSS’ LAW

We can do the same thing for magnetic flux density. If we now consider the magnetic flux flowing out of a closed surface, we get a similar picture, shown in Figure 2.1-2.

![Figure 2.2](image)

Figure 2.2  Geometry for illustration of the magnetic equivalent of Gauss’ Law.

Comparison with Figure 2.1-1 shows that the resulting integral equation must have a similar form. However, no magnetic monopoles (the magnetic equivalent of electric charges) have ever been found experimentally, despite extensive searches. Evidence of this curious fact is provided by the simple bar magnet, which has both north and south poles. However, if the magnet is divided into two, each half will also have two poles, and no amount of further subdivision will isolate a monopole. In this case, therefore, the right-hand side of the equation is zero, giving:

\[ \iint \mathbf{B} \cdot \, da = 0 \]  

2.1-7

Where \( \mathbf{B} \) is the magnetic flux density.

The second pair of laws (Faraday’s and Ampère’s laws) describe relations which are of greater significance for time-varying fields.
3. FARADAY’S LAW OF MAGNETIC INDUCTION

Consider a time-varying magnetic flux passing through a closed loop \( L \), defining the rim of an open surface, as shown in Figure 2.1-3.

![Figure 2.1-3 Geometry for illustration of Faraday’s Law.](image)

Now, the flux of magnetic induction \( \psi_B \) through the open surface is:

\[
\psi_B = \iiint_A B \cdot d\alpha
\]

The electromotive force \( E \) induced round the loop is therefore:

\[
E = -\frac{\partial \psi_B}{\partial t}
\]

However, we know from the integral relationship between electric potential and field strength that this can also be written as a line integral, in the form:

\[
E = \int_L E \cdot dL
\]

Here \( E \) is the electric field strength, \( dL \) is a small element of the closed loop in Figure 2.1-3, and the integral is taken round the whole loop. By comparing Equations 2.1-9 and 2.1-10, and using Equation 2.1-8 we obtain Faraday’s law, named after Michael Faraday (1791-1867):

\[
\int_L E \cdot dL = -\iiint_A \frac{\partial B}{\partial t} \cdot d\alpha
\]

Faraday’s law implies that a time-varying magnetic field must have an electric field associated with it. This feature is of great importance for electromagnetic waves, as we shall see later on.

4. AMPÈRE’S LAW

Now consider the flow of current through a closed loop, of the same geometry, as shown in Figure 2.1-4. Ampère’s law (named after André Ampère, 1775-1836) states that:

\[
\int_L H \cdot dL = \iint_A J \cdot d\alpha
\]

where \( H \) is the magnetic field strength and \( J \) is the current density. This implies that moving charges give rise to a magnetic field, a notion in pleasing symmetry with Faraday’s law.
In addition to these laws, there are a set of relationships known as the material equations, which link field strengths with flux densities, through a set of material coefficients that are representative of the bulk properties of matter. There are three of them, written:

\[
\begin{align*}
    \mathbf{J} &= \sigma \mathbf{E} \\
    \mathbf{D} &= \varepsilon \mathbf{E} \\
    \mathbf{B} &= \mu \mathbf{H}
\end{align*}
\]

2.1-13

Here \(\sigma\) is the conductivity, \(\varepsilon\) is the permittivity, and \(\mu\) is the permeability. The physical origin and significance of some of these quantities will be discussed in Chapter 3. Note that the first equation is a vectorial form of Ohm’s law, commonly encountered in electrical engineering. For the dielectric materials used to guide high frequency electromagnetic waves (which we will mainly consider here) the conductivity is typically zero, while the permeability is that of free space. The former implies that the current density \(\mathbf{J}\) is zero; there are also no free charges, so that \(\rho\) is zero.

However, this is not quite the end of the story, because experiments showed that magnetic fields can also be measured in free space - for example, between the plates of a capacitor, while it is being charged. Consequently, electromagnetic theory was modified by Maxwell (in what amounts to a stroke of genius) to cope with this observation.

5. GENERALISED FORM OF AMPÈRE’S LAW

Consider the parallel-plate capacitor shown in Figure 2.1-5, which is linked to a circuit by wires carrying a current \(I\). Clearly, time-varying currents can somehow travel round the circuit, despite the absence of conducting material in the region between the plates.

Figure 2.1-5  Geometry for derivation of Maxwell’s displacement current.
To account for the flow of current across this apparent break in the circuit, Maxwell suggested the existence of a new type of current. This is known as the displacement current, and is calculated as follows. If A is the area of each plate, and Q is the charge on it, then the electric field E between the plates is:

$$E = \frac{Q}{\varepsilon A}$$

2.1-14

As the charge varies, the electric field changes, so that $\varepsilon \frac{dE}{dt} = I/A$ is effectively a current density. Maxwell therefore defined a vector displacement current density $J_D$ as:

$$J_D = \varepsilon \frac{\partial E}{\partial t} = \frac{\partial D}{\partial t}.$$  

2.1-15

This displacement current must be added into any calculation involving the 'normal' conduction current. The only example we have come across so far is Ampère's law. Including the displacement current, Ampère's law must be restated as:

$$\int_L \mathbf{H} \cdot dL = \iint_A [\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}] \cdot dA$$

2.1-16

This simple modification was of great historical importance, since it showed that time-varying electric fields can exist even in the absence of normal currents (i.e. when $\mathbf{J} = 0$). It allowed Maxwell to justify theoretically the electromagnetic waves discovered experimentally by Heinrich Hertz in 1888, thus demonstrating at a stroke that static electric and magnetic fields, and radio and light waves, are all part of the wider phenomenon of electromagnetism.

### 2.2 THE DIFFERENTIAL FORM OF MAXWELL'S EQUATIONS

We can use two standard vector theorems to transform the integral equations above into an alternative differential form, which will be of great use in later calculations. The theorems are described in detail in many mathematics texts, and so will simply be stated here.

#### 1. GAUSS' THEOREM

The first is Gauss' theorem (not to be confused with Gauss' law), which states that for any vector field $\mathbf{F}$ the following relation holds:

$$\iiint_A \mathbf{F} \cdot d\mathbf{a} = \iiint_V \text{div}(\mathbf{F}) \, dv$$

2.2-1

Here the divergence of a vector field $\mathbf{F}$ is an important new quantity, a scalar term, which is defined in cartesian coordinates by:

$$\text{div}(\mathbf{F}) = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$$

2.2-2

div($\mathbf{F}$) is also often written as $\nabla \cdot \mathbf{F}$, where the vector operator $\nabla$ (or 'del') is defined in cartesian coordinates as:

$$\nabla = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z}$$

2.2-3
Here \( \mathbf{i} \), \( \mathbf{j} \) and \( \mathbf{k} \) are unit vectors in the x-, y- and z-directions, respectively. Hence, Gauss’ law becomes:

\[
\iint_{\Delta} \mathbf{D} \cdot d\mathbf{a} = \iiint_{\gamma} \text{div}(\mathbf{D}) \, dv = \iiint_{\gamma} \rho \, dv
\]

By examining the latter part of Equation 2.2-4, we see that an integral-free relation between \( \mathbf{D} \) and \( \rho \) can be obtained, in the form:

\[
\text{div}(\mathbf{D}) = \rho
\]

Similarly, we can show from the magnetic equivalent of Gauss’ law that \( \text{div}(\mathbf{B}) = 0 \).

2. STOKES’ THEOREM

A similar type of theorem, due to Stokes, states that for any vector field \( \mathbf{F} \):

\[
\int_{L} \mathbf{F} \cdot d\mathbf{L} = \iint_{\Lambda} \text{curl}(\mathbf{F}) \cdot d\mathbf{a}
\]

Here we have introduced another new quantity, the \text{curl} of a vector field \( \mathbf{F} \). This is itself a vector, and is defined in cartesian coordinates as:

\[
\text{curl}(\mathbf{F}) = \mathbf{i} \left\{ \frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right\} + \mathbf{j} \left\{ \frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \right\} + \mathbf{k} \left\{ \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right\}
\]

Equation 2.2-7 is a rather long-winded expression. A more compact version (which is also rather easier to remember) can be written in the form of a determinant, as shown below:

\[
\text{curl}(\mathbf{F}) = \begin{vmatrix}
\mathbf{i} & \mathbf{j} & \mathbf{k} \\
\frac{\partial F_y}{\partial z} & \frac{\partial F_z}{\partial x} & \frac{\partial F_x}{\partial y} \\
F_x & F_y & F_z
\end{vmatrix}
\]

Here \( |\mathbf{M}| \) represents the determinant of \( \mathbf{M} \), where \( \mathbf{M} \) is a general matrix. Curl(\( \mathbf{F} \)) is also often written as \( \nabla \times \mathbf{F} \), where the symbol \( \times \) denotes a further new operation, the \text{vector product}.

The vector product \( \mathbf{F} \times \mathbf{G} \) between two general fields \( \mathbf{F} \) and \( \mathbf{G} \) is defined as:

\[
\mathbf{F} \times \mathbf{G} = \mathbf{i} \left\{ F_y G_z - F_z G_y \right\} + \mathbf{j} \left\{ F_z G_x - F_x G_z \right\} + \mathbf{k} \left\{ F_x G_y - F_y G_x \right\}
\]

This can also be written in determinantal form, as:

\[
\mathbf{F} \times \mathbf{G} = \begin{vmatrix}
\mathbf{i} & \mathbf{i} & \mathbf{k} \\
F_x & F_y & F_z \\
G_x & G_y & G_z
\end{vmatrix}
\]

Note that the vector product (unlike the scalar product) is not commutative, so that \( \mathbf{F} \times \mathbf{G} \neq \mathbf{G} \times \mathbf{F} \). However, with the definition given above, \( \nabla \times \mathbf{F} \) reduces to both Equations 2.2-7 and 2.2-8.

Using Stokes’ theorem, Faraday’s law transforms to:

\[
\int_{l} \mathbf{E} \cdot d\mathbf{L} = \iint_{\Lambda} \text{curl}(\mathbf{E}) \cdot d\mathbf{a} = -\iint_{\Lambda} \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a}
\]
Again, examining the latter part of Equation 2.2-11, we see that:

\[
\text{curl}(\mathbf{E}) = -\frac{\partial \mathbf{B}}{\partial t}
\]

Similarly, we can show from Ampère's law that:

\[
\text{curl}(\mathbf{H}) = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}
\]

We have now derived the differential form of Maxwell's equations. Because they contain only differential operators, they are often much easier to manipulate than the integral form. For completeness, we show all the new equations grouped together below:

\[
\begin{align*}
\text{div}(\mathbf{D}) &= \rho \quad (1) \\
\text{div}(\mathbf{B}) &= 0 \quad (2) \\
\text{curl}(\mathbf{E}) &= -\frac{\partial \mathbf{B}}{\partial t} \quad (3) \\
\text{curl}(\mathbf{H}) &= \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (4)
\end{align*}
\]

Note that when \( \rho = 0 \) and \( \mathbf{J} = \mathbf{0} \) (which is the case for electromagnetic waves in dielectric media) the equations for electric and magnetic quantities appear interchangeable.

2.3 HARMONICALLY-VARYING FIELDS AND THE WAVE EQUATION

One striking success of Maxwell's equations is to predict the existence of harmonically-varying fields, otherwise known as electromagnetic waves. We shall now perform a similar demonstration, with the following assumptions: a) We restrict ourselves to non-magnetic materials, so that \( \mu = \mu_0 \), where \( \mu_0 \) is the permeability of free space (known from electrostatic experiment to have the value \( 4\pi \times 10^{-7} \text{ m kg/C}^2 \)). b) We assume that there are no currents flowing, and no charges present, so that \( \mathbf{J} \) and \( \rho \) are both zero.

THE WAVE EQUATION

We start by deriving a suitable wave equation. If we put \( \mathbf{D} = \varepsilon \mathbf{E} \) and \( \mathbf{B} = \mu_0 \mathbf{H} \) in Equation 2.2-14, then Equations 3 and 4 contain only the two variables \( \mathbf{E} \) and \( \mathbf{H} \). We can therefore eliminate one or other by direct manipulation. Taking the curl of Equation 3 gives:

\[
\text{curl} [\text{curl}(\mathbf{E})] = -\mu_0 \frac{\partial^2 \mathbf{D}}{\partial t^2}
\]

We now simplify Equation 2.3-1 using a standard vector identity:

\[
\text{curl} [\text{curl}(\mathbf{F})] = \text{grad} [\text{div}(\mathbf{F})] - \nabla^2 \mathbf{F}
\]

Here we have introduced two new operators. The first is the gradient of a scalar function \( \phi \) (which in the case of Equation 2.3-2 is \( \text{div}(\mathbf{F}) \)). This is defined in cartesian coordinates as:

\[
\text{grad}(\phi) = \nabla \phi = \frac{\partial \phi}{\partial x} \mathbf{i} + \frac{\partial \phi}{\partial y} \mathbf{j} + \frac{\partial \phi}{\partial z} \mathbf{k}
\]
The second is the operator $\nabla^2$ (known as the **Laplacian**), defined in cartesian coordinates by:

$$\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$$  \hspace{1cm} 2.3-4

Using this identity, Equation 2.3-1 can be reduced to:

$$\text{grad} \left[ \text{div}(\mathbf{E}) \right] - \nabla^2 \mathbf{E} = -\mu_0 \partial^2(\varepsilon\mathbf{E})/\partial t^2$$  \hspace{1cm} 2.3-5

This is now a form of wave equation. This be simplified considerably, by making assumptions and approximations that are frequently valid:

a) Since, with no charges present, $\text{div}(\mathbf{D}) = 0$, it follows that $\text{div}(\varepsilon\mathbf{E}) = 0$. If the medium is homogeneous and isotropic, then $\varepsilon$ is independent of position and direction, so $\text{div} \mathbf{E} = 0$. If $\varepsilon$ varies only slowly with position, we can still put $\text{div}(\mathbf{E}) = 0$ to a reasonable approximation. Consequently, the first term in Equation 2.3-5 is often zero.

b) If $\varepsilon$ is not time-dependent, then $\partial^2(\varepsilon\mathbf{E})/\partial t^2 = \varepsilon \partial^2 \mathbf{E}/\partial t^2$. This is normally a good assumption; even if $\varepsilon$ varies with time, it needs to do so significantly in a period of oscillation of the field before it matters. Non-linear optics is concerned precisely with circumstances in which $\varepsilon$ varies with $t$, but high fields are required for significant effects to occur.

Assuming approximations a) and b), Equation 2.3-5 becomes:

$$\nabla^2 \mathbf{E} = \mu_0 \varepsilon \partial^2 \mathbf{E}/\partial t^2$$  \hspace{1cm} 2.3-6

This is also a **vector wave equation**. In cartesian coordinates, it can be written as a set of three independent scalar equations - one for each co-ordinate. For the $E_x$ component, for example, we have:

$$\partial^2 E_x/\partial x^2 + \partial^2 E_x/\partial y^2 + \partial^2 E_x/\partial z^2 = \mu_0 \varepsilon_0 \varepsilon_r \partial^2 E_x/\partial t^2$$  \hspace{1cm} 2.3-7

Here we have written the permittivity $\varepsilon$ as a product, in the form:

$$\varepsilon = \varepsilon_0 \varepsilon_r$$  \hspace{1cm} 2.3-8

Here $\varepsilon_0$ is the permittivity of free space, again known from experiments to have the value $8.85 \times 10^{-12} \text{ s}^2 \text{ C}^2/\text{m}^3 \text{ kg}$, while $\varepsilon_r$ is the **relative permittivity** or **relative dielectric constant** of the material concerned.

Equation 2.3-7 is now in the form of a simple classical scalar wave equation, which is adequate for introducing many aspects of electromagnetic wave propagation. It should be compared with the one-dimensional wave equation for waves on a string, which is often written as:

$$\partial^2 y/\partial x^2 = (1/c^2) \partial^2 y/\partial t^2$$  \hspace{1cm} 2.3-9

Here $y$ is the displacement of the string, $x$ is the distance along it, and $c$ is the wave velocity. Both equations are essentially similar in character, involving second derivatives of some quantity with respect to space (on the left-hand side) and second derivatives with respect to time (on the right). The main difference is that electromagnetic waves are not constrained to
travel in any particular direction, so that all three spatial coordinates appear in Equation 2.3-7.

Note that an exactly similar equation to 2.3-6 can be obtained for the magnetic field $\mathbf{H}$:

$$\nabla^2 \mathbf{H} = \mu_0 \varepsilon_0 \varepsilon_r \frac{\partial^2 \mathbf{H}}{\partial t^2}$$

This follows from the 'interchangeability' of the equations for electric and magnetic field quantities discussed earlier. There will be occasions when it is easier to work with Equation 2.3-6 than 2.3-10, and vice versa.

THE TIME-INDEPENDENT WAVE EQUATION

Now, the electric field $\mathbf{E}$ is, in general, a function of $x$, $y$, $z$ and $t$. However, it is often the case that all the fields involved will be harmonically varying, at some single angular frequency $\omega$. This will occur in many situations involving monochromatic light. It is then convenient to eliminate any time-dependence from the problem. One possibility is to assume cosinusoidally-varying solutions to the wave equation, but it is generally far more convenient to use the complex exponential form:

$$\mathbf{E}(x, y, z, t) = \mathbf{E}(x, y, z) \exp(j\omega t)$$

Here the function $\mathbf{E}(x, y, z)$ accounts for the spatial dependence of the field, while the exponential $\exp(j\omega t)$ describes the time-variation. The use of such complex notation is standard in electromagnetic theory, but it is important to note that ultimately only the real parts of Equation 2.3-11 are significant.

Generally, entire calculations are performed in books with the simple understanding that real parts are implied throughout. With this assumption, we can find time-derivative terms as:

$$\frac{\partial \mathbf{E}}{\partial t} = j\omega \mathbf{E}; \quad \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\omega^2 \mathbf{E}$$

and so on, so that the wave equation reduces to:

$$\nabla^2 \mathbf{E} = -\omega^2 \mu_0 \varepsilon_0 \varepsilon_r \mathbf{E}$$

Equation 2.3-13 is a time-independent vector wave equation, which is valid for fields oscillating at a single angular frequency $\omega$. We will now try to find some solutions to it.

2.4 PLANE WAVES

The simplest form of solution is a plane wave, i.e. a wave whose surfaces of constant phase are infinite planes, perpendicular to the direction of propagation. Figure 2.4-1 shows the geometry for a wave travelling in the $+z$-direction.

Clearly, in this case no field quantities can vary with the transverse coordinates $x$ and $y$; the only spatial variation is caused by changes in $z$. Considering the electric field to begin with, we must therefore have:

$$\frac{\partial \mathbf{E}}{\partial x} = \frac{\partial \mathbf{E}}{\partial y} = 0$$

so that $\mathbf{E}$ must be a function of $z$ only. We may therefore replace $\frac{\partial \mathbf{E}}{\partial z}$ by $\frac{d\mathbf{E}}{dz}$, and so on.
For simplicity, we now consider the electric field to have one single component, say in the x-direction. The vectorial wave equation 2.3-13 then reduces to the scalar equation:

\[ \frac{d^2E_x}{dz^2} + \omega^2 \mu_0 \varepsilon_0 \varepsilon_r E_x = 0 \]

2.4-2

We now guess that a possible solution has the form:

\[ E_x = E_{x+} \exp(-jkz) \]

2.4-3

where \( E_{x+} \) is a constant. Direct substitution into 2.4-2 shows that this solution is valid, provided:

\[ k = \omega \sqrt{\mu_0 \varepsilon_0 \varepsilon_r} \]

2.4-4

Hence, the full solution, including the time variation, is:

\[ E_x = E_{x+} \exp[j(\omega t - kz)] \]

2.4-5

Equation 2.4-5 represents a travelling wave, moving in the +z-direction. \( E_{x+} \) is clearly the wave amplitude. What do the other constant terms represent? Well, \( \omega = 2\pi f \) is the angular frequency of the wave; typically, the corresponding temporal frequency \( f \) is \( \approx 10^{15} \) Hz for light waves. We could also write \( \omega = 2\pi/T \), where \( T \) is the period of the oscillating field. By analogy, we could write \( k = 2\pi/\lambda \), introducing the new quantity \( \lambda \), the spatial wavelength. This is the distance separating planes of equal phase, as shown in Figure 2.4-1. Typically, \( \lambda \) lies in the approximate range 0.4 - 0.8 \( \mu m \) for visible light; the lower limit corresponds to the ultra-violet or blue end of the spectrum, while the latter corresponds to deep red or near infrared wavelengths. \( k \) is an important new parameter, known as the propagation constant, which we will refer to often later on.

Finally, the velocity of the wave (known as the phase velocity) is given by:

\[ v_{ph} = \frac{\omega}{k} \]

2.4-6

This can also be written in the more familiar form:

\[ v_{ph} = \frac{f}{\lambda} \]

2.4-7

From Equation 2.4-4, we can find the phase velocity as:

\[ v_{ph} = 1/\sqrt{\mu_0 \varepsilon_0 \varepsilon_r} \]
In a vacuum, $\varepsilon_r = 1$, so $v_{ph} = 1/\sqrt{\mu_0\varepsilon_0}$. Putting in the numbers for $\mu_0$ and $\varepsilon_0$, we get:

$$v_{ph} = \frac{1}{\sqrt{(8.85 \times 10^{-12} \times 4\pi \times 10^{-7})}} \approx 3 \times 10^8 \text{ m/s}.$$  

This is the velocity of light, written as $c$, and one of the major successes of Maxwell's equations was the discovery that this directly measurable quantity could be derived so simply. In 1849, Fizeau measured $c$ as $3.153 \times 10^8$ m/s, while in 1983 the value was fixed at $2.99792458 \times 10^8$ m/s. In other media, $\varepsilon_r$ is normally greater than unity, so that $v_{ph} < c$; light therefore generally travels slower in matter than in free space. The refractive index $n$ (a useful parameter in optics) is then defined as:

$$n = \frac{c}{v_{ph}} = \sqrt{\varepsilon_r}.$$  

Often, quantities measured in a particular material are referred to those measured in free-space. For example, the spatial wavelength $\lambda$ could be related to the free-space wavelength $\lambda_0$ by:

$$\lambda = \lambda_0 / n.$$  

Similarly, the propagation constant $k$ could be related to the free-space propagation constant $k_0$ by:

$$k = nk_0.$$  

Typically, the relative permittivity $\varepsilon_r$ and the refractive index $n$ are both functions of frequency $\omega$, as we shall see in the next chapter.

THE TRANSVERSE NATURE OF ELECTROMAGNETIC WAVES

We will now consider the other field components that must accompany the solution we have found for the electric field. First, we note that with the assumptions we have made so far, Maxwell's equations can be rewritten in the following time-independent form:

\[
\begin{align*}
\text{div}(\varepsilon \mathbf{E}) &= 0 \quad (1) \\
\text{div}(\mu_0 \mathbf{H}) &= 0 \quad (2) \\
\text{curl}(\mathbf{E}) &= -j\omega\mu_0 \mathbf{H} \quad (3) \\
\text{curl}(\mathbf{H}) &= j\omega\varepsilon_0 \mathbf{E} \quad (4)
\end{align*}
\]

Now, our solution has so far contained only an $x$-component of the electric field. In this case:

$$\text{curl}(\mathbf{E}) = j \frac{\partial E_x}{\partial z} - k \frac{\partial E_x}{\partial y}.$$  

Given that $E_x = E_{x+} \exp(-jkz)$, we find:

$$\text{curl}(\mathbf{E}) = -jk E_{x+} \exp(-jkz) j.$$  

Now, from Equation (3) in 2.4-13, we must have $\text{curl}(\mathbf{E}) = -j\omega\mu_0 \mathbf{H}$. Hence, the magnetic field accompanying our solution only has a component in the $y$-direction. Writing this as:

$$\mathbf{H} = H_y \exp(-jkz) j.$$  

R.R.A. Sym s and J.R. Cozens  Optical Guided Waves and Devices  12
we can obtain the following constant relation between the electric and magnetic field amplitudes:

\[ \frac{H_y}{E_x} = \frac{k}{\omega \mu_0} \]  

The solution therefore really consists of two travelling waves - an electric and a magnetic component. Both are in-phase, but the field directions are at right angles to each other. We can represent the complete solution at any given instant in time as in Figure 2.4-2, which shows the real parts of the two components together. Both exhibit similar cosinusoidal variations with distance.

![Figure 2.4-2 A plane electromagnetic wave.](image)

Is this solution the only one possible? It would seem reasonable to repeat the analysis, starting with the assumption that the electric field only has a component in the y-direction. In this case, we find that if \( E = E_{y+} \exp(-jkz) \), then \( H = H_{x+} \exp(-jkz) \), so the magnetic field now only has a component in the x-direction. As before, we can find a relation between the two field amplitudes. This time, we get:

\[ \frac{H_{x+}}{E_{y+}} = \frac{-k}{\omega \mu_0} \]  

Apart from the minus sign, the amplitude ratio is as before.

What happens if we assume instead that the electric field only has a component in the z-direction? Well, in an isotropic medium, \( \text{div}(\varepsilon E) = \varepsilon \text{div}(E) \), so Equation (1) in 2.4-13 must reduce to \( \text{div}(E) = 0 \). Remember that we can expand this as:

\[ \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0 \]  

However, since we have already assumed that \( \frac{\partial E_x}{\partial x} \) and \( \frac{\partial E_y}{\partial y} = 0 \), it follows that \( \frac{\partial E_x}{\partial x} = \frac{\partial E_y}{\partial y} = 0 \). Hence, \( \frac{\partial E_z}{\partial z} \) must be zero, so \( E_z \) must be a constant independent of \( z \). We therefore do not find travelling-wave solutions for \( E_z \), and a similar argument can be used to show that there are no wave solutions for \( H_z \). Plane electromagnetic waves are therefore strictly transverse. They are therefore often described as **TEM** (standing for Transverse ElectroMagnetic) waves.
OPTICAL POLARIZATION

We now consider some of the wider properties of the solutions found so far, beginning with the important feature of optical polarization. We start by noting that the two independent travelling wave solutions discussed above can be combined into a more general solution, in the form:

\[ E = E_x \exp[j(\omega t - k z + \phi_x)] i + E_y \exp[j(\omega t - k z + \phi_y)] j \]

2.4-20

where \( \phi_x \) and \( \phi_y \) are arbitrary (but constant) phase factors. The nature of the resulting wave then depends on the values of \( E_x, E_y, \phi_x \) and \( \phi_y \). Several combinations are particularly important.

(i) If \( \phi_x = \phi_y \), the solution can be written as:

\[ E = [E_x i + E_y j] \exp[j(\omega t - k z + \phi)] \]

= \( E_0 \exp[j(\omega t - k z + \phi)] \)

2.4-21

In this solution, the direction of the electric field vector is independent of time and space, and is defined by a new vector \( E_0 \), which is the vectorial sum of \( E_x i + E_y j \) as shown in Figure 2.4-3. This type of wave is known as a **linearly polarized** wave, and the direction of the electric field vector \( E_0 \) represents the **direction of polarization**. Linearly polarized light is particularly important in engineering optics. It can be produced from natural light (which has random polarization) by passing it through a **polarizer**. More importantly, it is emitted directly by many types of laser.

![Figure 2.4-3 Construction of the polarization vector.](image)

(ii) If \( E_x = E_y \), and \( \phi_y = \phi_x \pm \pi/2 \), the solution can be written as:

\[ E = E_0 \exp[j(\omega t - k z + \phi)] j + E_0 \exp[j(\omega t - k z + \phi \pm \pi/2)] j \]

2.4-22

Or, alternatively, as:

\[ E = E_0 (i \pm j) \exp[j(\omega t - k z + \phi)] \]

2.4-23

Now, ultimately, we are interested in the real part of \( E \). This is given by:

\[ \text{Re}\{E\} = E_0 (\cos(\omega t - k z + \phi) i \pm \sin(\omega t - k z + \phi) j) \]

2.4-24

In this case, the amplitude of the electric field vector is still constant (and equal to \( E_0 \)), but the direction of polarization is not. Instead, it rotates as a function of space and time. This
solution is known as a **circularly polarized** wave, because the locus traced of the electric field vector as a function of time (at a given point) is a circle, as shown in Figure 2.4-4a. Right- and left-hand circular polarizations are both possible, depending on the sign of the $\pi/2$ phase-shift. If $E_x \neq E_y$, the locus becomes an ellipse, and the wave is described as being **elliptically polarized** (Figure 2.4-4b).

![Figure 2.4-4 Loci of the electric field vector for a) circular and b) elliptic polarization.](image)

**OBLIQUE WAVES**

Simple solutions can also be found for plane waves travelling in different directions. For example, Figure 2.4-5 shows a wave, travelling in the $x$ - $z$ plane at an angle $\theta$ to the $z$-axis. If the wave is linearly polarized in the $y$-direction, i.e. perpendicular to the plane of the figure, the time-independent electric field is given simply by:

$$E = j E_y \exp[-j(kx \sin(\theta) + kz \cos(\theta))]$$

This result is obtained simply by rotating the co-ordinate system of Figure 2.4-1 about the $y$-axis. We can check that it is correct by putting $\theta = 0$, whereupon Equation 2.4-25 reduces to our original solution, $E = j E_y \exp(-jkz)$. Similarly, if $\theta = \pi/2$, we get $E = j E_y \exp(-jx) - this is also as expected, as the wave is now travelling in the $+x$ direction.

![Figure 2.4-5 An obliquely-travelling plane wave.](image)

However, we note that in the time it takes the wave to travel a distance $\lambda$ in the direction of propagation, the phase-fronts have advanced further in the $z$-direction, a distance $\lambda' = \lambda / \cos(\theta)$. The effective phase velocity in the $z$-direction is therefore greater than $v_{ph}$, by the factor $1/\cos(\theta)$. Some thought is required fully to appreciate this point, since it implies faster motion of the disturbance in a direction off-axis to the direction of propagation.
IMPEDEANCE

We can find the impedance of a medium through which a wave is propagating, as follows. Since \( \omega/k = 1/\sqrt{\mu_0 \varepsilon_0 \varepsilon_r} \), the ratio \( E_x/H_y \) can be found as \( \sqrt{\mu_0 / \varepsilon_0 \varepsilon_r} \). Since this ratio has the dimensions of ohms, it is called the characteristic impedance \( Z \) of the medium. Thus, we may put:

\[
Z = \sqrt{\mu_0 / \varepsilon_0 \varepsilon_r}
\]

For free space, \( \varepsilon_r = 1 \), so that \( Z_0 = \sqrt{\mu_0 / \varepsilon_0} = 377 \Omega \). Anyone owning a radio tuner with an external aerial socket should note the input impedance - it will be around this value. The impedance of any other material may be related to the impedance of free space using:

\[
Z = Z_0 / \sqrt{\varepsilon_r} = Z_0 / n
\]

Impedance is therefore inversely proportional to refractive index.

WAVES IN LOSSY DIELECTRICS

We may also extend the analysis to describe the behaviour of plane waves propagating in slightly lossy dielectric media, as follows. We begin by including loss in a phenomenological way, by assuming that the relative dielectric constant of the material is complex-valued. We shall justify this more rigorously in Chapter 3, but for the time being we will simply take \( \varepsilon_r \) to be defined by:

\[
\varepsilon_r = \varepsilon_r' - j \varepsilon_r''
\]

For x-polarized waves travelling in the +z-direction, the scalar wave equation we must solve can be found by substituting Equation 2.4-28 into Equation 2.4-2. We get:

\[
d^2E_x/dz^2 + \omega^2 \mu_0 \varepsilon_0 [\varepsilon_r' - j \varepsilon_r'`] E_x = 0
\]

Assuming a solution in the form used previously, namely \( E_x = E_x \exp(-jkz) \), we find that the propagation constant \( k \) is now given by:

\[
k^2 = \omega^2 \mu_0 \varepsilon_0 [\varepsilon_r' - j \varepsilon_r'`] \]

Or:

\[
k = \omega \sqrt{\mu_0 \varepsilon_0 \varepsilon_r'} \sqrt{1 - j \varepsilon_r'' / \varepsilon_r'}
\]

Making the additional assumption that the loss is small, so that \( \varepsilon_r'' \ll \varepsilon_r' \), we may use a binomial approximation for Equation 2.4-31, which gives:

\[
k = \omega \sqrt{\mu_0 \varepsilon_0 \varepsilon_r'} \left[ 1 - j \varepsilon_r'' / 2 \varepsilon_r' \right] = k' - jk''
\]

where the real and imaginary parts of \( k \) are given by:

\[
k' = \omega \sqrt{\mu_0 \varepsilon_0 \varepsilon_r'} , \text{ and } k'' = \varepsilon_r'' / 2 \varepsilon_r'
\]

Since the propagation constant is complex, we can now rearrange the solution as the product of two exponentials, as:
Equation 2.4-34 has the form of a plane wave, whose amplitude decays exponentially with distance \( z \). The real part of the propagation constant, \( k' \), defines the phase variation of the wave, while the imaginary part \( k'' \) defines the amplitude variation. \( k'' \) is known as the absorption constant, and is often given the symbol \( \alpha \). The presence of even a small amount of loss in a material (which is often unavoidable) therefore causes the exponential decay of a propagating wave.

We may also hypothesise the existence of media with a negative value of \( \varepsilon'' \). In this case, the solution corresponds to an exponentially growing wave, of the form:

\[
E_x = E_{x*} \exp(+gz) \exp(-jk'z)
\]

where \( g \), the gain constant of the medium, is defined by \( g = -k'' \). The growth of an optical wave as it passes through a medium with gain is the key to the operation of all laser devices.

### WAVES IN METALS

Finally, we can extend the analysis to include materials with non-zero conductivity - for example, metals. All that is required is to work through Equations 2.3-1 to 2.3-13 again, assuming to start with that \( \text{curl}(\mathbf{H}) = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \), rather than simply \( \frac{\partial \mathbf{D}}{\partial t} \). If this is done, a slightly revised time-independent wave equation can be obtained for monochromatic waves:

\[
\nabla^2 \mathbf{E} + \left\{ \frac{\omega^2 \mu_0}{\varepsilon} - j \frac{\omega \mu_0}{\sigma} \right\} \mathbf{E} = 0
\]

Note that this has exactly the same form as Equation 2.3-13, but the term \( \omega^2 \mu_0 \varepsilon \) has been replaced here by \( \omega^2 \mu_0 \varepsilon - j \omega \mu_0 \sigma \). Since the former was previously interpreted in terms of the propagation constant by putting \( \omega^2 \mu_0 \varepsilon = k^2 \), it is reasonable to write in this case:

\[
k^2 = \omega^2 \mu_0 \varepsilon - j \omega \mu_0 \sigma = \omega^2 \mu_0 (\varepsilon - j \sigma/\omega)
\]

If we wish, we can interpret all the terms inside the bracket in Equation 2.4-37 as a modified dielectric constant, given by:

\[
\varepsilon = \varepsilon_\infty - j \frac{\sigma}{\omega}
\]

where \( \varepsilon_\infty \) is the value as \( \omega \) tends to infinity. If this is done, we see that the effective dielectric constant is complex once again. However, we cannot immediately equate the real and imaginary parts of \( \varepsilon \) to \( \varepsilon_\infty \) and \( -\sigma/\omega \), since this implies an assumption that the conductivity \( \sigma \) is real. While this is justified for low values of \( \omega \), it is certainly not the case at optical frequencies. We shall see why this should be so in Chapter 3.

### 2.5 POWER FLOW

It is clearly important to form an expression for power flow in terms of \( \mathbf{E} \) and \( \mathbf{H} \), since this is something we can measure. This is done by considering the work done on charges in a volume, as a combination of a change in stored energy and a flow of power into or out of the volume. We shall consider initially time-dependent fields.
The rate of work done by the fields per unit volume is $\mathbf{J} \cdot \mathbf{E}$, by analogy with $W = I \mathbf{V}$ in electrical circuits. From Equation 4 in 2.2-14 we get:

$$\text{curl}(\mathbf{H}) - \partial \mathbf{D}/\partial t = \mathbf{J} \quad 2.5-1$$

We can then write:

$$\mathbf{J} \cdot \mathbf{E} = \mathbf{E} \cdot \text{curl}(\mathbf{H}) - \mathbf{E} \cdot \partial \mathbf{D}/\partial t \quad 2.5-2$$

We now introduce a further vector identity, again described more fully in mathematics texts:

$$\text{div}(\mathbf{E} \times \mathbf{H}) = \mathbf{H} \cdot \text{curl}(\mathbf{E}) - \mathbf{E} \cdot \text{curl}(\mathbf{H}) \quad 2.5-3$$

We also need another relation, from Equation 3 in 2.2-14:

$$\text{curl}(\mathbf{E}) = -\partial \mathbf{B}/\partial t \quad 2.5-4$$

Combining Equations 2.5-2 to 2.5-4, we get:

$$\mathbf{J} \cdot \mathbf{E} = -\mathbf{H} \cdot \partial \mathbf{B}/\partial t - \mathbf{E} \cdot \partial \mathbf{D}/\partial t - \text{div}(\mathbf{E} \times \mathbf{H}) \quad 2.5-5$$

We can now rewrite this in the following standard form:

$$\partial U/\partial t + \nabla \cdot \mathbf{S} + \mathbf{J} \cdot \mathbf{E} = 0 \quad 2.5-6$$

Where:

$$\partial U/\partial t = [\mathbf{E} \cdot \partial \mathbf{D}/\partial t + \mathbf{H} \cdot \partial \mathbf{B}/\partial t] \quad 2.5-7$$

And:

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \quad 2.5-8$$

Equation 2.5-6 (known as Poynting's Theorem, after John Poynting, 1852-1914) is effectively a power conservation relation, since it relates the rate of change of stored energy to the outward energy flow and the energy dissipated. $U$ is the density of the energy stored in the electromagnetic fields (measured in J/m$^3$). This can be divided conveniently into two components: $U_E = 1/2 \mathbf{E} \cdot \mathbf{D}$ represents the electrically-stored energy, while $U_M = 1/2 \mathbf{H} \cdot \mathbf{B}$ is the energy in the magnetic field. Similarly the vector $\mathbf{S}$, known as the Poynting vector, describes the power flow (measured in W/m$^2$). However, it should be noted that $\mathbf{S}$ is an extremely fast-varying function. In fact, $\mathbf{S}$ contains components at $2\omega$ (or about $10^{10}$ Hz for optical waves) which are clearly not measurable by any practical technique. It is therefore convenient to define an alternative quantity related to power, that is directly measurable.

**IRRADIANCE**

For oscillating fields, we have been using the alternative notation $\mathbf{E}(x, y, z, t) = \mathbf{E}(x, y, z) \exp(j\omega t)$, where the real part is implied. In this case, we can calculate an associated time-averaged Poynting vector or irradiance $\mathbf{S}$. This is the mean of $\mathbf{S}$, over many oscillations, defined as:
where $T$ is large compared with the period of the oscillations. Substituting harmonic solutions for $\mathbf{E}$ and $\mathbf{H}$ into Equation 2.5-9, we obtain:

$$S = \frac{1}{T} \int_{T} \Re \{ E \exp(j\omega t) \} \times \Re \{ H \exp(j\omega t) \} \, dt$$

It is then simple to show that the irradiance is given by:

$$S = \frac{1}{2} \Re \left[ \mathbf{E} \times \mathbf{H}^* \right]$$

The expression for $S$ therefore contains no time variation, as required.

**TIME-AVERAGED POWER**

Irradiance is still not a directly measurable quantity. However, we can measure the time-averaged power $P$ flowing through a given surface. This is found as the integral of the normal component of $S$ over the surface, as shown in Figure 2.5-1. Using Equation 2.5-11, $P$ may be evaluated as:

$$P = \frac{1}{2} \Re \left\{ \iint_{A} \left[ \mathbf{E} \times \mathbf{H}^* \right] \cdot d\mathbf{a} \right\}$$

Time-averaged power is extremely important, since it is one of the few parameters of a high-frequency electromagnetic field that can actually be measured. Suitable detectors are the human eye, solar cells and semiconductor p-n junction photodiodes.

**DESIGN EXAMPLE**

As an example, we shall calculate the power carried by a plane wave travelling in the $+z$-direction. Assuming that components of both polarizations are present, the electric field can be written as:

$$\mathbf{E} = [i E_{x+} + j E_{y+}] \exp(-jkw)$$

The corresponding magnetic field can be found from Equations 2.4-17 and 2.4-18 as:

$$\mathbf{H} = [E_{y+} (-k/\mu_0 \omega) i + E_{x+} (k/\mu_0 \omega) j] \exp(-jkw)$$

The time-averaged power flow in the $z$-direction, per unit area, is therefore:
\[ P = \frac{1}{2} \text{Re} \left[ \mathbf{E} \times \mathbf{H}^* \right] \cdot \mathbf{k} \]
\[= \frac{1}{2} \text{Re} \left[ j \left( \mathbf{E}_y \mathbf{H}_x^* - \mathbf{E}_x \mathbf{H}_y^* \right) + j \left( \mathbf{E}_x \mathbf{H}_y^* - \mathbf{E}_y \mathbf{H}_x^* \right) + k \left( \mathbf{E}_y \mathbf{H}_z^* - \mathbf{E}_z \mathbf{H}_y^* \right) \right] \cdot \mathbf{k} \]

which in this case reduces to:
\[ P = \frac{1}{2} \text{Re} \left[ \mathbf{E}_x \mathbf{H}_y^* - \mathbf{E}_y \mathbf{H}_x^* \right] \]

Substituting in the necessary values, we get:
\[ P = \left( \frac{1}{2} \right) \sqrt{\left( \varepsilon \varepsilon_0 / \mu \right) \left( E_x^2 + E_y^2 \right)} \]

This expression allows us to relate electric field strengths in V/m to power density in W/m². Now, each of the two terms above clearly corresponds to one polarization component. Writing \( E^2 = E_x^2 + E_y^2 \), where \( E \) is the amplitude of the combined electric field, and using the definitions of impedance given in Equations 2.4-26 and 2.4-27 we obtain:
\[ P = \frac{1}{2} E^2 / Z \]
\[ = \frac{1}{2} n E^2 / Z_0 \]

The power carried by a plane wave is therefore proportional to the product of \( E^2 \) and the refractive index of the medium.

### 2.6 THE PROPAGATION OF GENERAL TIME-VARYING SIGNALS

So far, we have concentrated on the behaviour of single-frequency electromagnetic fields, which in optics correspond to monochromatic light. Throughout this book, however, we will be interested in the use of optical devices in an engineering environment. One of the most obvious applications is a communications system, which might loosely be defined as an arrangement for the transmission of information between different points. Unfortunately, a perfectly monochromatic wave (which, strictly speaking, must exist for all time without changing its frequency or amplitude) cannot carry any information. Only the modulation of such a wave - for example, by switching it on and off - can do so.

In general, therefore, we will be interested in the transmission of time-varying signals. A suitable framework for their analysis is provided by Fourier transform theory, which states that any signal \( f(t) \) may be decomposed into an infinite sum of single-frequency terms. Conventionally, this relationship is written as an integral transformation, of the form:
\[ f(t) = \left( \frac{1}{2\pi} \right) \int_{-\infty}^{\infty} F(\omega) \exp(j\omega t) \, d\omega \]

where \( F(\omega) \) represents the amplitude of the component at angular frequency \( \omega \), which itself may be found from the signal using the inverse transform:
\[ F(\omega) = \int_{-\infty}^{\infty} f(t) \exp(-j\omega t) \, dt \]

### DESIGN EXAMPLE

As an example, we may compute the frequency spectrum of a signal consisting of a short burst of a monochromatic carrier, of unity amplitude and angular frequency \( \omega_c \), as in Figure 2.6-1.
Assuming that the duration of the burst is $\Delta T$, Equation 2.6-2 reduces to:

$$F(\omega) = \frac{1}{\Delta T/2} \int_{-\Delta T/2}^{\Delta T/2} \exp(j \omega c t) \exp(-j \omega t) \, dt$$

Evaluation of the integral then gives:

$$F(\omega') = \Delta T \text{sinc}(\omega' \Delta T/2)$$

where $\omega' = \omega_c - \omega$ and $\text{sinc}(x) = \sin(x) / x$. Figure 2.6-2 shows a plot of the frequency spectrum; this peaks at $\omega = \omega_c$ and decays away on either side of this point with a typical filter envelope.

We may obtain an estimate of the width of this frequency spectrum by noting that the first zeros in Equation 2.6-4 are reached when $\omega' \Delta T/2 = \pi$. This allows the definition of an approximate signal bandwidth $\Delta \omega$ (which is then about half the width of the main lobe) in the form:

$$\Delta \omega \Delta T = 2 \pi$$

The signal bandwidth is therefore inversely proportional to the duration of the burst. This implies that high bit-rate data transmission will involve large bandwidths.

DISPERSION

The Fourier relations above may be used to analyse communication channels, as follows. Given a specified input $f(t)$, Equation 2.6-2 may be used to identify the frequency components of the signal and their corresponding amplitudes $F(\omega)$. These components may
then mentally be passed through the channel in turn; naturally, a component of angular frequency $\omega$ will propagate as a travelling wave of the same angular frequency. On arrival at the far end of the channel, the amplitudes of the components may well have changed, so that an amplitude $F(\omega)$ might be received as $F'(\omega)$. However, the total received signal $f'(t)$ may be reconstructed from these modified constituents by using a simple adaptation of Equation 2.6-1, namely:

$$f'(t) = (1/2\pi) \int_\infty^{-\infty} F'(\omega) \exp(j\omega t) \, d\omega$$

By comparing the received signal $f'(t)$ with the transmitted signal $f(t)$, the effect of the channel may be assessed. For example, it will be important to know in advance what type of signals may be passed through the channel and still arrive in recognisable form. Before this can be done, however, we must identify the signal distortions that are possible. By and large, there are just two. Firstly, the relative amplitudes of the frequency components may alter. This could occur in a channel with frequency-dependent attenuation. Secondly, the relative phases of the components might change, if the phase velocity of the channel is frequency-dependent. Generally, the latter effect is the most significant. It is known as dispersion, and we will now consider some of its features.

We start by returning to Section 2.4, where the phase velocity of a wave of angular frequency $\omega$ travelling in a homogeneous medium was defined as $v_{ph} = c/n$. From this, we may infer that the medium will be dispersive if there is any dependence of the refractive index on frequency. In fact, this is the case in all matter. The result is therefore normally described as material dispersion (to distinguish it from other effects that occur in more complicated transmission channels, particularly waveguides). We will examine the underlying reasons for this dependence in Chapter 3. However, we also obtained the alternative definition $v_{ph} = \omega/k$. Consequently, in a dispersive medium we would expect a more complicated relation between $\omega$ and $k$ than just $\omega/k = \text{constant}$.

One of the simplest examples of a dispersive medium is provided by an ionised medium (such as the ionosphere), for which it can be shown that:

$$\omega = \sqrt{\omega_p^2 + c^2k^2}$$

where $\omega_p$ is a constant, the plasma frequency, which will be introduced properly in Chapter 3. This type of relation can be represented as a plot of $\omega$ against $k$ called a dispersion diagram, as shown in Figure 2.6-3. In this case, the diagram shows that $\omega$ tends to $\omega_p$ for small values of $k$, while for large $k$, $\omega$ tends to the dashed line $\omega = ck$. 

![Dispersion Diagram](image)
The phase velocity $\omega/k$ may then be found either from the dispersion diagram, or directly from Equation 2.6-7. The latter process yields:

$$v_{ph} = \sqrt{\left(\frac{\omega_p^2}{k^2} + c^2\right)}$$  \hspace{1cm} 2.6-8

From this, it can be seen that the phase velocity is not constant; it tends to infinity as $k$ tends to zero, and to $c$ as $k$ becomes large. To assess the effect of this variation, we shall consider the propagation of an elementary compound signal, consisting of components at just two distinct angular frequencies $\omega + d\omega$ and $\omega - d\omega$, where $d\omega$ is small. For simplicity, we take the amplitudes of the two waves to be the same. However, we assume that the phase velocities at the two frequencies are unequal, so that the corresponding propagation constants must be written as $k + dk$ and $k - dk$. Assuming $y$-polarization and $z$-propagation, the electric field of the signal might then be written:

$$E_y = E_{y_0} \cdot \left[ \exp\{j((\omega + d\omega)t - (k + dk)z)\} + \exp\{j((\omega - d\omega)t - (k - dk)z)\} \right]$$  \hspace{1cm} 2.6-9

Figure 2.6-4 shows a snapshot of the combined field at $t = 0$. Note that the two waves sum together to give what amounts to a carrier of constant period, modulated by an envelope (shown dashed).

We may illustrate this mathematically, by combining the two components slightly differently. If the common factor $\exp\{j(\omega t - kz)\}$ is taken out, we get:

$$E_y = 2E_{y_0} \cdot \exp\{j(\omega t - kz)\} \cdot \cos\{d\omega t - dk z\}$$  \hspace{1cm} 2.6-10

This suggests that we may view propagating signals in two different ways. We may either regard them as a sum of a number of separate travelling waves (as in Equation 2.6-9) or as a single modulated wave (Equation 2.6-10). However, the latter viewpoint shows clearly that the information-carrying component of the signal - the modulation envelope - is also propagating as a travelling wave, defined by the term $\cos(d\omega t - dk z)$. This envelope must therefore also have a velocity of propagation, which in general is distinct from that of the carrier wave. Since it refers to a group of waves, rather than a single wave, it is known as the group velocity $v_g$, and is defined as:

$$v_g = \frac{d\omega}{dk}$$  \hspace{1cm} 2.6-11

For the ionosphere, we may find the group velocity by differentiating Equation 2.6-7, to get:
\[ v_g = \frac{c^2}{\sqrt{\omega_p^2/k^2 + c^2}} \]  

This is clearly different from the expression derived earlier for the phase velocity. However, for large \( k \) (and thus very high frequencies), \( v_g \) tends to \( v_{ph} \).

The analysis above can be used to find the velocity of information carried by groups of waves of similar frequency. However, for groups comprising a wider range of frequencies, \( v_g \) may not be considered constant, so different groups of a signal will travel at different speeds. This can result in a damaging effect, known as **pulse broadening**, which limits the rate of data transmission. We can illustrate this by considering the problem of transmitting data over a distance \( L \) via successive bursts of a high-frequency carrier. Each ‘one’ in the data stream corresponds to a burst of duration \( \Delta T \), and the separation between successive bursts is \( T \). Figure 2.6-5a shows the modulation envelope of a typical section of a message.

![Figure 2.6-5](image)

Figure 2.6-5  A short section of a message, a) as sent and b) after travelling some distance in a dispersive medium.

From our earlier example, we may estimate the required bandwidth to be \( \Delta \omega \), so that the frequencies comprising the signal range approximately from \( \omega_1 = \omega_c - \Delta \omega/2 \) to \( \omega_2 = \omega_c + \Delta \omega/2 \). At these extremes, the group velocity may have different values - say, \( v_{g1} \) and \( v_{g2} \). Consequently, different constituents of the signal must arrive at the far end of the channel at times ranging from \( t_1 = L/v_{g1} \) to \( t_2 = L/v_{g2} \). Assuming that \( t_2 > t_1 \), this spread in arrival times will amount to a lengthening of the information-carrying envelope from \( \Delta T \) to \( \Delta T + \Delta t \), where \( \Delta t = (t_2 - t_1) \). If \( L \) is large enough, we may even find that \( \Delta t \approx T \). Information from successive bursts will then become mixed (as is beginning to happen in Figure 2.6-5b) so that the message is beyond recovery by simple means.

Now, the effective broadening of the signal may be calculated by writing:

\[ \Delta t = L\{1/v_{g2} - 1/v_{g1}\} = L \frac{d(1/v_g)}{d\omega} \Delta \omega \]

Since \( 1/v_g = \frac{dk}{d\omega} \), this yields:

\[ \Delta t = L \Delta \omega \frac{d^2 k}{d\omega^2} \]

Clearly, pulse broadening may be minimised by ensuring that the term \( \frac{d^2 k}{d\omega^2} \) is as small as possible, and preferably zero. We may see how this relates to material parameters by noting that \( k = n\omega/c \). Consequently, for a dispersive medium we can write:

\[ \frac{dk}{d\omega} = \frac{1}{c} \{n + \omega \frac{dn}{d\omega}\} \]
Similarly, since $\lambda_0 = \frac{2\pi c}{\omega}$, we may put:

$$\frac{dn}{d\omega} = \frac{dn}{d\lambda_0} \left( \frac{d\lambda_0}{d\omega} \right) = -\frac{2\pi c}{\omega} \frac{dn}{d\lambda_0}$$

so that:

$$\frac{dk}{d\omega} = \frac{1}{c} \left\{ n - \lambda_0 \frac{dn}{d\lambda_0} \right\}$$

Differentiating a second time, we get:

$$\frac{d^2k}{d\omega^2} = \left\{ \frac{d(dk/d\omega)}{d\lambda_0} \right\} \left( \frac{d\lambda_0}{d\omega} \right) = -\frac{\lambda_0}{c} \left( \frac{d^2n}{d\lambda_0^2} \right) \left( \frac{d\lambda_0}{d\omega} \right)$$

Finally, if we put $\frac{d\lambda_0}{d\omega} \approx \Delta\lambda_0 / \Delta\omega$, where $\Delta\lambda_0$ is the wavelength range of the signal, we can obtain a variant of Equation 2.6-14 that emphasises the contribution of the medium to pulse broadening:

$$\Delta t = -\frac{(L\lambda_0 \Delta\lambda_0 / c)}{d^2n/d\lambda_0^2}$$

This suggests that dispersion can be minimised by operating near a wavelength where $d^2n/d\lambda_0^2 = 0$. In silica (the main constituent of most optical fibres) this occurs at $\lambda_0 \approx 1.27 \mu m$. Consequently, optical communications systems almost always operate at near infrared wavelengths.
CHAPTER TWO

PROBLEMS

2.1. Starting from the differential form of Maxwell’s equations, derive a time-dependent vector wave equation, valid for a uniform dielectric medium and containing only the magnetic field $\mathbf{H}$.

2.2. A plane electromagnetic wave oscillates at an angular frequency of $2.978 \times 10^{15}$ rad/sec. Find its wavelength in (a) free space, and (b) a material of relative dielectric constant 2.25. 
   [(a) $0.633 \, \mu\text{m}$; (b) $0.422 \, \mu\text{m}$]

2.3. A plane electromagnetic wave is defined by $\mathbf{E} = E_{y0} \exp\{-jk_0(ax + bz)\} \mathbf{j}$, where $k_0$ is the propagation constant of free space, and the constants $a$ and $b$ are given by $a = 1$, $b = \sqrt{3}$. Find (a) the wave amplitude, (b) the direction of polarization, (c) the direction of travel, and (d) the refractive index of the medium. 
   [(c) $30^\circ$ to the z-axis, in the x - z plane; (d) 2]

2.4. The expression $E_y = E_{y0} \exp(-jkz)$ is a solution to the scalar equation $\nabla^2 E_y + \omega^2 \mu_0 \varepsilon E_y = 0$ that represents a plane wave travelling in the z-direction. What is $k$? Show that the inhomogeneous wave $E_y = E_{y0} \exp(\gamma x) \exp(-j\beta z)$ is also a solution. What relation must be satisfied by $\gamma$ and $\beta$?

2.5. Sketch the loci of (a) constant phase, and (b) constant amplitude for the inhomogeneous wave in Question 2.4. What direction is it travelling in? Does it travel faster or slower than the plane wave? What do you think it represents?

2.6. A optical wave of wavelength $\lambda_0 = 0.633 \, \mu\text{m}$ is travelling in a material of complex relative dielectric constant $\varepsilon_r = 2.25 - j10^{-8}$. What are the values of the propagation constant and the absorption coefficient? Find the distance the wave must travel in the medium before the power it carries decays to 1/e of its initial value. 
   [14.89 x 10^6 m^-1; 0.0331 m^-1; 15.11 m]

2.7. Show that the irradiance $S = \frac{1}{2} \text{Re} \{\mathbf{E} \times \mathbf{H}^*\}$ does indeed represent the time average of the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, for harmonically varying fields.

2.8. Determine the z-component of the irradiance, for the wave in Question 2.3. What is the direction of maximum irradiance?

2.9. The time-averaged power density of a linearly polarized plane wave in free space is 100 mW/cm^2. Find the peak amplitude of the electric field. 
   [868 V/m]

2.10. The relationship between $\omega$ and $k$ for electromagnetic waves in the ionosphere may be shown to have the form $\omega = \sqrt{[\omega_p^2 + c^2 k^2]}$, where $c$ is the velocity of light and $\omega_p$ is the plasma frequency. Sketch the variation of the phase and group velocity with $k$. Find the value of $v_g$ when $\omega \rightarrow \omega_p$. What is the significance of this result? Find a solution for $k$ when $\omega < \omega_p$. What form of wave do you think that this solution describes?
SUGGESTIONS FOR FURTHER READING

Hecht E., "Optics", 2nd Edn., Addison-Wesley Publishing Co., Reading MA (1987); Chapters 1, 2, 3, & 8.
3.1 INTRODUCTION

In Chapter 2, we described the basic characteristics of plane electromagnetic waves by making use of the parameters $\varepsilon$, $\sigma$ and $\mu$. Although these can account phenomenologically for the properties of the medium in which the wave is propagating, they treat it as if it were totally uniform and continuous. Consequently, they provide no help in visualising the detailed interaction of the wave and the medium, and even the simplest model will allow an improvement on this rather superficial concept. We will now add such a model, basing the discussion again on a classical picture.

Materials may obviously be classified in several different ways. On the basis of their phase, we may describe them as solids, liquids or gases. Here, we will naturally be interested mainly in solids. Even these can exist in a variety of forms, however, ranging from those entirely lacking in detailed structure (amorphous materials) to those with a detailed, long-range order (crystals). In between, there are media which have local structure, known as polycrystalline materials. Furthermore, their electronic properties may be widely different. At one extreme, there are insulators (which do not conduct electricity), and at the other there are conductors (which clearly do). Somewhere in the middle, we find the semiconductors, which have the most complicated properties of all.

Faced with all this, we might suppose that a quite separate and complex model would be required for each class of solid. For simplicity, therefore, we will begin by ignoring any detailed structure, and concentrate entirely on the electronic aspects. Our argument runs like this. Since all materials are simply collections of atoms and molecules, they can all be described as particular distributions of positive and negative charge. A model for the optical properties of a material can then be developed by considering this charge distribution and its interaction with an electromagnetic wave. Naturally, we may improve the picture later on, by inserting structural features where relevant.

The electronic model (originally due mainly to Hendrik Lorentz, 1853-1928, and Paul Drude, 1863-1906) is based on a consideration of the forces acting on the charges inside the material. For a free charge $e$ moving with a velocity $v$ through an electric field $E$ and a magnetic flux density $B$, the force is known from low-frequency electromagnetic experiments to be:

$$ F = e \left[ E + v \times B \right] $$

3.1-1

However, all charge velocities will be small in the discussions that follow, so we may safely neglect any forces due to magnetic flux, concentrating exclusively on the interaction of the electric field with the charges. This view is reinforced by the fact that the permeability of most optically significant materials is essentially $\mu_0$, the value for free space. We can use this approach to explain the origin of the dielectric constant in insulators and the conductivity of metals; the properties of semiconductors will be treated in a different way (using the language of solid state theory) in Chapter 11.

3.2 THE ORIGIN OF THE DIELECTRIC CONSTANT

We will start by considering an insulating, transparent, dielectric solid, whose optical properties are determined mainly by the behaviour of the outermost, bound electrons.
orbiting each atom. Now, in Chapter 2 we showed that the electric field at any point will oscillate sinusoidally for a harmonic wave. From Equation 3.1-1, this field must impose forces in opposite directions on any positive and negative charges located there. Consequently, we would expect the field to separate the charges in a periodic manner, producing a local distribution of dipoles. Figure 3.2-1 shows this separation for a static field.

![Figure 3.2-1 Separation of charges by an electric field.](image)

In addition, we might expect that as the charges separate, the local charge imbalance will generate another force, which would tend to restore charge neutrality. Finally, the oscillating charges will act as microscopic radiators, giving rise to energy loss. This will also cause some disturbance to the neighbouring atoms of the structure. We can model the whole as a damping effect, which describes the transfer of energy from our particular charges into the whole structure of the solid.

Combining these effects together enables us to write an expression for the equation of motion of a typical charge. This will be one of the bound electrons - the positive charges, being considerably more massive (and anyhow locked into the structure, since it is solid) would tend not to move significantly at optical frequencies. For simplicity, we will also ignore the vectorial nature of the field, and consider motion in the x-direction only. The components of force are therefore:

1. **THE DRIVING FORCE DUE TO THE HARMONIC FIELD**

Since a harmonic electric field is given by \( E = E \exp(j\omega t) \), it might initially be thought that this force will have the form:

\[
F_e = eE = eE \exp(j\omega t)
\]

However, the local field acting on an electron will be determined not only by the driving field \( E \) but also by the separation of all the other charges in its immediate neighbourhood. The net effect of this charge separation is described as an electronic polarization of the material, and is represented by the symbol \( P \) (not to be confused with the optical polarization of the previous Chapter - it is an unfortunate coincidence that two such different phenomena should have the same name). Defining the displacement of a typical charge from its equilibrium position by \( x \), we can write \( P \) as:

\[
P = N\vec{x}
\]
\[ P = P \exp(j\omega t) \]  

In general, \( P \) might be a complex quantity, implying a possible phase difference between \( E \) and \( P \).

The effect of the polarization is to modify the local electric field, increasing its value above that in free space (to, say, \( E_{\text{loc}} \)). For a non-polar material (i.e., one with no permanent dipoles in the absence of a driving field), Lorentz showed that the increase can be found very simply as:

\[ E_{\text{loc}} - E = P/3\varepsilon_0 \]

Consequently, the force acting on an electron is not as given in Equation 3.2-1, but is instead:

\[ F_e = eE_{\text{loc}} = e(E + P/3\varepsilon_0) \]

Typical non-polar materials include Si, Ge, C (in diamond form) and liquid hydrogen. Polar materials, which do have permanent dipoles, include NaCl, TiO\textsubscript{2} and SiO\textsubscript{2}; in addition, there is also a class of weakly polar materials.

2. THE RESTORING FORCE DUE TO THE CHARGE SEPARATION

Since dipoles have been produced by the driving field, we might expect a restoring force due to attraction between the separated charges. Modelling the system as a spring with extension \( x \), this force would have the form:

\[ F_r = -sx \]

Here \( s \) is a constant of proportionality, which depends on the exact nature of the material. For example, materials with a highly polarizable structure will have high values of \( s \). Now, the linear dependence of \( F_r \) on \( x \) might be regarded as an unjustifiable assumption (indeed, we might have expected an inverse square law dependence, since this is the norm for two point charges). However, there are many interacting charges, and it is the superposition of large numbers of inverse square forces that produce this linear effect. Furthermore, we can argue that for the moment we will treat the linear term as but the first (and hopefully dominant) term in a general power series. We will consider the significance of higher order terms later on.

3. THE DAMPING FORCE

We must also include a force that accounts for energy loss. The justification for this ultimately hangs on any success of our theory in describing the observable effects. The simplest such force is the viscous damping force of a simple harmonic oscillator, which has the form:

\[ F_d = -r\, \text{d}x/\text{d}t \]

The sum of these forces must yield the product of the mass \( m \) of our electron (9.1 x 10\textsuperscript{-31} kg) and its acceleration, so combining all the terms we obtain the following typical equation of motion:

\[ m \frac{d^2x}{dt^2} = -sx - r \, \text{d}x/\text{d}t + e(E + P/3\varepsilon_0) \]
Eliminating the displacement $x$ using Equation 3.2-2, we then obtain:

$$m \frac{d^2 P}{dt^2} + r \frac{dP}{dt} + sP = Ne^2 (E + P/3\varepsilon_0)$$

Re-arranging Equation 3.2-9 slightly then gives:

$$\frac{d^2 P}{dt^2} + \gamma \frac{dP}{dt} + (\omega_0^2 - Ne^2/3m\varepsilon_0)P = Ne^2E/m$$

where we have written $s/m = \omega_0^2$, and $r/m = \gamma$.

Now, both $E$ and $P$ vary harmonically. Performing the necessary differentiation, we can therefore eliminate the time-dependence to get:

$$-\omega^2 P + j\omega\gamma P + (\omega_0^2 - Ne^2/3m\varepsilon_0)P = Ne^2E/m$$

and hence find $P$ as:

$$P = E \frac{Ne^2/m}{[(\omega_0^2 - \omega^2) - Ne^2/3m\varepsilon_0] + j\omega\gamma}$$

We have now found a relationship between the polarization and the driving field. For simplicity, we can combine the terms $\omega_0^2$ and $Ne^2/3m\varepsilon_0$. Defining the **plasma frequency** $\omega_p$ using:

$$\omega_p^2 = Ne^2/m\varepsilon_0$$

we obtain the slightly modified resonant frequency $\omega_m$, given by:

$$\omega_m^2 = \omega_0^2 - \omega_p^2/3$$

so that Equation 3.2-12 becomes:

$$P = E \varepsilon_0 \omega_p^2 / [(\omega_m^2 - \omega^2) + j\omega\gamma]$$

Equation 3.2-15 shows that the dipoles will oscillate in response to the driving field $E$, but generally neither the polarization nor the displacement are in phase with $E$. We note that the magnitude of the polarization depends on the relation of the driving frequency $\omega$ to $\omega_m$. Close to resonance, when $\omega = \omega_m$, $P$ will reach a maximum, whose magnitude is limited only by the damping term $j\omega\gamma$.

We now need to find out how the polarization $P$ relates to the permittivity $\varepsilon$. We can do this as follows. From Gauss' law, we know that:

$$\iiint \mathbf{D} \cdot d\mathbf{a} = \iiint \rho \, dv$$

This implies (among other things) that the surface charge density on the plates of a parallel-plate capacitor must be $\mathbf{D}$. If the plates are spaced by vacuum, we have of course that $\mathbf{D} = \varepsilon_0 \mathbf{E}$. However, the introduction of a dielectric spacer must lead to a change in the surface charge density, since the dielectric material is polarized by the field inside the capacitor. This leads to a motion of charges towards the plates. The quantity of charge moved per unit area is $Ne$, so in general we have:
A dielectric spacer therefore improves the charge storage capability of a capacitor, a discovery originally made by Faraday. Since \( P \) is proportional to \( E \), then writing \( P = \varepsilon_0 \chi E \), where \( \chi \) is the dielectric susceptibility, leads to:

\[
D = \varepsilon_0 (1 + \chi) E
\]

Since in Chapter 2 we defined \( D = \varepsilon_\varepsilon E \), it is clear that the relative dielectric constant must be:

\[
\varepsilon_\varepsilon = 1 + \chi = 1 + P/\varepsilon_0 E
\]

Using Equation 3.2-14, we then obtain:

\[
\varepsilon_\varepsilon = 1 + \omega_p^2 / \{(\omega_m^2 - \omega^2) + j\omega\gamma\}
\]

We have therefore found a direct connection between the dielectric constant and the displacements of dipoles. There are two interesting points to note. Firstly, we see that the expression for \( \varepsilon_\varepsilon \) is independent of the driving field, as we might expect. This has come about essentially because of our assumption that the restoring force on a general electron was linearly proportional to displacement. Secondly, the expression for \( \varepsilon_\varepsilon \) is complex. In the next section, we will discuss the implications of this, but before doing so we will briefly consider one final aspect: the relation between polarizability and the dielectric constant.

**POLARIZABILITY**

We have seen that the origin of the dielectric constant lies in the separation of charges by an electric field. We may therefore connect the polarization with the local field using:

\[
P = N\alpha E_{loc}
\]

Here \( \alpha \) is a constant, known as the polarizability. Its magnitude essentially describes the ease or otherwise with which charges may be separated by the field. Using the definition of the local field given in Equation 3.2-4 we can obtain:

\[
P = \{N\alpha / (1 - N\alpha/3\varepsilon_0)\} E
\]

However, since we have already defined \( P = \varepsilon_0 \chi E \), it is clear that the dielectric susceptibility is related to the polarizability by:

\[
\chi = (N\alpha/\varepsilon_0) / (1 - N\alpha/3\varepsilon_0)
\]

Equally, since \( \varepsilon_\varepsilon = 1 + \chi \), we must have:

\[
\varepsilon_\varepsilon = (1 + 2N\alpha/3\varepsilon_0) / (1 - N\alpha/3\varepsilon_0)
\]

Equation 3.2-24 shows that, in general, materials with high polarizability will have a large dielectric constant, and vice versa. Often, the equation is inverted to give \( \alpha \) in terms of \( \varepsilon_\varepsilon \) as:

\[
\alpha = (3\varepsilon_0/N) (\varepsilon_\varepsilon - 1) / (\varepsilon_\varepsilon + 2)
\]
This is known as the Lorentz-Lorenz equation. It is an interesting relation, because it enables the calculation of a microscopic quantity (α) from a macroscopic observable (ε_r).

### 3.3 MATERIAL DISPERSION

The dependence of the dielectric constant on frequency is known as **material dispersion**. We shall now examine the consequence of this feature, and of the complex nature of ε_r.

In the previous chapter, we assumed that a lossy dielectric could be described by a complex relative permittivity of the form ε_r = ε'_r - jε''_r. Multiplying the top and bottom of Equation 3.2-20 by \((ω^2_m - ω^2) - jωγ\), we obtain the following expressions for the real and imaginary parts of ε_r:

\[
ε'_r = 1 + \frac{ω_p^2 (ω^2_m - ω^2)}{((ω^2_m - ω^2)^2 + (ωγ)^2)}
\]
\[
ε''_r = \frac{ω_p^2 ωγ}{((ω^2_m - ω^2)^2 + (ωγ)^2)}
\]

If the damping factor γ is negligibly small, then these may be approximated by:

\[
ε'_r \approx 1 + \frac{ω_p^2}{(ω^2_m - ω^2)}
\]
\[
ε''_r \approx \frac{ω_p^2 ωγ}{(ω^2_m - ω^2)^2}
\]

From this, we see that the real part ε'_r will be approximately unity, unless ω is close to the resonant frequency ω_m. Exactly at ω = ω_m, it is discontinuous. Similarly, the imaginary part ε''_r is small except near ω = ω_m, where it becomes infinite. This behaviour is plotted in Figure 3.3-1. Note that it is quite possible for ε'_r to become negative, just beyond the resonant frequency.

Also shown in Figure 3.3-1 are the predictions of the full expression 3.3-1, which does not assume that γ = 0. The introduction of damping introduces two new features. The first is that while ε'_r normally increases with ω (this behaviour is called **normal dispersion**) there is now a narrow range of frequency near ω_m where ε'_r decreases with ω. This is known as **abnormal dispersion**. The second is that ε''_r now has a finite maximum value. The overall shape of this so-called 'absorption line' is now a function that occurs often in physics, a **Lorentzian lineshape**.

For optical propagation it is often convenient to deal with the refractive index instead of ε_r. Since n = √ε_r, n must also be complex. Writing n = n' - jn'', we have:

\[
n^2 = (n^2 - n''^2) - 2jn'n'' = ε'_r - j ε''_r
\]

Making the approximation that n'' is small, we may put n^2 ≈ ε'_r and 2n'n'' ≈ ε''_r. Assuming that Equation 3.3-2 is also valid, we can then write:

\[
n^2 ≈ 1 + \frac{ω_p^2}{(ω^2_m - ω^2)}
\]

and so on. Hence, it is clear that the refractive index must depend on frequency.
Figure 3.3-1 Typical variation of the complex dielectric constant with frequency.

Now, the situation in real materials is normally more complicated, because there may be more than one resonance involved. For example, there are often resonances that correspond to excitation of vibrational oscillation modes of molecules, in addition to the electronic resonance considered so far. The latter usually lie at the ultraviolet end of the optical spectrum, while the former appear at the infrared end. However, by slightly modifying Equation 3.3-4, it is simple to accommodate these extra features in a phenomenological way. All we need to do is to identify the appropriate resonant frequencies, together with the density of oscillators involved, and incorporate these terms as a summation in the expression for refractive index. If this is done, we get:

\[
n' \approx 1 + \omega_p^2 \sum_j f_j / (\omega_m^2 - \omega^2)
\]

3.3-5

Here we have replaced the plasma frequency \(\omega_p^2\) by the more general term \(\omega_p^2 f_j\), where \(f_j\) is the oscillator strength (an indication of the density of electrons involved), and the resonant frequency \(\omega_m\) by \(\omega_m\). Substituting \(2\pi c/\lambda_0\) for \(\omega\), and so on, Equation 3.3-5 may then be transformed into its more common incarnation, the Sellmeier equation:

\[
n' \approx 1 + \sum_j A_j \lambda_0^2 / (\lambda_0^2 - \lambda_j^2)
\]

3.3-6

Values of \(A_j\) and \(\lambda_j\) are often fitted to dispersion data, to provide an estimate of refractive index over a wide spectral range.

DESIGN EXAMPLE

Table 3.1-1 shows values of \(A_j\) and \(\lambda_j\) for silica, following I.H. Maletson (JOSA 55, 1205 (1965)). Note that only three resonances need to be considered. Two of these (at 0.068 \(\mu\text{m}\)}
and 0.116 µm wavelength) lie in the ultraviolet, while the third (at 9.9 µm) lies in the infrared. The corresponding dispersion curve is shown in Figure 3.3-2. In this case, experimental values are indistinguishable from the theoretical curve, which provides convincing support for the rather heuristic theory we have used here. Note that there is an inflexion in the curve at around $\lambda_0 = 1.27$ µm, so that $d^2n/d\lambda_0^2$ will be zero at this point. According to the discussion at the end of Chapter 2, this will prove a suitable operating point for minimum signal dispersion.

<table>
<thead>
<tr>
<th>i</th>
<th>$\lambda_1$</th>
<th>$\lambda_1$ (µm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6961663</td>
<td>0.0684043</td>
</tr>
<tr>
<td>2</td>
<td>0.4079426</td>
<td>0.1162414</td>
</tr>
<tr>
<td>3</td>
<td>0.8974794</td>
<td>9.8961610</td>
</tr>
</tbody>
</table>

Table 3.1-1  Sellmeier coefficients for silica

![Dispersion curve for silica.](image)

VALIDITY OF THE ‘CLASSICAL’ MODEL

Figure 3.3-1 shows absorption loss increasing strongly in the vicinity of a resonance. While it adequately describes the sense of the loss, the classical treatment we have used fails completely to account for the actual transition between one bound state and another that must occur. This can only be described effectively using the ideas of quantum mechanics and solid state theory, which we will discuss in Chapter 11.

3.4 THE ORIGIN OF ELECTRICAL CONDUCTIVITY

We now move on to consider the origin of a different parameter, conductivity. In this case, the dominant effects are caused by charges that are free to move, rather than bound to an atomic site as in a dielectric. For simplicity, we shall restrict ourselves to the case of a single type of charge carrier, free electrons in a metal. An alternative title for this topic might therefore be ‘the optics of metals’.
We begin by considering the response of the free electrons to the oscillating field of a propagating wave. The major difference in the behaviour of these electrons (compared with bound electrons) is that as a free electron moves under the action of the field, another one moves into the space vacated. Consequently, there are no unbalanced charges, and there can be no restoring electrostatic forces. Similarly, there is little in the way of local polarization. However, we can still expect damping, due to collisions of the electrons among themselves and with the lattice. The controlling equation of motion for each electron is therefore:

$$m \frac{d^2x}{dt^2} = -r \frac{dx}{dt} + eE$$  \hspace{1cm} (3.4-1)$$

This should be compared with Equation 3.2-8; clearly, the two are similar, but Equation 3.4-1 lacks the spring and local polarization terms.

Assuming again that $E = E \exp(j\omega t)$, we may solve Equation 3.4-1 to obtain:

$$x = eE \exp(j\omega t) / (rj\omega - m\omega^2)$$  \hspace{1cm} (3.4-2)$$

Now, motion of charges implies a flow of current. If the density of electrons is $N$, then the resulting current density is:

$$J = Nev = Ne \frac{dx}{dt}$$  \hspace{1cm} (3.4-3)$$

Differentiating Equation 3.4-2, we get:

$$J = Ne^2E \exp(j\omega t) / (r + jm\omega)$$  \hspace{1cm} (3.4-4)$$

Now, conductivity is defined as a parameter that relates current density to electric field, according to $J = \sigma E$. We may therefore divide Equation 3.4-4 by $E$ to extract the conductivity as:

$$\sigma = Ne^2 / (r + jm\omega)$$  \hspace{1cm} (3.4-5)$$

Clearly, electrical conductivity is also in general a complex quantity. At low frequencies, it is real (which corresponds with everyday experience) but at high enough frequencies its character changes quite dramatically, and it becomes purely imaginary.

At this point, we shall revert to the notation of Chapter 2, where we showed that a material with non-zero conductivity could be described by a modified dielectric constant of the form:

$$\varepsilon = \varepsilon_\infty - j\sigma/\omega$$  \hspace{1cm} (3.4-6)$$

In a metal, we have already mentioned that the contribution of bound electrons is extremely small, so $\varepsilon_\infty$ is very close to the value obtained in free space, $\varepsilon_0$. Consequently, a metal may be described by a modified relative dielectric constant:

$$\varepsilon_r = 1 - \left(\frac{Ne^2/m\varepsilon_0}{\omega (\omega - jr/m)}\right)$$  \hspace{1cm} (3.4-7)$$

Now, the term $Ne^2/m\varepsilon_0$ is clearly equal to $\omega_p^2$, where $\omega_p$ is once again the plasma frequency. Similarly, $r/m$ corresponds to the damping factor $\gamma$ of the previous sections. However, because damping in metals is essentially due to collisions, this is generally given the alternative name $\omega_c$, the **collision damping frequency**.
We can easily extract the real and imaginary parts of $\epsilon_r$ by multiplying the top and bottom of Equation 3.4-7 by $\omega + j\omega$. If this is done, we get:

$$
\epsilon_r' = 1 - \frac{\omega_p^2}{\omega^2 + \omega_c^2} \\
\epsilon_r'' = -\frac{\omega_p^2}{\omega_c} \frac{\omega_c}{\omega (\omega^2 + \omega_c^2)}
$$

Examining Equation 3.4-8, we first note the very striking result that $\epsilon_r'$ can actually be negative over a wide spectral range, when $\omega_p^2 / (\omega^2 + \omega_c^2) > 1$. For this to be possible, however, we require $\omega_p > \omega_c$. Values of $\omega_p$ and $\omega_c$ are shown in Table 3.4-1 for some common metals. Note that the ratio $\omega_p / \omega_c$ is greater than about 20 for all the members of this group.

<table>
<thead>
<tr>
<th>Metal</th>
<th>$\omega_p$</th>
<th>$\omega_c$</th>
<th>$\omega_p / \omega_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluminium</td>
<td>$19.8 \times 10^{15}$</td>
<td>$1.01 \times 10^{15}$</td>
<td>19.6</td>
</tr>
<tr>
<td>Copper</td>
<td>$10.1 \times 10^{15}$</td>
<td>$0.31 \times 10^{15}$</td>
<td>32.6</td>
</tr>
<tr>
<td>Gold</td>
<td>$9.9 \times 10^{15}$</td>
<td>$0.44 \times 10^{15}$</td>
<td>22.5</td>
</tr>
<tr>
<td>Silver</td>
<td>$12.2 \times 10^{15}$</td>
<td>$0.09 \times 10^{15}$</td>
<td>135.6</td>
</tr>
</tbody>
</table>

Table 3.4-1 Values of $\omega_p$, $\omega_c$, and $\omega_p / \omega_c$ for some common metals.

In the case of silver (Ag), which is widely used in optical surface wave devices (see Ch. 5), $\epsilon_r'$ will be negative for all angular frequencies up to $\omega \approx 1.2 \times 10^{16}$ rad/sec. Since this corresponds to a wavelength of $\lambda_0 = 150$ nm (deep in the UV), $\epsilon_r'$ will be negative through the visible and infrared spectra. This result has enormous influence on the types of electromagnetic wave that may be supported by this material. Furthermore, the exceptionally high value of $\omega_p / \omega_c$ suggests that the damping of these waves may be relatively low.

**DESIGN EXAMPLE**

We can use Equation 3.4-8 to plot the variation of $\epsilon_r'$ and $\epsilon_r''$ with wavelength for silver through the visible spectrum. The results are as shown in Figure 3.4-1.

![Figure 3.4-1](image-url)

Figure 3.4-1 Comparison between experimental and theoretical data for $\epsilon_r'$ and $\epsilon_r''$ in silver.
CHAPTER THREE

Also shown on the figure are some experimental points (after Ordal et al., Appl. Opt. 22, 1099 (1983)). The theoretical model has been matched to the data by choosing \( \omega_p \) and \( \omega_\tau \) to give exact agreement for \( \varepsilon'_r \) at \( \lambda_0 = 0.6 \mu \text{m} \). Clearly, the simple theory outlined here provides broadly the correct magnitude for \( \varepsilon'_r \), but the wavelength dependence is not quite correct. This is mainly due to the fact that our model does not completely describe the mechanisms for electron energy loss through collisions. However, the agreement is almost exact for \( \varepsilon''_r \).

Note that if damping can be neglected (as in an ionised gas, which has no lattice for the electrons to collide with), Equation 3.4-8 reduces to \( \varepsilon'_r = n^2 = 1 - \frac{\omega_p^2}{\omega^2} \). Since \( \omega = kc/n \), we may then obtain \( \omega = \sqrt{[\omega_p^2 + c^2k^2]} \) - the dispersion characteristic introduced in Chapter 2 for the ionosphere.

PLASMA CONTRIBUTION TO THE REFRACTIVE INDEX

Of course, there are materials in which the dielectric constant is determined by contributions from both bound and free electrons. The examples we have in mind are semiconductors, and though the electronic effects occurring in these media are vastly more complicated than we have encountered so far, we shall take the liberty of extending our crude model to estimate one very important quantity: the change in refractive index that follows from the introduction of significant numbers of free carriers to an otherwise poorly-conducting material.

We start by returning to Equations 3.4-5 and 3.4-6. Assuming this time that the contribution of bound electrons is not negligible, we may write:

\[
\varepsilon_r = \varepsilon_{\infty} - j\sigma/\varepsilon_0 \omega \tag{3.4-9}
\]

where \( \varepsilon_{\infty} \) is the relative dielectric constant in the absence of free electrons. This is no longer unity, but we take it for simplicity to be real. In the presence of free electrons, \( \sigma \) is non-zero, but we shall assume the frequency is sufficiently high that \( \omega \gg \omega_\tau \). \( \sigma \) may then be approximated by:

\[
\sigma \approx Ne^2/jm\omega \tag{3.4-10}
\]

In this case, Equation 3.4-9 becomes:

\[
\varepsilon_r = \varepsilon_{\infty} \left\{ 1 - \frac{Ne^2/m\varepsilon_0\varepsilon_{\infty} \omega^2}{\omega^2} \right\} \tag{3.4-11}
\]

The refractive index is clearly given by the square root of Equation 3.4-11. If we now assume that the right-hand term in the brackets is small, we may use a binomial approximation to obtain:

\[
n \approx n_{\infty} - \frac{Ne^2}{2m\varepsilon_0\varepsilon_{\infty} \omega^2} \tag{3.4-12}
\]

where we have defined \( n_{\infty} = \sqrt{\varepsilon_{\infty}} \) as the refractive index in the absence of free electrons. Equation 3.4-12 now shows that the addition of carriers causes an effective decrease in index \( \Delta n \), given by:

\[
\Delta n \approx - \frac{Ne^2}{2m\varepsilon_0\varepsilon_{\infty} \omega^2} \tag{3.4-13}
\]

Despite the crudity of our assumptions, it turns out that the calculation above gives the correct answer, provided we replace the mass of the electron by its effective mass \( m^* \) in the
semiconductor material. This quantity will be defined fully in Chapter 11. Writing $\omega = 2\pi c/\lambda_0$, we may then present Equation 3.4-13 in its more common form:

$$\Delta n \approx -Ne^2\lambda_0^2/8\pi^2e^2m^*\epsilon_0n_\infty$$

3.4-14

A similar result may be obtained if the carriers are holes (the other type of charge carrier present in a semiconductor, which we will also introduce in Chapter 11) instead of electrons.

**DESIGN EXAMPLE**

For n-type GaAs, $n_\infty = 3.5$ and $m^* = 0.067m$. Consequently, $\Delta n \approx -1.9 \times 10^{-21} N\lambda_0^2$ (when $N$ is measured in cm$^{-3}$ and $\lambda_0$ is in $\mu$m). For $N \approx 5 \times 10^{18}$ cm$^{-3}$ and $\lambda_0 = 1\mu$m, we then obtain $\Delta n \approx -0.01$. Although the carrier density needed to obtain such a value of $\Delta n$ is very large, the reduction in index caused by the plasma contribution is extremely useful. For example, fixed changes in index may be induced by doping semiconductor material to change the carrier concentration. Alternatively, carriers may be injected to cause a dynamic change in refractive index. Both effects are employed in semiconductor integrated optics, as we shall see later in Chapter 12.

### 3.5 ANISOTROPIC MEDIA

We now consider the additional effect of the actual **structure** of materials, which we have ignored in favour of their electronic behaviour. Essentially, we have treated all materials as isotropic until now. This assumption implies that the result of applying a field will be independent of its direction. Furthermore, it assumes implicitly that the material polarization will also be parallel to the applied field. Although these assumptions are sometimes valid (e.g. in a truly amorphous material), in crystals we should generally allow for an **anisotropic** response. Since many of the materials used in optoelectronics are crystalline, this is clearly an important feature.

The origins of anisotropy lie in any lack of symmetry in the molecules forming the material. In general, these do not have neat spherical shapes - instead, they have configurations that are characteristic of their composition, electronic structure and chemical activity. The molecular configuration in turn determines the broad classification of any crystalline solid they may form. Generally, there are seven possibilities: a crystal may be cubic, tetragonal, hexagonal, trigonal, orthorhombic, monoclinic or triclinic in structure. These systems represent the only shapes of unit cell (the basic element of a crystal lattice) that can fill all space by repetition.

Figure 3.5-1 shows a particularly simple example, the structure of the cubic ionic crystal NaCl. Here, the Na$^+$ and Cl$^-$ ions are arranged in a regular lattice, which consists of three equal **crystal axes**, all at right angles to one another. This high degree of symmetry suggests that cubic crystals may have optical properties similar to those of amorphous media. However, the axes of the other systems are generally more complicated.

The complete set is shown in Figure 3.5-2 for comparison, together with some examples of common optical materials. As can be seen, the lengths of the axes may be different, and they need not be oriented at right angles. Within these systems, there are a number of subcategories, based on the possible symmetry operations that leave the lattice invariant. Each of these is called a **point group**, and there are a total of 32 of them. Of these, 21 are non-centrosymmetric (i.e., they lack a centre of symmetry).
Because of these differences in symmetry, we would guess that the ease with which an electron may move in the various directions may be different in some crystals. An electronic
model of optical anisotropy could therefore be developed from our previous treatment, for example by assuming several spring constants, each one for electron displacement in a particular direction. If this is done, it is found that the result of applying a field is not independent of direction; instead, the polarization depends on the field direction relative to the crystal axes. The relation between the polarization and the electric field must then be written (at least, for small fields) as:

\[ \mathbf{P} = \varepsilon_0 [\chi] \mathbf{E} \]  

Although Equation 3.5-1 appears more complicated than the definition \( \mathbf{P} = \varepsilon_0 \chi \mathbf{E} \) we have used till now, it has a simple explanation. First, we have replaced the scalar fields by vector ones, to introduce some significance to field directions. Second, we have exchanged the scalar dielectric susceptibility \( \chi \) for the second rank tensor \([\chi]\); this allows the response of the material to be different in different directions relative to the crystal axes.

According to the rules of tensor arithmetic, Equation 3.5-1 may be written out in longhand as:

\[ \begin{align*}
\mathbf{P}_x &= \varepsilon_0 (\chi_{xx} E_x + \chi_{xy} E_y + \chi_{xz} E_z) \\
\mathbf{P}_y &= \varepsilon_0 (\chi_{yx} E_x + \chi_{yy} E_y + \chi_{yz} E_z) \\
\mathbf{P}_z &= \varepsilon_0 (\chi_{zx} E_x + \chi_{zy} E_y + \chi_{zz} E_z)
\end{align*} \]

Thus, Equation 3.5-1 may be interpreted by saying that if the three components of \( \mathbf{P} \) are written as a column vector, this may be found by multiplying a similar vector describing \( \mathbf{E} \) by a 3 x 3 matrix, whose elements are those of the dielectric susceptibility tensor.

If we consider one component of polarization (say, \( \mathbf{P}_x \)), then Equation 3.5-2 shows that there will be contributions to \( \mathbf{P}_x \) from all three of the electric field components \( E_x, E_y \) and \( E_z \). Now, the relative magnitudes of the components of the susceptibility tensor depend in practice on our choice of coordinate system, with respect to the axes of the crystal. For the particular case in which the coordinate system is chosen to coincide with these principal axes, the off-axis components of the tensor are zero. We are then left with the simpler relations:

\[ \begin{align*}
\mathbf{P}_x &= \varepsilon_0 \chi_x E_x \\
\mathbf{P}_y &= \varepsilon_0 \chi_y E_y \\
\mathbf{P}_z &= \varepsilon_0 \chi_z E_z
\end{align*} \]

Here we have written \( \chi_x \) for \( \chi_{xx} \) in this special case, and so on; we will continue to use this reduced notation in other similar situations. Even here, \( \mathbf{P} \) will only be parallel to \( \mathbf{E} \) when:

\[ \chi_x = \chi_y = \chi_z \]

It turns out that this equality holds for amorphous materials, and also for crystals with cubic symmetry. However, for all the other crystal groups, it does not hold. We can interpret this as establishing that the dielectric constant will be different for field components along each of the crystal axes. Thus, we will also need to replace the simple relation \( \mathbf{D} = \varepsilon_0 (1 + \chi) \mathbf{E} \) by:

\[ \mathbf{D} = \varepsilon_0 \{1 + [\chi]\} \mathbf{E} = [\varepsilon] \mathbf{E} \]

where \([\varepsilon] = \varepsilon_0 \{1 + [\chi]\}\) is a second rank tensor definition of the dielectric constant, with \( \varepsilon_{xx} = \varepsilon_0 \{1 + \chi_{xx}\}, \varepsilon_{xy} = \varepsilon_0 \chi_{xy}, \varepsilon_{xz} = \varepsilon_0 \chi_{xz}, \) and so on. Similarly, we may define a tensor \([\varepsilon r]\)
for the relative dielectric constant, with elements \( \varepsilon_{ij} = \varepsilon_i / \varepsilon_0 \). This allows the following classification of crystals in terms of their optical properties. Materials in which two of the components \( \varepsilon_x, \varepsilon_y, \) and \( \varepsilon_z \) are equal (e.g. \( \varepsilon_x = \varepsilon_y \neq \varepsilon_z \)) are termed uniaxial. This class includes tetragonal, hexagonal and trigonal crystals. Similarly, materials in which no two components are equal (e.g. \( \varepsilon_x \neq \varepsilon_y \neq \varepsilon_z \)) are called biaxial. This class contains the orthorhombic, monoclinic and triclinic crystals.

WAVE PROPAGATION IN ANISOTROPIC MEDIA

A logical conclusion from the above is that the phase velocity of waves in anisotropic materials will also depend on the orientation of the electric field vector with respect to the crystal axes, so we would expect some complicated propagation effects to arise. We shall illustrate this with simple examples, beginning with the case of an optical wave travelling in a direction parallel to one of the crystal axes. We shall assume for simplicity that these axes are parallel to the x-, y- and z-directions, and that the wave is travelling in the z-direction. Figure 3.5-3 then shows one possible orientation of the electric field vector in relation to the x- and y-axes.

![Figure 3.5-3 Transverse field of a plane wave travelling in an anisotropic crystal.](image)

Clearly, the component of the transverse field in the x-direction (\( E_x \)) will experience the dielectric constant \( \varepsilon_x \), while \( E_y \) sees \( \varepsilon_y \). Consequently, the wave is divided into two parts, one polarized in the x-direction and travelling with a phase velocity \( c/n_x \), where \( n_x = \sqrt{\varepsilon_x} \), the other y-polarized and moving with a velocity \( c/n_y \). If \( n_x \neq n_y \), the two travel in the same direction, but at different speeds.

DESIGN EXAMPLE - RETARDATION PLATES

This feature is exploited in retardation plates, which are devices that can manipulate the state of polarization of a wave. Assuming the geometry of Figure 3.5-3, with a crystal thickness of d, the phase delays \( \phi_x \) and \( \phi_y \) introduced into each component in travelling through the material are:

\[
\phi_x = k_0 n_x d \quad \text{and} \quad \phi_y = k_0 n_y d
\]

The phase difference introduced (i.e. \( \phi_x - \phi_y \)) will then in general convert the incident plane polarized wave into an elliptically polarized output. There are two particular cases of interest. The first occurs when there is an effective phase difference of \( \pi \) radians, so that:

\[
\phi_x - \phi_y = (2\nu + 1)\pi
\]
Now, the introduction of a $\pi$ radians phase difference is equivalent to changing the sign of one of the components (say, $E_y$). The output electric field is therefore as shown in Figure 3.5-4; as can be seen, the direction of polarization has simply rotated by $2\theta$, where $\theta$ is the angle of polarization of the input wave relative to the x-axis. Consequently, this component (known as a half-wave plate) allows the orientation of the polarization vector to be altered, simply by rotating the crystal. It is made simply by polishing a crystal substrate to the required thickness.

![Figure 3.5-4 Transverse field components after a relative phase delay of $\pi$ radians.](image)

The second case of interest occurs when there is an effective phase difference of $\pi/2$ rads, so that:

$$\phi_x - \phi_y = (2\nu + 1/2)\pi$$

3.5-8

In general, the output will then be elliptically polarized. However, if we arrange for $\theta$ to equal $\pi/4$, then $E_x = E_y$ for the input wave. According to our discussion in Chapter 2 (Equation 2.4-22), the output will then be circularly polarized. The quarter-wave plate is thus a convenient component for converting a plane polarized input into a circularly polarized output, and vice versa.

OFF-AXIS WAVE PROPAGATION

A much more complicated situation occurs when the direction of propagation of the input wave does not lie along a crystallographic axis, when the material polarization must be described by the full tensor expression (Equation 3.5-1). Similarly, the full tensor description of the permittivity must be used. Fortunately, some simplifications are possible. Firstly, it can be shown via the principle of energy conservation that the dielectric tensor must be symmetric, so that $\varepsilon_{ij} = \varepsilon_{ji}$. Secondly, it is possible to devise a simple geometrical interpretation of many aspects of propagation, known as the index indicatrix (or index ellipsoid).

The argument runs as follows. Consider the scalar product $1/2 \mathbf{E} \cdot \mathbf{D}$; as we saw in Chapter 2, this is a measure of the stored electrostatic energy $U_E$. If we form this product with the relevant field components, we can show that:

$$\mathbf{E} \cdot \mathbf{D} = 2U_E = D_x^2/\varepsilon_{xx} + D_y^2/\varepsilon_{yy} + D_z^2/\varepsilon_{zz} + 2D_x D_y/\varepsilon_{xy} + 2D_y D_z/\varepsilon_{yz} + 2D_z D_x/\varepsilon_{zx}$$

3.5-9
Here we have used the symmetry of the dielectric tensor to combine terms of the form $D_x D_y / \varepsilon_{xy}$ and $D_y D_x / \varepsilon_{yx}$. Now, by applying suitable scaling (which merely involves writing $D^2 / 2 \varepsilon_0 U_{ii} = x$, and so on), we may transform Equation 3.5-9 to:

$$\frac{x^2}{\varepsilon_{rx}} + \frac{y^2}{\varepsilon_{ry}} + \frac{z^2}{\varepsilon_{rz}} + \frac{2xy}{\varepsilon_{rxy}} + \frac{2yz}{\varepsilon_{ryz}} + \frac{2zx}{\varepsilon_{rzx}} = 1$$

This is the equation of an ellipsoid, the index indicatrix mentioned earlier. In general, its orientation in space is determined by the magnitudes of the off-axis elements $\varepsilon_{rxy}$, $\varepsilon_{ryz}$ and $\varepsilon_{rzx}$. A typical view of the indicatrix is shown in Figure 3.5-5.

For the special case when the off-axis elements of the dielectric tensor are all zero, Equation 3.5-10 reduces to:

$$\frac{x^2}{\varepsilon_{rx}} + \frac{y^2}{\varepsilon_{ry}} + \frac{z^2}{\varepsilon_{rz}} = 1$$

The index indicatrix is then oriented along the coordinate axes, as shown in Figure 3.5-6, and the semi-major axes of the ellipsoid are equal to $\sqrt{\varepsilon_{rx}}$, $\sqrt{\varepsilon_{ry}}$ and $\sqrt{\varepsilon_{rz}}$ (and thus to $n_x$, $n_y$ and $n_z$). For a uniaxial crystal (for which $n_x = n_y$, say), the indicatrix reduces to an ellipsoid of revolution, whose axis of symmetry (in this case, the $z$-axis) is called the **optic axis**. This particular ellipsoid can be characterised by just two values of refractive index: $n_o$ (the **ordinary index**), which corresponds to $n_x$ or $n_y$, and $n_e$ (the **extraordinary index**), corresponding to $n_z$.

Typical values of $n_o$ and $n_e$ are given in Table 3.5-1 for some common optical materials. Note that in some cases $n_o$ may be greater than $n_e$, while the reverse is true in others.
### Table 3.5-1  Indices of refraction of some uniaxial crystals.

<table>
<thead>
<tr>
<th>Crystal</th>
<th>$n_0$</th>
<th>$n_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcite (CaCO$_3$)</td>
<td>1.658</td>
<td>1.486</td>
</tr>
<tr>
<td>Lithium niobate (LiNbO$_3$)</td>
<td>2.286</td>
<td>2.200</td>
</tr>
<tr>
<td>Lithium tantalate (LiTaO$_3$)</td>
<td>2.176</td>
<td>2.180</td>
</tr>
<tr>
<td>Quartz (SiO$_2$)</td>
<td>1.544</td>
<td>1.553</td>
</tr>
<tr>
<td>Rutile (TiO$_2$)</td>
<td>2.616</td>
<td>2.903</td>
</tr>
</tbody>
</table>

We have seen that the dimensions of the ellipsoid can be usefully interpreted when it is oriented parallel to the coordinate axes. If this is not the case, or if the direction of propagation is not along a principal axis, the interpretation is more complicated. Let us return to the general ellipsoid, Figure 3.5-5, which also shows the direction of propagation of a plane wave, marked $k$. It can be shown that such a wave can be split into two components, plane-polarized at right angles to each other. The directions of polarization are themselves parallel to the axes of the shaded ellipse, which is formed by the intersection of a plane through the origin, normal to $k$, with the index ellipsoid. The effective indices for the two component waves are then given by the semi-major axes of this ellipse.

This general phenomenon (the multi-valued nature of the refractive index) is called **birefringence**, and was first observed by Erasmus Bartolinus (1625-1692). However, from our previous electronic model, we might expect the following additional effects to occur. Firstly, a frequency-dependence of the refractive indices, which results in dispersion of the birefringence. Secondly, a polarization-dependence of the absorption, known as **dichroism**, which can be explained by the introduction of a different damping factor for electronic motion in different directions. If the effect is sufficiently strong, dichroic material can be used to polarize natural light, by preferential absorption of the unwanted polarization. The most famous examples of an artificial dichroic medium are the molecular ‘Polaroid’ sheets invented by Edwin Land in 1928.

### 3.6  NON-LINEAR EFFECTS

In our model of the interaction of radiation and matter, we have supposed that a dipole can be created by an electric field, and that the restoring force on the charges is proportional to displacement. This led to a form of simple harmonic motion, and to a polarization proportional to the driving field. It is reasonable to suppose that this is merely a first approximation to the true behaviour of a dipole, since all real oscillators demonstrate simple harmonic behaviour only for vanishingly small oscillations. We might therefore expect that the relation between $\mathbf{P}$ and $\mathbf{E}$ should really be written as a power series, of which Equation 3.5-1 is simply the first term. If this is done, our model will describe **non-linear** optical effects.

Now, any of Equations 3.5-2 may be written using the following notation:

$$\mathbf{P}_i = \varepsilon_0 \sum_j \chi_{ij} \mathbf{E}_j$$

3.6-1

More generally, if we allow higher-order dependence of the polarization on the electric field, we must write:
\[ P_i = \varepsilon_0 \left\{ \sum_j \chi_{ij} E_j + \sum_j \sum_k \chi_{ijk} E_j E_k + \sum_j \sum_k \sum_l \chi_{ijkl} E_j E_k E_l + \ldots \right\} \]

This shows that (for example), the product of any pair of field components \( E_j \) and \( E_k \) can contribute to \( P_i \) through a third-rank susceptibility tensor whose elements are \( \chi_{ijk} \). The first point to note is that the higher-order terms normally get progressively smaller, so that the elements \( \chi_{ij} \) are the dominant factors determining the dielectric constant. The importance of the higher-order terms lies in a number of new and entirely different effects they make possible.

Consider, for example, an optical field (say, in the x-direction), which at any point can be described as \( E_x = E_x \exp(j\omega t) \). The third-rank susceptibility tensor now makes an additional contribution \( P'_x \) to the total polarization \( P_x \), of the form:

\[ P'_x = \varepsilon_0 \sum_j \sum_k \chi_{xjk} E_j E_k \]

\[ = \varepsilon_0 \chi_{xxx} E_x^2 \exp(2j\omega t) \]

The interesting point here is that \( P'_x \) now has a contribution at an angular frequency 2\( \omega \), i.e. at twice the frequency of the incident wave. This effect is known as second harmonic generation, because this component of \( P \) can be arranged to generate a new wave at the doubled frequency. In fact, frequency doublers based on this principle are now widely used in optoelectronics. For example, a nonlinear crystal pumped at near infrared wavelengths (\( \lambda \approx 0.84 \mu m \)) can be used to generate visible, blue light (at \( \lambda \approx 0.42 \mu m \)), a truly remarkable feat. However, we note here that merely having a significant value of the appropriate coefficient \( \chi_{ijk} \) is not a sufficient condition for a strong second harmonic wave - a further requirement is that the phase velocities of the waves must be matched. We shall find this an important condition in other situations later on.

Another possibility occurs when the field components are associated with different sources. For example, we might have \( E_x = E_x \exp(j\omega t) \) and \( E_y = E_s \exp(j\omega_s t) \). While the former may still be a component of an optical field, the latter might be of much lower frequency, derived from an external source such as a signal generator. In this case, the new contribution to the polarization is:

\[ P'_x = \varepsilon_0 \chi_{xxy} E_x \exp(j\omega t) E_s \exp(j\omega_s t) \]

We can interpret this as follows. In our earlier, first-order derivation we had \( P = \varepsilon_0 [\chi] E_x \) with the dielectric tensor being given by \([\varepsilon] = \varepsilon_0 \{ 1 + [\chi] \}\). From Equation 3.6-4, we see that \( \chi_{xx} \) has effectively been replaced by \( \chi_{xx} + \chi_{xxy} E_s \exp(j\omega_s t) \). Thus, the dielectric constant has now become a function of the applied low-frequency field \( E_s \). This effect (known as the linear electro-optic or Pockels effect) has extremely important consequences, since it allows the modification of the phase velocity of a wave through the application of an external field. The magnitude of the effect in absolute terms is rather small. Typical changes in refractive index are often less than \( 10^{-4} \) (with \( n \) in the broad range 1.45 to 2.50), so that propagation distances of the order of \( 5 \times 10^3 \) wavelengths (= 1 mm) are required for a phase change of \( \pi \) radians. This implies that useful electro-optic devices will be physically large (at least, in comparison to an integrated electronic device).

### 3.7 THE ELECTRO-OPTIC EFFECT

We shall now look at this effect in more detail. The first point to note is that many materials show no electro-optic effect at all! The basic consideration is whether or not the point group has a symmetry which leaves the structure invariant after a change of the coordinates of all
the elements from \( r \) to \(-r\). This can be illustrated by considering the index change that might result from the application of fields \( \mathbf{E} \) and \(-\mathbf{E}\) to a material. In the first case, \( \Delta n = k \mathbf{E} \), where the constant \( k \) is a measure of the strength of the electro-optic effect, while in the second, we would expect \( \Delta n = k (-\mathbf{E}) \). However, if the material has inversion symmetry, then changing from \( \mathbf{E} \) to \(-\mathbf{E}\) should make no difference. This implies that \( k \) must be zero. The electro-optic effect can therefore only exist in materials lacking inversion symmetry. These are called non-centro-symmetric. Of the 32 point groups within the 7 crystal systems, 21 are of this type. All amorphous materials have inversion symmetry, and therefore show no linear electro-optic effect.

We note, however, that regardless of whether a material is centro-symmetric or not (or indeed of whether it is crystalline), the coefficient of the third term in the expansion for the polarization (the fourth-rank susceptibility tensor, whose elements are \( \chi_{ijkl} \)) is never zero. This term gives rise to the Kerr effect, in which we can imagine one component of an applied field generating an asymmetry, while a second produces a form of electro-optic modification of the refractive index, which then characterises the phase velocity of a third (optical) component. However, since the higher-order non-linear coefficients become progressively smaller, higher fields are needed to exploit the Kerr effect. Consequently, for practical devices, the electro-optic effect has been almost exclusively used. This may change as materials with higher Kerr coefficients are discovered.

Now, the tensor describing the electro-optic effect will have a total of 27 elements \( \chi_{ijk} \) (for each of the three possible directions of the polarization \( \mathbf{P} \), there will be three directions for the optical field \( \mathbf{E} \), and three for the low frequency field \( \mathbf{E}_k \)). However, the argument that showed that the dielectric tensor must be symmetric (so that \( \varepsilon_{ij} = \varepsilon_{ji} \) and \( \chi_{ij} = \chi_{ji} \)) can also be shown to lead to \( \chi_{ijk} = \chi_{ikj} \). Thus, the number of different elements needed is only 18. However, the \( \chi_{ijk} \) terms themselves are not normally used to characterise the effect. Since they modify the polarization (and hence the dielectric constant), we expect that the presence of an external electric field will alter the shape of the index indicatrix. It is these changes that are traditionally used instead.

Returning to the characteristic equation of the index ellipsoid, Equation 3.5-10, we note that it could be written in the following alternative form:

\[
B_{xx} x^2 + B_{yy} y^2 + B_{zz} z^2 + 2B_{xy} xy + 2B_{yz} yz + 2B_{zx} zx = 1
\]

where \( B_{xx} = 1/\varepsilon_{xx} \), and so on. Changes in these B-coefficients due to the external field can then be written as:

\[
\Delta B_{ij} = \sum_k r_{ijk} E_k
\]

The terms \( r_{ijk} \) are called the electro-optic or Pockels coefficients (after Carl Pockels, 1865-1913). Clearly, these new elements are simply related to the \( \chi_{ijk} \) terms. Following the arguments above, we would therefore expect \( r_{ijk} = r_{ikj} \). It is therefore customary to contract the indices \( i \) and \( j \) into a single index \( I \), following the set of 'rules' shown in Table 3.7-1.

<table>
<thead>
<tr>
<th>( ij )</th>
<th>( I )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( xx )</td>
<td>1</td>
</tr>
<tr>
<td>( xy )</td>
<td>6</td>
</tr>
<tr>
<td>( xz )</td>
<td>5</td>
</tr>
<tr>
<td>( yx )</td>
<td>2</td>
</tr>
<tr>
<td>( yy )</td>
<td>4</td>
</tr>
<tr>
<td>( yz )</td>
<td>3</td>
</tr>
<tr>
<td>( zx )</td>
<td></td>
</tr>
<tr>
<td>( zy )</td>
<td></td>
</tr>
<tr>
<td>( zz )</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.7-1 Contraction rules for the Pockels coefficients.
The notation $11$ for $xx$, $12$ for $xy$ is also often employed. Hence, $r_{yx}$ (which can also be written $r_{xy}$) becomes contracted to $r_{41}$. By symmetry, this is also equal to $r_{21}$. Similarly, $r_{52}$ corresponds to $r_{12}$ ($= r_{31}$). This notation may also be applied to the indices of the $B_{ij}$ coefficients. If this is done, Equation 3.7-2 may be written in matrix form as:

$$
\begin{bmatrix}
\Delta B_1 \\
\Delta B_2 \\
\Delta B_3 \\
\Delta B_4 \\
\Delta B_5 \\
\Delta B_6
\end{bmatrix} =
\begin{bmatrix}
r_{11} & r_{12} & r_{13} \\
r_{21} & r_{22} & r_{23} \\
r_{31} & r_{32} & r_{33} \\
r_{41} & r_{42} & r_{43} \\
r_{51} & r_{52} & r_{53} \\
r_{61} & r_{62} & r_{63}
\end{bmatrix}
\begin{bmatrix}
E_x \\
E_y \\
E_z
\end{bmatrix}
$$

Thus, for example, we have $\Delta B_1 = r_{11} E_x + r_{12} E_y + r_{13} E_z$, and so on. One important feature of the electro-optic matrix is that it is rather sparse, since the possible elements that can be non-zero are governed entirely by the symmetry of the appropriate point group. Table 3.7-2 is a standard representation of these elements, for the 21 non-centro-symmetric groups.

### TRICLINIC

<table>
<thead>
<tr>
<th>Group 1</th>
<th>$r_{11}$</th>
<th>$r_{12}$</th>
<th>$r_{13}$</th>
<th>$r_{21}$</th>
<th>$r_{22}$</th>
<th>$r_{23}$</th>
<th>$r_{31}$</th>
<th>$r_{32}$</th>
<th>$r_{33}$</th>
<th>$r_{41}$</th>
<th>$r_{42}$</th>
<th>$r_{43}$</th>
<th>$r_{51}$</th>
<th>$r_{52}$</th>
<th>$r_{53}$</th>
<th>$r_{61}$</th>
<th>$r_{62}$</th>
<th>$r_{63}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### MONOCLINIC

| Group 2 | 0 | $r_{21}$ | 0 | $r_{11}$ | 0 | $r_{13}$ | $r_{21}$ | 0 | $r_{23}$ | $r_{31}$ | 0 | $r_{33}$ | $r_{41}$ | $r_{43}$ | 0 | $r_{42}$ | 0 | $r_{52}$ | 0 | $r_{53}$ | 0 | $r_{62}$ | 0 |
|---------|---|---------|---|---------|---|---------|---------|---|---------|---------|---|---------|---------|---------|---|---------|---|---------|---|---------|---|---------|---|---------|
|         |   |         |   |         |   |         |         |   |         |         |   |         |         |         |   |         |   |         |   |         |   |         |   |         |

### ORTHORHOMBIC

<table>
<thead>
<tr>
<th>Group 222</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>$r_{11}$</th>
<th>0</th>
<th>0</th>
<th>$r_{13}$</th>
<th>0</th>
<th>0</th>
<th>$r_{23}$</th>
<th>$r_{31}$</th>
<th>0</th>
<th>$r_{33}$</th>
<th>$r_{41}$</th>
<th>$r_{43}$</th>
<th>0</th>
<th>$r_{42}$</th>
<th>0</th>
<th>$r_{52}$</th>
<th>0</th>
<th>$r_{53}$</th>
<th>0</th>
<th>$r_{63}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R.R.A.Syms and J.R.Cozens  Optical Guided Waves and Devices 21
### TETRAGONAL

<table>
<thead>
<tr>
<th>Gp 4</th>
<th>0</th>
<th>0</th>
<th>r_{13}</th>
<th>Gp 4</th>
<th>0</th>
<th>0</th>
<th>r_{13}</th>
<th>Gp 422</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{13}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-r_{13}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{33}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r_{41}</td>
<td>r_{51}</td>
<td>0</td>
<td>r_{41}</td>
<td>-r_{51}</td>
<td>0</td>
<td>r_{41}</td>
<td>0</td>
<td>0</td>
<td>r_{41}</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r_{51}</td>
<td>-r_{41}</td>
<td>0</td>
<td>r_{51}</td>
<td>r_{41}</td>
<td>0</td>
<td>0</td>
<td>-r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{63}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gp 4mm</th>
<th>0</th>
<th>0</th>
<th>r_{13}</th>
<th>Gp 42m</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{13}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{33}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>r_{51}</td>
<td>0</td>
<td>0</td>
<td>r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r_{51}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{63}</td>
<td>0</td>
</tr>
</tbody>
</table>

### TRIGONAL

<table>
<thead>
<tr>
<th>Gp 3</th>
<th>r_{11}</th>
<th>-r_{22}</th>
<th>r_{13}</th>
<th>Group 32</th>
<th>r_{11}</th>
<th>0</th>
<th>0</th>
<th>Group 3m</th>
<th>0</th>
<th>-r_{22}</th>
<th>r_{13}</th>
</tr>
</thead>
<tbody>
<tr>
<td>-r_{11}</td>
<td>r_{22}</td>
<td>r_{13}</td>
<td>-r_{11}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{22}</td>
<td>r_{13}</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>r_{33}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{33}</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r_{41}</td>
<td>r_{51}</td>
<td>0</td>
<td>r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{51}</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r_{51}</td>
<td>-r_{41}</td>
<td>0</td>
<td>0</td>
<td>-r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{51}</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-r_{22}</td>
<td>-r_{11}</td>
<td>0</td>
<td>0</td>
<td>-r_{11}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-r_{22}</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### HEXAGONAL

<table>
<thead>
<tr>
<th>Gp 6</th>
<th>0</th>
<th>0</th>
<th>r_{13}</th>
<th>Gp 6</th>
<th>r_{11}</th>
<th>-r_{22}</th>
<th>0</th>
<th>Gp 622</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{13}</td>
<td>-r_{11}</td>
<td>r_{22}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{33}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r_{41}</td>
<td>r_{51}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r_{51}</td>
<td>-r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-r_{41}</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-r_{22}</td>
<td>-r_{11}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gp 6mm</th>
<th>0</th>
<th>0</th>
<th>r_{13}</th>
<th>Gp 62m</th>
<th>0</th>
<th>-r_{22}</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{13}</td>
<td>0</td>
<td>0</td>
<td>r_{22}</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>r_{33}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>r_{51}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
A second point is that the magnitudes of the coefficients can vary widely. This limits the usability of some materials, and places restrictions on the crystal cut and on the orientation of the fields that must be used, if significant effects are to be obtained from fields of reasonable strength. Table 3.7-3 shows the values of the electrooptic coefficients for representative materials. Clearly, LiNbO$_3$ would be a strong contender for any device applications, due to the high value of its $r_{33}$ coefficient. New organic materials (such as the crystal 2-methyl-4-nitroaniline, or MNA) have also been found to have large coefficients. However, the relatively weak coefficients of the semiconductors GaAs and InP (which form the basis of many optoelectronic devices) should be noted.

<table>
<thead>
<tr>
<th>Material</th>
<th>Point Group</th>
<th>$r_{11}$, $10^{-12}$ m/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>GaAs</td>
<td>43m</td>
<td>$r_{41} = -1.5$</td>
</tr>
<tr>
<td>InP</td>
<td>43m</td>
<td>$r_{41} = -1.4$</td>
</tr>
<tr>
<td>KD$_2$PO$_4$ (KDP)</td>
<td>42m</td>
<td>$r_{41} = 8$; $r_{63} = 11$</td>
</tr>
<tr>
<td>LiNbO$_3$</td>
<td>3m</td>
<td>$r_{33} = 30.8$; $r_{51} = 28$; $r_{13} = 8.6$; $r_{22} = 3.4$</td>
</tr>
<tr>
<td>MNA</td>
<td>m</td>
<td>$r_{11} = 67$</td>
</tr>
<tr>
<td>SiO$_2$ (Quartz)</td>
<td>32</td>
<td>$r_{41} = 1.4$; $r_{11} = 0.59$</td>
</tr>
</tbody>
</table>

Table 3.7-3  Electro-optic coefficients for representative materials.

How do we make use of all this information? The simplest way is through the index indicatrix. Let us assume that (in the absence of a field) we have chosen a co-ordinate system that is aligned with the principal axes of the crystal, so that the equation for the ellipsoid is of the form:

$$B_x x^2 + B_y y^2 + B_z z^2 = 1$$  \[3.7-4\]

and the ellipsoid is as shown in Figure 3.5-4. Let us also assume the material is LiNbO$_3$, a trigonal crystal with point group 3m. From Tables 3.7-2 and 3.7-3, we see that there are a
number of non-zero elements in the electro-optic matrix. Consequently, several additional terms may be generated in the equation for the indicatrix, depending on the direction of the applied field.

Assuming that the field is applied in the x-direction, the significant elements are $r_{51}$ and $r_{22}$. Performing the matrix multiplication in Equation 3.7-3, we can find the equation of the new index indicatrix as:

$$B_{xx} x^2 + B_{yy} y^2 + B_{zz} z^2 + 2r_{51} E_x xz - 2r_{22} E_x xy = 1$$  

3.7-5

Note that the element $-r_{22}$ in the above is really $r_{61}$ (which is in turn equivalent to $r_{131}$). However, this coefficient has the same magnitude but opposite sign to $r_{22}$ (equivalent to $r_{222}$) for the 3m point group. These extra terms will produce a new ellipsoidal shape, which is shifted in both orientation and scale from the original one.

Alternatively, assuming the field is applied in the z-direction, we obtain:

$$(B_{xx} + r_{13} E_z) x^2 + (B_{yy} + r_{13} E_z) y^2 + (B_{zz} + r_{33} E_z) z^2 = 1$$  

3.7-6

Once again, Equation 3.7-6 describes a new ellipsoid, but this time the result is different. The ellipsoid is still aligned with the coordinate system, but the lengths of the principal axes have altered slightly. We can interpret these changes as follows. Recalling that $B_{xx} = 1/\varepsilon_{xx} = 1/n_x^2$, to a first approximation we can write:

$$\Delta B_{xx} = -2\Delta n_x /n_x^3$$  

3.7-7

However, from Equation 3.7-6 we known that $\Delta B_{xx} = r_{13} E_z$. Equating these two definitions, we can show that the effect of the field $E_z$ is to change the index $n_x$ by an amount:

$$\Delta n_x = -n_x^3 r_{13} E_z /2$$  

3.7-8

The result of the applied field is therefore an index change proportional to the field strength. Note that the important feature is $n_x^3 r_{13}$, rather than simply $r_{13}$. A relatively weak electrooptic coefficient can therefore be compensated for to some extent, if the material has a high refractive index. Similar changes can be found for $n_y$ and $n_z$, as:

$$\Delta n_y = -n_y^3 r_{13} E_z /2$$  

3.7-9

We can use these index changes to construct phase modulators (devices which can alter the phase of an optical wave under electrical control), as described below.

**LONGITUDINAL PHASE MODULATORS**

Figure 3.7-1 shows a longitudinal phase modulator. Here, transparent electrodes have been added to the input and output faces of a slab of electrooptic crystal. This has length $L$ in the z-direction, so the application of a voltage $V$ to the electrodes will lead to a longitudinal field $E_z = V/L$.

Light is passed through the crystal in the z-direction, with the optical field polarized in either the x- or y-directions. For LiNbO$_3$, we can see from Equations 3.7-8 and 3.7-9 that the static electric field will cause an equal refractive index change for both components. In length $L$, this will induce phase shifts:
CHAPTER
THREE

\[
\phi_x = \phi_y = k_p \Delta n L = -\pi n_x r_{13} V / \lambda_0
\]

Note that \( L \) has disappeared from Equation 3.7-10, because a longer crystal results in a
greater phase shift for a given index change, but a smaller electric field for a given voltage.
Consequently, the drive voltage must be independent of the modulator dimensions.

Figure 3.7-1 A longitudinal phase modulator.

DESIGN EXAMPLE

We may calculate the magnitude of the voltage \( V_\pi \) required for a phase shift of \( \pi \) radians
from Equation 3.7-10 as:

\[
V_\pi = \lambda_0 / (n_x r_{13})
\]

For LiNbO\(_3\), \( n_x = 2.286 \) and \( r_{13} = 8.6 \times 10^{-12} \) m/V. Hence, for \( \lambda_0 = 1 \) μm, \( V_\pi \approx 9.7 \) kV. This
is a very high value, presenting an inherent problem for longitudinal phase modulators.

TRANSVERSE PHASE MODULATORS

Improvements are offered by an alternative modulator design, which we now describe.
Figure 3.7-2 shows a transverse phase modulator. Here, electrodes have been deposited
on the top and bottom faces of the crystal, which has been rotated so that the static electric
field is still in the z-direction. The electrode separation is taken to be \( g \), so that the field is
now \( E_z = V / g \), but the length of the crystal in the direction of optical propagation remains \( L \).
This time, the input optical wave may be polarized in the x- or z-directions. Equations 3.7-8
and 3.7-9 show that the resulting phase shifts are now different, being given by:

\[
\phi_x = k_0 \Delta n_x d = -\pi n_x r_{13} V L / (\lambda_0 g)
\]

\[
\phi_z = k_0 \Delta n_z d = -\pi n_z r_{33} V L / (\lambda_0 g)
\]

Since \( r_{33} > r_{13} \) in LiNbO\(_3\), the greatest effect will be experienced by the z-polarized wave.
The magnitude of the voltage required for \( \pi \) radians phase shift of this component is now:

\[
V_\pi = \lambda_0 g / (n_z r_{33} L)
\]

Equation 3.7-13 is no longer independent of the crystal dimensions, since it contains \( g \) and \( L \). This suggests that if the ratio \( g / L \) can be arranged to be \( \approx 10^{-3} \) (say, by using an electrode
gap of a few microns and a length of several millimetres), \( V_\pi \) will reduce from kilovolts to
volts. Since this can be done in guided wave optics, the door has been opened to the widespread use of electro-optic devices.

Figure 3.7-2 A transverse phase modulator.

3.8 STRESS-RELATED EFFECTS

In this section, we consider a number of stress-related effects, all of which have been exploited for optical device applications. We begin with the photoelastic effect, whereby a transparent, isotropic material may become optically anisotropic under the influence of an applied stress. This phenomenon, also known as stress-induced birefringence, was first investigated by David Brewster in 1816.

The stress to which a solid body is subjected may be characterised by a second-rank stress tensor \([T]\), whose elements \(T_{ij}\) represent the force applied along the \(i\) direction per unit area of a plane normal to the \(j\) direction. The diagonal components of \([T]\) therefore represent normal components of stress, while the off-diagonal components are shear stresses. By considering moments, it is simple to show that \([T]\) must be symmetric, to avoid any net torque on the body.

Similarly, the elastic strain of the body due to applied stresses may be described by a symmetric, second-rank tensor \([S]\), whose diagonal elements \(S_{ii}\) (defined as \(S_{ii} = \partial U_i / \partial x_i\), where \(U_i\) is the local displacement of the body in the \(i\) direction, and \(x_i\) is the coordinate in the same direction) represent normal strains, and whose off-diagonal elements \(S_{ij}\) (defined as \(S_{ij} = \{ \partial U_i / \partial x_j + \partial U_j / \partial x_i \}\)) are shear strains. Note that we have avoided the use of the notation \(\sigma\) and \(\varepsilon\) for stress and strain, to avoid confusion with other optical parameters.

Because of the symmetry of \([T]\) and \([S]\), a contracted notation is often used to describe the elements \(T_{ij}\) and \(S_{ij}\), as was done previously for the Pockels coefficients. Since the rules for this have already been given in Table 3.7-1, we shall not repeat them.

Suffice it to say that if this notation is employed, \([T]\) and \([S]\) may be represented by two six-element vectors \(T = (T_1, T_2, \ldots T_6)\) and \(S = (S_1, S_2, \ldots S_6)\), which may be used to characterise the photoelastic effect in terms of the changes in the B-coefficients of the index ellipsoid induced by stresses or strains. Two alternative definitions are used. Firstly, in terms of stresses, we may write in longhand:
where the terms $q_{ij}$ are elements of a second-rank **stress-optical tensor** $[q]$, which have units of $m^2/N$. Secondly, in terms of strains, and using a more compact notation, we may put:

$$\Delta B = [p] S$$

$$3.8-2$$

where $\Delta B = (\Delta B_1, \Delta B_2, \ldots, \Delta B_6)$ and $[p]$ is a corresponding **strain-optical tensor**, whose elements $p_{ij}$ are dimensionless. Now, for elastic deformation, there is a linear relation between stress and strain. With the contracted notation used above, this may be written as:

$$T = [c] S$$

$$3.8-3$$

where $[c]$ is a second-rank **stiffness tensor** with elements $c_{ij}$. Clearly, the stress- and strain-optical tensors are related, since $[p] = [q][c]$. The two representations given above are therefore entirely complementary. With knowledge of the coefficients of $[p]$ or $[q]$, the effect of any stress or strain distribution may easily be computed, as we have done before for the electro-optic effect.

It can be shown that the non-zero elements in the photoelastic tensors are generally in different positions to those of the electro-optic tensor, but in the same positions as those of the Kerr effect, so the photoelastic effect is observable in centro-symmetric crystals, liquids (e.g. water) and amorphous solids (e.g. plastics, silica glass). Table 3.8-1 shows the stress-optical tensor for an isotropic medium; this is particularly simple, since it is described entirely by two coefficients, $q_{11}$ and $q_{12}$.

We may use this example to evaluate the effect of various applied stresses. From Equation 3.8-1, it is apparent that a uniaxial stress (say, a non-zero value of $T_1$) will result in the following changes to the $B$-coefficients:

$$\Delta B_1 = q_{11} T_1, \quad \Delta B_2 = \Delta B_3 = q_{12} T_1 \quad \text{and} \quad \Delta B_4 = \Delta B_5 = \Delta B_6 = 0.$$  

$$3.8-4$$

Clearly, if $q_{11}$ and $q_{12}$ are unequal (as they often are), the material will now be birefringent, with principle axes located parallel and perpendicular to the direction of the applied stress. In each case, the index change is of the form $\Delta n = -n^3 q T/2$, where $q$ the relevant stress-optical coefficient. By analogy, we may obtain the corresponding expression in terms of strain, as $\Delta n = -n^3 p S/2$, where $p$ is the corresponding strain-optical coefficient.
\[
[q] = \begin{bmatrix}
q_{11} & q_{12} & 0 & 0 & 0 \\
q_{12} & q_{11} & 0 & 0 & 0 \\
q_{12} & q_{11} & 0 & 0 & 0 \\
0 & 0 & q_{11} - q_{12} & 0 & 0 \\
0 & 0 & 0 & q_{11} - q_{12} & 0 \\
0 & 0 & 0 & 0 & q_{11} - q_{12}
\end{bmatrix}
\]

Table 3.8-1 Photoelastic tensor for an isotropic medium.

Similarly, if the stress is now applied hydrostatically (so that \(T_1 = T_2 = T_3 = T\), say), the index changes become:

\[
\Delta n_x = \Delta n_y = \Delta n_z = -n^3(q_{11} + 2q_{12})T/2
\]

This time, the refractive index change is isotropic, so the material does not become birefringent. By convention, compressive stresses are defined as being negative in sign (since they cause negative strain), while tensile stresses are positive. An increase in hydrostatic pressure will therefore result in a uniform increase in refractive index that is linearly proportional to the pressure.

THE ACOUSTO-OPTIC EFFECT

Now, an **acoustic wave** is a form of travelling wave, consisting of moving variations in stress. From the above, it is clear that the passage of an acoustic wave through a photoelastic medium should be accompanied by a travelling refractive index variation, with peaks and troughs in index correlated with points of high and low stress in the wave. The variation must therefore be periodic, with the same periodicity as the acoustic wave (\(\lambda = v/f\), where \(v\) is the acoustic velocity and \(f\) is the frequency). In 1922, Brillouin suggested that this variation could be used to diffract an optical beam, a prediction that was verified almost simultaneously in 1932 by Debye and Sears in America and Lucas and Biquard in France.

In this form, the photoelastic effect is commonly known as the **acousto-optic effect**. We shall describe the diffraction process itself in Chapter 7; in the mean time, we shall simply calculate the peak index change expected from an acoustic wave of a given power.

Acoustic waves may take several different forms, depending on whether they propagate in solids or liquids, and in the bulk of a medium or near its surface. In solids, bulk waves can exist as either longitudinal or transverse waves (which have mechanical displacements parallel and normal to the direction of propagation). Figure 3.8-1 shows a longitudinal wave, which has a strain variation only in the direction of propagation. Only one component of strain (say, \(S_x\)) is non-zero, so we can describe the wave by the travelling variation 

\[
S_x = S_{\text{max}} \cos(\omega t - kx), \text{ where } \omega = 2\pi f \text{ and } k = 2\pi/\lambda.
\]
Note that it is also possible for surface acoustic waves to exist; these are propagation modes that have both longitudinal and transverse components, and are constrained to propagate with finite amplitude only close to a surface boundary. They are of considerable importance in guided wave optics, because they can interact efficiently with optical waves propagating in a guide near a surface.

Now, it can be shown that the power per unit area carried by the longitudinal wave above is given by $S_{\text{max}}^2 \rho v^3/2$, where $\rho$ is the density of the medium supporting the acoustic wave. Since this equals $P/A$, where $P$ is the total power in the acoustic beam and $A$ is its cross-section, we must have:

$$S_{\text{max}} = \sqrt{2P/\rho v^3}$$  

3.8-6

We may then obtain the peak index change in a given direction as:

$$\Delta n_{\text{max}} = n^3 p S_{\text{max}}^2 = \sqrt{(Pp^2 n^6/2A \rho v^3)}$$  

3.8-7

where $p$ is the relevant strain-optical coefficient. This result is often written as $\Delta n_{\text{max}} = \sqrt{(P M_2^2/2A)}$, where $M_2 = p^2 n^6/\rho v^3$ is one of a number of acousto-optic figures of merit (actually, the most important for many practical purposes), useful for comparing different acousto-optic materials.

**DESIGN EXAMPLE**

We may estimate $M_2$ for a typical material, fused silica, as follows. Fused silica is isotropic, with non-zero strain-optical coefficients of $p_{11} = 0.121$ and $p_{12} = 0.27$. Its refractive index, density and acoustic velocity are 1.46, $2.2 \times 10^3$ kg/m$^3$ and 5960 m/s, respectively. The largest value of $M_2$ is obtained from $p_{12}$, as $M_2 = 0.272 \times 1.466 / (2.2 \times 10^3 \times 5960^3) = 1.51 \times 10^{-15}$ s$^3$/kg. An acoustic beam of power 1 W and cross-sectional area 1 mm$^2$ would then yield a peak index change of $\Delta n_{\text{max}} = \sqrt{1 \times 1.51 \times 10^{-15} / (2 \times 10^{-6})} = 2.75 \times 10^{-5}$. Table 3.8-2 shows values of $M_2$ for some other acousto-optic materials. One of the most widely used is LiNbO$_3$, since it is also piezoelectric. This allows simple excitation of an acoustic wave by an electrical drive signal, as we shall describe below.

<table>
<thead>
<tr>
<th>Material</th>
<th>$v$ (m/s)</th>
<th>$\rho$ (g/cm$^3$)</th>
<th>$M_2$ ($10^{-15}$ s$^3$/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fused silica</td>
<td>5960</td>
<td>2.2</td>
<td>1.51</td>
</tr>
<tr>
<td>LiNbO$_3$</td>
<td>6570</td>
<td>4.64</td>
<td>6.95</td>
</tr>
<tr>
<td>PbMoO$_4$</td>
<td>3630</td>
<td>6.95</td>
<td>35.8</td>
</tr>
<tr>
<td>TeO$_2$</td>
<td>4200</td>
<td>6.0</td>
<td>34.6</td>
</tr>
<tr>
<td>GaAs</td>
<td>5150</td>
<td>5.34</td>
<td>104.19</td>
</tr>
</tbody>
</table>

Table 3.8-2 Parameters of typical acousto-optic materials (from I.C. Chang, IEEE Trans. Sonics & Ultrasonics SU-23, 2-21 (1976)).

**THE PIEZOELECTRIC EFFECT**

All crystals that exhibit the linear electro-optic effect possess one further stress-related characteristic: they are piezoelectric. The piezoelectric effect was first noted by the brothers Pierre and Jacques Curie in 1880, and is strong in materials such as BaTiO$_3$, LiNbO$_3$, and...
PLZT, quartz and ZnO. It has two manifestations. In the first, the application of an electric field to a piezo-electric crystal may give rise to an internal strain. If the field is a DC one, the crystal dimensions will therefore alter slightly.

This particular effect has found widespread application in guided wave optics, since it allows the construction of devices capable of moving small components through short distances (say, 10 µm), with a resolution of the order of 1 nm. Using a three-axis piezoelectric micropositioner, it is possible to position one component (e.g. a single-mode optical fibre) relative to another (perhaps a semiconductor laser) to high accuracy, so that transmission between the two is optimised. Often, this is done under closed-loop control. Piezoelectric devices are therefore essential to the construction of low-loss jointed systems.

For an AC electric field, on the other hand, the strain distribution inside the crystal will be time-varying. If the crystal is arranged as a parallel-sided slab, it may act as an acoustic resonator, so that the overall pattern is equivalent to a standing acoustic wave. This can be used to excite travelling acoustic waves in a non-piezoelectric material in close mechanical contact with the crystal. A piezoelectric crystal may therefore be used as an extremely convenient transducer between an electrical drive signal and an acoustic wave, for example in an acousto-optic device.

In the second manifestation of the piezoelectric effect, the application of a stress may result in an internal electric field, so that a voltage is developed across the crystal. Here, the major consequence is that the effect of an acoustic wave is more complicated than we have described so far. For example, the travelling strain distribution can produce a change in refractive index directly, through the photoelastic effect. However, it can also generate an electric field (via the piezo-electric effect), which in turn produces a change in index (via the electro-optic effect). Depending on the material, the index change produces by this second, indirect mechanism may either add to or subtract from that produced directly.

Transducers for surface acoustic waves are rather different, and will be described in Chapter 7.
Problems

3.1 Using the Sellmeier coefficients given in Table 3.1-1, compute the refractive index of silica at wavelengths of 1.3 and 1.55 µm.
[1.447; 1.444]

3.2 Sellmeier's equation \( n^2 = 1 + \sum_j A_j \lambda_0^2 / (\lambda_0^2 - \lambda_j^2) \) can be used to model material dispersion over a wide spectral range. Show that, for the case of a single resonance located in the U.V., dispersion over the visible spectrum can be approximated by **Cauchy's equation** \( n' = A + (B/\lambda_0^2) + (C/\lambda_0^4) \), and find A, B and C in terms of the Sellmeier coefficients. What phenomenon would you expect Cauchy's equation to be unable to describe?

3.3 Using the data given in Table 3.4-1, estimate the density of free electrons in silver. Over what spectral range would the real part of the dielectric constant \( (\varepsilon_r') \) be negative?
[4.68 x 10^{28} \text{ /m}^3; \lambda_0 > 0.155 \text{ µm}]

3.4 By constructing a simple model for the behaviour of electrons in a high frequency electromagnetic field, show that a reasonable approximation to the dispersion relation for a low-density, ionised gas (e.g. the ionosphere) is given by the expression \( \omega^2 = \sqrt{\omega_p^2 + c^2 k^2} \)

3.5 The dielectric constant of a semiconductor may be modified by the introduction of free carriers, through the plasma contribution. Find an approximate expression for the change in complex relative dielectric constant, assuming that collision damping is not negligible. What consequence will follow from the resulting finite value of \( \varepsilon_r'' \)?

3.6 The figure below shows an optical experiment involving a birefringent wave plate and a polarizer. The former is constructed from a lithium niobate slab of thickness 3.684 mm (\( n_0 = 2.286, n_e = 2.2 \) for LiNbO₃). You may assume that the latter is perfect, i.e. that it blocks the unwanted polarization absolutely, while allowing the desired polarization through without attenuation. A plane wave (of wavelength 0.633 µm) is passed through the two components, which are initially rotated about the optical axis to maximise the output power. What will be the percentage transmission if the birefringent plate is now rotated by (a) 22.5°, and (b) 45°?

[50%; zero]

3.7 The general equation for the index indicatrix in anisotropic media may be written as \( B_{xx} x^2 + B_{yy} y^2 + B_{zz} z^2 + 2B_{xy} xy + 2B_{xz} yz + 2B_{zx} zx = 1 \), where \( B_{xx} = 1/\varepsilon_{xxx} \) and so on.
Show that, if the coordinates are represented by a radius vector $\mathbf{r}$, and the B-coefficients are assembled into a symmetric, second-rank tensor $[B]$, this equation may be rewritten in the alternative \textbf{quadratic form} $\mathbf{r}^T [B] \mathbf{r} = 1$, where $\mathbf{r}^T$ represents the transpose of $\mathbf{r}$.

3.8 The $3 \times 3$ matrix $[B]$ possesses a set of eigenvalues $\lambda_i$ and eigenvectors $\mathbf{v}_i$ defined using $[B] \mathbf{v}_i = \lambda_i \mathbf{v}_i$. Since $[B]$ is symmetric, the eigenvectors are orthogonal, so $\mathbf{v}_i^T \mathbf{v}_j = 0$. It may also be assumed that the eigenvectors are \textit{normalised} so that $\mathbf{v}_i^T \mathbf{v}_i = 1$.

Show that, if the eigenvalues and eigenvectors are assembled into two new matrices $[\Lambda]$ (whose diagonal elements are $\lambda_i$) and $[V]$ (which contains the vectors $\mathbf{v}_i$ arranged in columns), $[B]$ may be written as $[B] = [V] [\Lambda] [V]^T$. Hence, find an expression for a new co-ordinate system $\mathbf{r}'$, which coincides with the axes of the index ellipsoid. What are the lengths of the semi-major axes of this ellipsoid?

3.9 A $\text{KH}_2\text{PO}_4$ (KDP) crystal is oriented with its crystal axes parallel to the $x$-, $y$- and $z$-directions. From the information given in Tables 3.7-2 and 3.7-3, determine the equation of the index ellipsoid when an electric field is applied in the $z$-direction.

3.10 Using the methods of Questions 3.7 and 3.8, show that the lengths of the semi-major axes of the index ellipsoid for the KDP crystal in Question 3.9 are unaltered by the application of the electric field.
SUGGESTIONS FOR FURTHER READING

Maletson I.H. "Interspecimen comparison of the refractive index of fused silica" JOSA 55, 1205-1209 (1965)
Yariv A. "Optical electronics" Holt, Rinehart & Winston, New York (1965); Chs. 1, 8 & 9
4.1 SCALAR THEORY

The analysis of Chapter 2 was based on the vectorial form of Maxwell’s equations, which we used to demonstrate the existence of plane waves. Clearly, other types of wave must exist - an infinite plane wave is only an idealisation, and real waves have the form of bounded beams. Unfortunately, vector theory rapidly becomes unmanageable when applied to more complicated geometries. We therefore introduce an alternative scalar theory, which can be used to illustrate the optics of beams.

Firstly, we impose the condition that the entire field is harmonically varying, at a single angular frequency $\omega$. Secondly, we assume that the electric field vector always points in the same direction. This is reasonable for linearly polarized, paraxial fields, i.e. beams with amplitudes that are only non-zero close to some axis of propagation. Finally, we take the medium to be free space. With these assumptions, the time-independent vector wave equation 2.3-13 reduces to the simple scalar equation:

$$\nabla^2 E + \omega^2 \mu_0 \varepsilon_0 E = 0$$

4.1-1

It is simple to repeat our previous analysis, and find plane wave solutions to Equation 4.1-1. For a wave travelling in the +z-direction, for example, we would obtain $E = E_0 \exp(-jk_0 z)$, where $E_0$ is some fixed amplitude and $k_0$ (the propagation constant) is again given by $k_0^2 = \omega^2 \mu_0 \varepsilon_0$. Now we shall find solutions for non-planar waves. We shall begin with another wave of simple geometrical form, the spherical wave.

4.2 SPHERICAL WAVES

Until now, we have worked in cartesian co-ordinates, for which the Laplacian in Equation 4.1-1 is given by $\nabla^2 = \partial^2 / \partial x^2 + \partial^2 / \partial y^2 + \partial^2 / \partial z^2$. Unfortunately, the evaluation of $\nabla^2$ in other coordinate systems is (in general) rather difficult. This makes the wave equation hard to set up, let alone solve. In this case, however, there is an observation we can use to simplify matters. It is that spherical waves are, by definition, spherically symmetric, so our field solution must depend only on a radial co-ordinate r. This implies that $E(x, y, z) = E(r)$, so the first derivative of E with respect to x can be found as:

$$\partial E/\partial x = dE/dr \partial r/\partial x$$

4.2-1

Similarly, we may obtain the corresponding second derivative as:

$$\partial^2 E/\partial x^2 = d^2 E/dr^2 (\partial r/\partial x)^2 + dE/dr \partial^2 r/\partial x^2$$

4.2-2

Now, since $r^2 = x^2 + y^2 + z^2$, we must have:

$$2r \, dr = 2x \, dx, \quad \text{or} \quad \partial r/\partial x = x/r$$

4.2-3

and:

$$\partial^2 r/\partial x^2 = (1/r) + x \, \partial(1/r)/\partial x$$

$$= (1/r) \left( 1 - x^2/r^2 \right)$$

4.2-4
so that:
\[ \partial^2 E/\partial x^2 = (x^2/r^2) \partial^2 E/\partial r^2 + (1/r) (1 - x^2/r^2) \partial E/\partial r \]
\[ \text{4.2-5} \]

We can of course find \( \partial^2 E/\partial y^2 \) and \( \partial^2 E/\partial z^2 \) in a similar way. By adding all three second derivative terms together, we can obtain an expression for the Laplacian that involves \( r \) only:
\[ \nabla^2 E = \partial^2 E/\partial r^2 + (2/r) \partial E/\partial r \]
\[ \text{4.2-6} \]

For spherically symmetric fields, the scalar wave equation therefore reduces to:
\[ \partial^2 E/\partial r^2 + (2/r) \partial E/\partial r + \omega^2 \mu_0 \epsilon_0 E = 0 \]
\[ \text{4.2-7} \]

This is still quite a hard equation to solve, so we now proceed as follows. Let us define a new variable \( F(r) \), such that:
\[ F(r) = r E(r) \]
\[ \text{4.2-8} \]

Derivatives of \( F \) may then be evaluated as:
\[ dF/dr = r \partial E/\partial r + E \]
\[ \text{4.2-9} \]
and:
\[ d^2 F/\partial r^2 = r \partial^2 E/\partial r^2 + \partial E/\partial r + \partial E/\partial r \]
\[ = r \partial^2 E/\partial r^2 + 2 \partial E/\partial r \]
\[ \text{4.2-10} \]

Using these, Equation 4.2-7 may be re-written as:
\[ (1/r) d^2 F/\partial r^2 + \omega^2 \mu_0 \epsilon_0 (1/r) F = 0 \]
\[ \text{4.2-11} \]

Apart from the common factor of \( 1/r \) (which we can remove by cancellation) Equation 4.2-11 looks just like a one-dimensional wave equation, with the radial co-ordinate \( r \) now being the only variable. By analogy with our previous results for plane waves, we can therefore guess that solutions travelling in the +\( r \) direction (i.e. outward-propagating waves) have the form:
\[ F(r) = E_0 \exp(-jk_0 r) \]
\[ \text{4.2-12} \]
Where \( E_0 \) is again a constant factor. Complete solutions for \( E(r) \) are therefore given by:
\[ E(r) = E_0 /r \exp(-jk_0 r) \]
\[ \text{4.2-13} \]

Equation 4.2-13 is the spherical wave solution we have been seeking. It has two components. The first is the term \( \exp(-jk_0 r) \), which describes the phase of the wave, and is similar to the complex exponential variation we have previously found in plane waves. However, the surfaces of constant phase (known as phase-fronts) are now spheres instead of planes, as shown in Figure 4.2-1. (Figure 4.2-1 illustrates another important concept, the ray. This is the local normal to a phase-front, and it effectively defines the direction the wave is moving in at any particular point. This idea is often used to deduce many of the features of an optical system, without invoking waves at all.)
The second is a $1/r$ dependence, which implies that the amplitude falls off with distance from the origin. The explanation for this is power conservation. We found for plane waves that the time-averaged power per unit area is proportional to the electric field amplitude squared, and we might expect the same to be true for a spherical wave. However, at any radius $r$, a spherical wave must cross a hypothetical surface of area $4\pi r^2$. In order for the total power passing through this surface to be independent of radius, the wave amplitude must vary inversely with $r$. The solution unfortunately predicts an infinite amplitude at the origin. However, we shall ignore this for the time being, since it only occurs with a perfect spherical wave. In practice, an additional effect, known as diffraction, places an inherent limit on the amplitude achievable with bounded beams, as we shall see shortly.

**PARAXIAL APPROXIMATION**

We shall now briefly introduce a paraxial approximation to the spherical wave, which will prove useful later on. Consider the wave amplitude in an $x$-$y$ plane, a distance $z$ from the origin, as shown in Figure 4.2-2.

Now, the phase variation of the wave is defined by the exponential $\exp(-jk_0 r)$, where $r^2 = z^2 + R^2$ and $R^2 = x^2 + y^2$. For small $R$, we can put:

$$r = z \sqrt{1 + R^2/z^2} \approx z + R^2/2z$$

so we can approximate the phase variation by:
\[ \exp(-jk_0 r) \approx \exp(-jk_0 z) \cdot \exp(-jk_0 R^2/2z) \]  

We could make a similar approximation for the \(1/r\)-dependence of the wave amplitude. However, because \(1/r\) is such a slowly varying function, we can make the approximation a cruder one in this case, and put \(r \approx z\). The complete form of the wave is then:

\[ E = E_0/z \exp(-jk_0 R^2/2z) \cdot \exp(-jk_0 z) \]  

We can now see clearly the significant features of our paraxial spherical wave. It travels mainly in the \(z\)-direction, because of the \(\exp(-jk_0 z)\) term in Equation 4.2-16. However, in any plane \(z = \text{constant}\) (i.e. in a plane normal to the main axis of propagation) the amplitude is roughly constant, while the phase varies parabolically with distance \(R\) from the axis. Thus we may write:

\[ E(R, z) \approx A(z) \exp(-jk_0 R^2/2z) \]  

It is easy to obtain a similar approximation for a converging wave, rather than a diverging one. The key information is contained in the sign of the exponential in Equation 4.2-17, which describes the curvature of the phase front. With the negative sign above, we have seen that the wave diverges from an origin a distance \(z\) behind the plane of observation. If the sign is made positive instead, the phase curvature is reversed. A converging wave may therefore be defined to vary as:

\[ E'(R, z) = A(z)' \exp(+jk_0 R^2/2z) \]  

This wave will converge on a different origin, a distance \(z\) in front of the plane of observation, as shown in Figure 4.2-3.

The astute reader will have noticed that the arguments used above are somewhat self-contradictory. We cannot have an uniform amplitude (which implies an unbounded beam) while maintaining the validity of a paraxial approximation. Some form of 'envelope' must be introduced to limit the beam cross-section in a truly paraxial wave, which then cannot be spherically symmetric ...! We will show how to find a more consistent solution later on. In the mean time, we shall continue with our approximate description of a spherical wave, and examine the way in which it can be manipulated by a lens.
4.3 LENSES

**Lenses** are transmissive optical components, which operate by altering the phase of an optical wave as it passes through. In the simplest approximation, their action can be described by a complex *transmissivity* or *transparency function*. To illustrate this, we shall consider the propagation of a plane wave at normal incidence through a dielectric slab. We assume for simplicity that the surround is air. The slab itself has refractive index \( n \) and thickness \( d \), as in Figure 4.3-1.

![Figure 4.3-1 A plane wave passing through a slab of dielectric.](image)

When the wave strikes the first boundary, at \( z = 0 \), it will actually be reflected and refracted. The first term implies that not all of the wave will penetrate the slab, and the second that the direction of the transmitted component may change. Both will be explained fully in Chapter 5, when we learn how to solve boundary-matching problems exactly. In the mean time, we will simply borrow a result from the full solution: at normal incidence, there is little reflection and no refraction. We may therefore consider to good approximation that the wave enters the slab unaltered.

Now, wave solutions in a medium of refractive index \( n \) have the form \( E = E_0 \exp(-jk_0 nz) \). By the time the wave emerges from the slab at \( z = d \), it will therefore have accumulated additional phase, defined by the exponential term \( \exp(-jk_0 nd) \). We can therefore relate the input and output wave amplitudes by:

\[
E_{\text{out}} = \tau_s E_{\text{in}}
\]

4.3-1

Where:

\[
\tau_s = \exp(-jk_0 nd)
\]

4.3-2

The complex transparency function \( \tau_s \) now describes the effect of the slab: it is to impart a phase-shift, which depends on the *optical thickness* \( nd \). Because the wave travels slower in a medium of refractive index greater than unity, this amounts to a retardation of the wave by the material.

The **thin lens** is a more complicated optical element, which works by the same principle. It consists of a lump of dielectric (often glass), of refractive index \( n \) and maximum thickness \( d \), with two highly polished spherical surfaces of radii \( r_1 \) and \( r_2 \). Figure 4.3-2 shows some of the possible cross-sectional shapes a thin lens may take, depending on the magnitude and sign of \( r_1 \) and \( r_2 \).
Once more, the effect of the dielectric is to retard a wave. Clearly, the retardation is greater in thicker regions of the lens. In a diverging lens, the dielectric is thicker in the outer regions; this retards the outermost parts of the wave, causing it to diverge. The reverse is true in a converging lens, which has thicker glass in its central region. If one of the lens surfaces is plane, the lens is described as plano-convex (for a converging lens) or plano-concave (for a diverging lens). In either case, one of the radii $r_1$ or $r_2$ must be infinite.

We can analyse a thin lens by first finding the relevant transparency function. Again, this can be done by evaluating the phase shift incurred by passing between two parallel planes, say at $z = 0$ and $z = d$. Figure 4.3-3 shows the more detailed diagram needed for this calculation, for a converging lens.

First, we work out the thickness $t$ of the glass at any radial distance $R$ from the axis as:

$$t = d - \{ r_1 [- \cos (\theta_1)] + r_2 [1 - \cos (\theta_2)] \}$$

where $\theta_1$ and $\theta_2$ are as shown in Figure 4.3-3.

Using small-angle approximations, we may put:

$$t \approx d - (R^2 / 2) (1/r_1 + 1/r_2)$$

Figure 4.3-2 Lens types.

Figure 4.3-3 Detailed geometry of a converging lens.
Now, in travelling from $z = 0$ to $z = d$, a wave will cover a distance $t$ in a medium of refractive index $n$, and a distance $(d - t)$ in a medium of unity refractive index. This yields a total optical thickness $\delta$ of:

$$\delta = nt + (d - t) = nd - (n - 1) \left( \frac{R^2}{2} \right) \left( \frac{1}{r_1} + \frac{1}{r_2} \right)$$

Equation 4.3-5 shows that the optical thickness varies parabolically with distance $R$ from the axis. This variation has two components. The first is a constant, which would arise if the lens were to be replaced by a uniform slab of thickness $d$; this is relatively unimportant. The second is a term proportional to $R^2$ - this is far more interesting, since it accounts for the focusing property of the lens.

The lens transfer function $\tau_L$ can now be found (by analogy with the slab) as:

$$\tau_L = \exp(-jk_0\delta) = \tau_s \exp[jk_0(n - 1) \left( \frac{R^2}{2} \right) \left( \frac{1}{r_1} + \frac{1}{r_2} \right)]$$

Equation 4.3-6

Here $\tau_s$ is the slab transfer function previously found in Equation 4.3-2. Note that we can write Equation 4.3-6 in the alternative form:

$$\tau_L = \tau_s \exp[jk_0R^2/2f]$$

Equation 4.3-7

Here we have introduced an important new parameter, the focal length $f$ of the lens, given by:

$$\frac{1}{f} = (n - 1) \left( \frac{1}{r_1} + \frac{1}{r_2} \right)$$

Equation 4.3-8 is known as the lens maker's equation, since it offers a useful formula for choosing the surface curvatures required in a lens of a given focal length. Using the transfer function, we can now examine the effect of the lens on a spherical wave. We shall see that a lens can be used to transform an incoming wave, whose origin is located at one particular point, into an outgoing wave with a different origin. This process is called imaging.

**IMAGING**

Figure 4.3-4 shows a thin lens, illuminated on-axis by a paraxial spherical wave originating from an object point $u$ away from the lens. We shall now show that the transmitted output is a wave converging on a different origin (or image point) a distance $v$ away from the lens.

![Figure 4.3-4 Transformation of a spherical wave by a lens.](image)
Now, the input electric field $E_{in}$ can be found in the plane of the lens from Equation 4.2-16 as:

$$E_{in} = A(u) \exp(-jk_0 R^2/2u)$$

4.3-9

The output field $E_{out}$ is then found by multiplying the input field by $\tau_L$, to give:

$$E_{out} = \tau_L E_{in} = A'(u) \exp(+jk_0 R^2/2f) \exp(-jk_0 R^2/2u)$$

4.3-10

where we have written $A'(u)$ for $A(u) \tau_s$. The important feature of this new field is that the phase still varies parabolically with the distance $R$ from the axis. This implies that we could write it as:

$$E_{out} = A'(u) \exp(+jk_0 R^2/2v)$$

4.3-11

Where $v$ is given by:

$$1/u + 1/v = 1/f$$

4.3-12

The exponential in Equation 4.3-11 has the positive sign associated with a converging wave, which will travel towards an origin or focus at a distance $v$ from the lens. This shows that a lens can be used to convert one paraxial spherical wave into another, and Equation 4.3-12 (known as the thin lens imaging equation) provides a convenient relation between the object and image points.

**DESIGN EXAMPLES**

We can illustrate lens imaging with a number of examples. If $u = \infty$ (so that the phase-front of the input wave is effectively plane), Equation 4.3-12 shows that $v = f$. A lens can therefore focus a plane wave to a point a distance $f$ away from the lens, as shown in Figure 4.3-5.

![Figure 4.3-5 Focusing of a plane wave.](image)

Similarly, if $u = 2f$, we find that $v = 2f$. It is therefore possible to design a repetitive sequence of identical lenses, each spaced a distance $4f$ apart, which will continually focus and refocus a wave to points equally spaced between the lenses, as in Figure 4.3-6. We will return to this structure later on, because it represents one of the earliest attempts to guide an optical wave (known, with startling originality, as the **lens waveguide**).
IMAGE TYPES

Note that $v$ is positive in the examples above, since $f$ is positive and $u > f$. The images formed are therefore real ones, formed on the opposite side of the lens to the object. If $v$ is negative, however, the image is created on the same side as the object. It is then known as a virtual image. This occurs when $u < f$. However, a virtual image is always formed when $f$ is negative, i.e. for a diverging lens.

It is possible to extend the argument above, and show that a lens will also image a spherical wave originating from an off-axis point source onto an off-axis image point. It can be shown that Equation 4.3-12 is still valid in this case, which implies that the entire plane passing through $u$ is imaged as a plane passing through $v$. This simple theory (known as the Gaussian theory of image formation) accounts for the formation of extended images by a lens.

4.4 DIFFRACTION THEORY

We now consider the effect of the bounded nature of real beams. We base the argument on Huygens' principle, a powerful idea originally devised by Christiaan Huygens (1629-1695). Huygens' principle is founded on the reasonable argument that every point on a (possibly complicated) wavefront can be considered to act as a source of secondary spherical waves. The position of the wavefront at a later time can then be found by allowing these secondary waves to propagate forward, and combining their contributions in a new composite field. The principle works very well for many diffraction problems, and for determining the properties of beams. It is generally used to determine the amplitude distribution in a plane distant to another one, for which the distribution is already known. The necessary mathematics was initially developed by Augustin Fresnel (1788-1827), using the physical arguments outlined above. However, it was subsequently shown to have a valid basis by Robert Kirchhoff (1824-1887), as an approximate solution to the scalar wave equation. Since Kirchhoff's procedure is rather abstract, we shall ignore it here, and follow Huygens and Fresnel instead.

Suppose the amplitude distribution $A(x, y) = A(x, y, 0)$ is known in the $x$-$y$ plane, and we wish to determine the new distribution $A(\xi, \eta) = A(\xi, \eta, z)$ in a plane a distance $z$ away, as in Figure 4.4-1. We can find the answer by integrating the contributions made by all the elementary areas $dxdy$ to the field at $P$. The input weight attached to each such contribution is $A(x, y) dxdy$. However, as the contribution travels forward, it propagates as a spherical wave. At a distance $r$ from the origin, it will have changed both its phase and its amplitude, becoming $A(x, y) (1/r) \exp(-jkzr) dxdy$. 

![Figure 4.3-6 An iterative sequence of lenses.](image-url)
In addition, Fresnel found it was necessary to multiply each elementary contribution by a further scaling factor \( f(\theta)/j\lambda_0 \), where \( f(\theta) \) is a slowly-varying function of \( \theta \), defined by:

\[
\phi(\theta) = 1 \text{ for } \theta = 0; \quad f(\theta) = 0 \text{ for } \theta = \pi
\]

The function \( f(\theta) \), which was later justified rigorously by Kirchhoff, has the effect of removing all backward-propagating waves from the analysis. Integrating all the contributions, we then get:

\[
A(\xi, \eta) = \int \int [f(\theta)/j\lambda_0 r] A(x, y) \exp(-jk_0 r) \, dx \, dy
\]

where the integration is taken over the whole area of the \( x - y \) plane of interest - typically, \( x \) and \( y \) will each range from minus infinity to plus infinity.

Now, for small \( \theta \), we can put \( f(\theta) \approx 1 \). In addition, we assume that the effective range of \( A(x, y) \) is much less than \( z \), so we have a fairly narrow beam. In this case \( z \gg x, y, \xi \) or \( \eta \) so we can put:

\[
r = \sqrt{z^2 + (\xi - x)^2 + (\eta - y)^2}
\approx z + (\xi^2 + \eta^2)/2z + (x^2 + y^2)/2z - (\xi x + \eta y)/z
\]

This approximation can be used in the exponential contained in Equation 4.4-2. Now, there is also a \( 1/r \) amplitude term in Equation 4.4-2; however, since this function is so slowly-varying, a cruder approximation can be used in this case, and we may put \( 1/r \approx 1/\xi \). These simplifications are effectively paraxial approximations for the spherical wave terms.

Equation 4.4-2 then becomes:

\[
A(\xi, \eta) = (1/j\lambda_0 \xi) \exp[-jk_0 (z + \{\xi^2 + \eta^2\}/2z)] \cdot \int \int A(x, y) \exp[-jk_0 (x^2 + y^2)/2z - (\xi x + \eta y)/z] \, dx \, dy
\]

Equation 4.4-4 is known as the **diffraction integral**, and forms the basis of scalar diffraction theory. Once \( A(x, y) \) is known, the integration can be carried out and \( A(\xi, \eta) \) can be found. To evaluate such integrals, we distinguish two important cases known as Fraunhofer and Fresnel diffraction.
4.5 FRAUNHOFER DIFFRACTION

When the distance $z$ is much larger than the size of the object field, the term $k_0(x^2 + y^2)/2z$ is always much less than one, so the exponential $\exp[-jk_0(x^2 + y^2)/2z]$ is always approximately unity. When this condition is satisfied, we talk of Fraunhofer diffraction. The consequences are not only that the integral is simplified, but that $A(\xi, \eta)$ looks very like the Fourier transform of $A(x, y)$. This distant field is called the far-field. We can estimate the minimum distance $z$ that satisfies the criterion above, by setting the maximum allowable phase angle to about $\pi/8$. We then have:

$$k_0(x^2 + y^2)/2z \leq \pi/8$$

4.5-1

Bearing in mind that $k_0 = 2\pi/\lambda_0$, we may rearrange this condition as:

$$z \geq 8(x^2 + y^2)/\lambda_0$$

4.5-2

Putting in some numbers, we find that if the input beam is about 1 millimetre in diameter ($x = y = 10^{-3}\, \text{m}$), and the wavelength is 0.5 $\mu\text{m}$ (for green light), then $z$ must be greater than 32 m. This is clearly a large distance! However, if the approximation is valid, we can write:

$$A(\xi, \eta) = (1/j\lambda_0 z) \exp\{-jk_0 (z + \{\xi^2 + \eta^2\}/2z)\} \int \int A(x, y) \exp\{+jk_0 (\xi x + \eta y)/z\} \, dx \, dy$$

4.5-3

Given that the term outside the integral is mainly a phase factor, we have that:

$$\left| A(\xi, \eta) \right|^2 = (1/\lambda_0^2 z^2) \left| \int \int A(x, y) \exp\{+j(\omega_x x + \omega_y y)\} \, dx \, dy \right|^2$$

4.5-4

where $\omega_x$ and $\omega_y$ are given by:

$$\omega_x = k_0 \xi / z \text{ and } \omega_y = k_0 \eta / z$$

4.5-5

If $\omega_x$ and $\omega_y$ are interpreted as spatial frequencies, Equation 4.5-4 looks very like a Fourier transform relationship. A useful two-dimensional signal processing operation can therefore be carried out on an image simply by allowing an optical beam to propagate through the required distance, and allowing diffraction to do all the work. This is the origin of the much-quoted parallel processing capability of light. The large propagation distance involved is, unfortunately, a major drawback, which renders the whole process impractical without modification.

FOURIER TRANSFORMATION USING A LENS

A useful Fourier transform can, however, be carried out if a lens is introduced into the scheme, which effectively moves the far field to a convenient nearby point. Figure 4.5-1 shows the setup. We can find the effect of the lens using the transfer function approach previously described. It is first assumed that an input amplitude distribution $A'(x, y)$ is prepared just behind the lens. This might be done, by passing a plane wave through a suitable transparency.

After passing through the lens, this distribution is modified to $A(x, y)$, given by:

$$A(x, y) = A'(x, y) \tau \exp[+jk_0(x^2 + y^2)/2f]$$

4.5-6
If Equation 4.5-6 is now substituted into Equation 4.5-4, we obtain:

\[ |A(\xi, \eta)|^2 = \left( \frac{1}{\lambda_0^2 z^2} \right) \int \int_A A'(x, y) \exp\left[ -jk_0 \left( \frac{x^2 + y^2}{2f} - \frac{\xi x + \eta y}{z} \right) \right] \, dx \, dy \]  

When \( z = f \), the first part of the exponent cancels, and we are left with:

\[ |A(\xi, \eta)|^2 = \left( \frac{1}{\lambda_0^2 z^2} \right) \int \int_A A'(x, y) \exp\left[ +jk_0 (\xi x + \eta y)/z \right] \, dx \, dy \]  

Equation 4.5-8 shows that we can obtain the far-field distribution at a distance \( f \) from the lens, in a plane known as the **Fourier transform plane**. This arrangement is much more convenient, and is often used in signal processing systems. Note, however, that the result is not an exact transform, because of the presence of extra phase terms outside the integral in Equation 4.5-3. This is often unimportant (especially if we are only interested in intensity variations), but if an exact transform is essential it can be obtained by a further simple rearrangement of the optical set-up.

![Figure 4.5-8 Optical Fourier transform using a lens.](image)

**4.6 FRESNEL DIFFRACTION AND GAUSSIAN BEAMS**

When the far-field condition does not hold, we speak instead of **Fresnel diffraction**. As an example application of scalar diffraction theory, we shall consider what happens if we have a **Gaussian amplitude distribution** in the \( x - y \) plane, defined by:

\[ A(x, y) = A_0 \exp\left( -\frac{x^2 + y^2}{a^2} \right) \]  

The amplitude distribution in a plane a distance \( z \) away is found by substituting this function into Equation 4.4-4, a process which gives:

\[ A(\xi, \eta) = \left( \frac{1}{j\lambda_0 z} \right) \exp\left[ -jk_0 \left( z + \frac{\xi^2 + \eta^2}{2z} \right) \right] \int \int_A A_0 \exp\left( -\frac{x^2 + y^2}{a^2} \right) \exp\left[ -jk_0 \left( \frac{x^2 + y^2}{2z} - \frac{\xi x + \eta y}{z} \right) \right] \, dx \, dy \]  

Equation 4.6-2 looks extremely complicated. However, because the variables are separated, it is actually easier to evaluate than it appears. We first note that \( A(\xi, \eta) \) can be written in the form:

\[ A(\xi, \eta) = \left( \frac{1}{j\lambda_0 z} \right) \exp\left[ -jk_0 \left( z + \frac{\xi^2 + \eta^2}{2z} \right) \right] I_1 \, I_2 \]  

**R.R.A.Syms and J.R.Cozens**

**Optical Guided Waves and Devices**

12
where $I_1$ and $I_2$ are two new integrals, defined by:

\[
I_1 = \int_x \exp\left[-(jk_0/2z + 1/a^2)x^2\right] \exp[jk_0 \xi x/z] \, dx
\]

\[
I_2 = \int_y \exp\left[-(jk_0/2z + 1/a^2)y^2\right] \exp[jk_0 \eta y/z] \, dy
\]

and where $x$ and $y$ each range from minus infinity to plus infinity. Because of the similar appearance of $I_1$ and $I_2$, we need only evaluate one of them, say $I_1$. We now make the substitutions:

\[
b = jk_0/2z + 1/a^2 \quad \text{and} \quad c_1 = jk_0 \xi /z
\]

which reduces $I_1$ to:

\[
I_1 = \int_x \exp[-bx^2 + c_1 x] \, dx
\]

We then follow with the additional variable changes:

\[
p = x \sqrt{b} - c_1/2\sqrt{b} \quad \text{and} \quad dp = \sqrt{b} \, dx
\]

which transform $I_1$ to:

\[
I_1 = (1/\sqrt{b}) \exp(c_1^2/4b) \int_p \exp(-p^2) \, dp
\]

where $p$ now ranges from minus infinity to plus infinity. Since:

\[
\int_p \exp(-p^2) \, dp = \sqrt{\pi/b}
\]

we then obtain:

\[
I_1 = \sqrt{(\pi/b)} \exp(c_1^2/4b)
\]

The $y$-integral $I_2$ may be evaluated in a similar way. If we define:

\[
c_2 = jk_0 \eta /z
\]

we get:

\[
I_2 = \sqrt{(\pi/b)} \exp(c_2^2/4b)
\]

In this way, we can obtain an expression for the field distribution $A(\xi, \eta)$ in the form:

\[
A(\xi, \eta) = \left(\frac{A_0}{j\lambda_0 z}\right) \frac{2\pi a^2}{(jk_0^2 + 2z)} \exp(-jk_0 z) \cdot \exp\left[-(k_0/2z)(j + k_0 a^2/(jk_0 a^2 + 2z))(\xi^2 + \eta^2)\right]
\]

Equation 4.6-13 still looks rather complicated. However, we note that, for any output plane, $z$ is a constant, so the first part of the expression must also be constant. Hence, we can write:

\[
A(\xi, \eta) = A(z) \exp\left[-(k_0/2z)(j + k_0 a^2/(jk_0 a^2 + 2z))(\xi^2 + \eta^2)\right]
\]
Here \( A(z) \) describes the \( z \)-variation of the amplitude, while the information about the amplitude distribution in the \( \xi - \eta \) plane is contained in a much simpler exponential. We now notice that the coordinates \( \xi \) and \( \eta \) only appear together, in the form \( f(\xi, \eta) = (\xi^2 + \eta^2) \), so the amplitude distribution must be circularly symmetric. If we put \((\xi^2 + \eta^2) = R^2\), we get:

\[
A(R) = A(z) \exp \left[ -(k_0/2z) \left\{ j4z^2/(4z^2 + k_0^2a^4) + 2k_0za^2/(4z^2 + k_0^2a^4) \right\} R^2 \right]
\]

4.6-15

We shall use this expression to examine the basic properties of Gaussian beams.

**AMPLITUDE VARIATIONS**

The real part of the exponential in Equation 4.6-15 must describe the amplitude distribution of \( A(R) \), while the imaginary part gives its phase variation. For the real part, we have:

\[
|A(R)| = |A(z)| \exp[-k_0^2a^2R^2 / (4z^2 + k_0^2a^4)]
\]

4.6-16

Since this still has the form of a Gaussian function, the overall shape of the beam must still be Gaussian, even after it has travelled the distance \( z \) from the \( x-y \) plane to the \( \xi-\eta \) plane. If we define the **beam width** \( w \) as the radius at which the amplitude falls to \( 1/e \) of its peak value, we get:

\[
w^2 = (4z^2 + k_0^2a^4) / k_0^2a^2
\]

4.6-17

Equation 4.6-17 shows that the beam width varies parabolically with \( z \). We can examine this behaviour in more detail, in two important regions:

a) **In the near field**, when \( 4z^2 \ll k_0^2a^4 \), we have \( w^2 \approx a^2 \). This implies that the beam has constant width, independent of \( z \). We can therefore usefully define \( w_0 = a \) as the radius of the beam waist.

b) **In the far field**, when \( 4z^2 \gg k_0^2a^4 \), we have \( w^2 \approx 4z^2 / k_0^2a^2 \). Here the beam width varies linearly with \( z \). In this region, we can define a **half-angle of divergence** \( \theta \), such that:

\[
\theta = w/z = 2/k_0w_0
\]

4.6-18

We can also define a **whole-angle of divergence** \( \Theta \) as:

\[
\Theta = 2\theta = 4\lambda_0/\pi D
\]

4.6-19

where \( D \) (the beam diameter at its waist) equals \( 2w_0 \). We can summarise the results above in normalised units. Suppose we define the transition between one regime and the other as the point \( z = z_0 \). This occurs when:

\[
4z_0^2 = k_0^2a^4 \quad \text{or} \quad z_0 = k_0w_0^2/2
\]

4.6-20

Equation 4.6-17 can then be rewritten as:

\[
(w/w_0)^2 = 1 + (z/z_0)^2
\]

4.6-21

The normalised beamwidth \( w/w_0 \) therefore depends parabolically on the normalised distance \( z/z_0 \). This is shown in Figure 4.6-1, where \( w/w_0 \) is plotted as a function of \( z/z_0 \), for both positive and negative \( z/z_0 \). In the far field, the variation tends to the linear dashed-line asymptote shown. Note that, as \( w_0 \) increases, \( z_0 \) increases and \( \theta \) decreases. A wide beam
(measured in wavelengths) therefore expands slowly, and travels a longer distance before starting to expand.

![Figure 4.6-1 Parabolic variation of beamwidth with distance.](image)

**DESIGN EXAMPLES**

We can illustrate this with some examples. A typical visible laser has the parameters $\lambda_0 = 0.633 \ \mu m$ and $w_0 \approx 0.3 \ mm$. This corresponds to a helium-neon laser emitting red light. In this case, the beam is roughly 1000 wavelengths wide, and we find that $z_0 = 0.45 \ m$ ($\approx 700,000$ wavelengths) and $\theta = 0.67 \times 10^{-3} \ rad$. Similarly, a microwave dish might have $\lambda_0 = 1 \ cm$ and $w_0 = 10 \ cm$ (so the beam is 20 wavelengths wide), which gives $z_0 = 3.14 \ m$ ($\approx 300$ wavelengths) and $\theta = 31.8 \times 10^{-3} \ rad$.

**PHASE VARIATIONS**

Now we return to the complex part of the exponential in Equation 4.6-15, which determines the phase of the Gaussian beam. The relevant part is the term:

$$A'(R) = \exp[-j2k_0zR^2 / (4z^2 + k_0^2a^4)]$$

4.6-22

This can be rewritten using our normalised variables as:

$$A'(R) = \exp[-jkR^2 / 2(z^2 + z_0^2)]$$

4.6-23

In any plane $z = \text{constant}$, the phase therefore varies parabolically with the distance $R$ from the axis. This is very similar to the behaviour previously demonstrated by a paraxial spherical wave. In fact, Equation 4.6-23 can be written in the alternative form:

$$A'(R) = \exp[-jkR^2/2r(z)]$$

4.6-24

Equation 4.6-24 should be compared with Equation 4.2-17. It shows that the effective radius of curvature $r(z)$ of the phase-front is given by:

$$r(z) = (z^2 + z_0^2) / z$$

4.6-25

We can now work out the local radius of curvature a number of important points along the beam.

a) In the near field (when $z \approx 0$) we find that $r \approx \infty$. The phasefront is therefore approximately plane, as might be expected from symmetry.
b) At the transition point (when \( z = z_0 \)) we find that \( r = 2z_0 \).

c) In the far-field (when \( z \to \infty \)) we see that \( r \to z \). The phasefront is therefore roughly spherical, and the field is equivalent to spherical wave emanating from a point source at \( z \).

These results are summarised in Figure 4.6-2. At the left, the wavefront is spherical and the beam is converging. As the beam passes through its waist, the phasefront slowly becomes plane. At the right, the wavefront gradually becomes spherical once more, but now the beam is diverging.

![Figure 4.6-2](image)

**DIFFRACTION OF ARBITRARY FIELD DISTRIBUTIONS**

We have seen how the width of a Gaussian beam spreads as we move away from the beam waist. This illustrates an important general point: an arbitrary bounded beam will, in general, spread through diffraction. Furthermore, the more the beam is confined, the worse this effect becomes. For example, Figure 4.6-3 shows a 'top-hat' intensity distribution, which rapidly broadens and loses its well-defined shape as it propagates, with a consequent reduction in power density. Efficient line-of-sight optical communication is therefore not practical over long distances. An exception to this general rule is the recently discovered set of **diffraction-free beams**. These have rather more complicated beam shapes, but can (in theory) propagate without any spreading at all. At the moment, it is too early to say if any practical use can be made of this feature.

![Figure 4.6-3](image)

**4.7 TRANSFORMATION OF A GAUSSIAN BEAM BY LENSES**

From the previous section, we know that the main features of a Gaussian beam are a Gaussian amplitude distribution and a spherical phase-front. We shall now see what
happens when such a beam is passed through a lens. Clearly, the amplitude distribution will be unaltered. Equally, the emerging wave will still be spherical, because the lens can only introduce a quadratic phase factor. However, the phase curvature will be modified, because the spherical wave will have different parameters. A lens can therefore only transform a Gaussian input beam into a Gaussian output. We shall now see how this feature can be exploited to control the parameters of a beam.

TRANFORMATION BY A SINGLE LENS

Firstly, we shall consider the transformation of a Gaussian beam by a single lens. Figure 4.7-1 shows the geometry. The input beam has a waist of width $2w_{10}$ located at a distance $u$ behind the lens.

![Figure 4.7-1 Transformation of a Gaussian beam by a lens.](image)

It is simplest to assume initially that $u \gg w_{10}$, so that the lens is in the far field of the beam. If this is the case, we know from Equation 4.6-25 that the radius of curvature of the phase-front at the lens is roughly $u$. Similarly, from Equation 4.6-17, we know that the beam width at the lens is $2w$, where:

$$w \approx \frac{2u}{k_0} w_{10}$$  \hspace{1cm} 4.7-1

From the thin-lens imaging equation, Equation 4.3-12, we can find the radius of curvature of the emerging beam as $v$. If we adopt a far-field approximation once again, we can assume that the new beam waist will be located at a distance $v$ in front of the lens. If the new waist has width $2w_{20}$, we know that $w_{20}$ is defined (by analogy with Equation 4.7-1) by:

$$w \approx \frac{2v}{k_0} w_{20}.$$  \hspace{1cm} 4.7-2

Combining Equations 4.7-1 and 4.7-2, we can relate $w_{20}$ to $w_{10}$ using:

$$w_{20} / w_{10} = v / u.$$  \hspace{1cm} 4.7-3

If $v < u$, the new waist will be smaller than the old. We can illustrate this with examples.

1) FOCUSSING OF A GAUSSIAN BEAM.

First we consider the focusing of a plane beam with a Gaussian envelope, as shown in Figure 4.7-2. In this case, we assume that the input beam is approximately parallel, with a Gaussian amplitude distribution of width $2w$ at the lens.
From the thin-lens imaging equation, we know that \( v = f \) for a plane input. In this case, we obtain:

\[
\frac{w_{20}}{180} = \frac{2f}{\lambda_0 w}
\]

Thus, the wider the input beam, the smaller the size to which it can be focused. For the typical parameters of \( f = 1 \text{ cm}, \lambda_0 = 0.633 \mu\text{m} \) and \( w = 1 \text{ mm} \), we obtain \( w_{20} = 2 \mu\text{m} \), a respectably small figure. This shows that a Gaussian beam can be focused down to small dimensions. However, because the beam is bounded, the energy density at focal spot remains finite. Generally, the lens diameter is roughly matched to the expected input beam width, and a lens with a short focal length compared with its diameter is then said to have a high **numerical aperture** or NA. For a medium of refractive index \( n_e \) external to the lens, this quantity is defined as:

\[
\text{NA} = n_e \sin(\theta_{\text{max}})
\]

where \( \theta_{\text{max}} \) is the maximum angle of the converging cone of rays (e.g. the largest value of \( \theta_2 \) in Figure 4.3-3).

Naturally, there is a trade-off between parameters. For example, the wider the beam, the larger the angle of convergence, and hence the smaller the extent of the focused region. This has important consequences for the design of electro-optic modulators (see Chapter 3). The voltage needed to drive a transverse modulator depends on the electrode gap, and if the beam has a small diameter, this gap can be small. As we have seen, a lens can be used to obtain a small beam diameter; however, if the depth of focus is very small, the usable interaction length is very short, and this in turn increases the drive voltage. To make a more useful modulator, more effective confinement of the beam is needed.

2) TELESCOPE BEAM EXPANSION

We now consider an application that involves the use of two lenses. When arranged as a **telescope**, these can be used to carry out beam expansion. Figure 4.7-3 shows the geometry. The input beam is again assumed to be approximately parallel, and of width \( 2w \) at the lens.

If the focal length of the first lens is \( f_1 \), then the focal spot is defined as before by \( w_{20} = 2f_1/\lambda_0 w \). However, if the second lens has focal length \( f_2 \), then \( w_{20} = 2f_2/\lambda_0 w' \). Equating the two, we can find the output beam width \( w' \) as:

\[
\frac{w_{20}}{180} = \frac{2f}{\lambda_0 w} = \frac{2f_2}{\lambda_0 w'}
\]
\[ w' = w \left( \frac{f_2}{f_1} \right) \]

The output beam is therefore an expanded version of the input, enlarged by the ratio of the lens focal lengths. This reduces any residual beam divergence by the same factor. In many cases, the input beam is not quite Gaussian, but suffers from imperfections caused by noise of high spatial frequency. These can be removed by placing a pinhole at the focus between the lenses. Since the field distribution in this plane is closely related to a Fourier transform of the input distribution, blocking the outermost regions removes the high frequency components and can greatly improve the beam quality. This process is known as \textit{spatial filtering}.

![Figure 4.7-3 A telescope beam expander.](image)

3) \textbf{THE LENS WAVEGUIDE}

One solution to the problem of confining a beam within a narrow cross-section over a long distance is provided by the iterative sequence of lenses, which we discussed earlier in Figure 4.3-6 in terms of simple imaging theory. If we now assume that the waveform propagating between each lens is a Gaussian beam, it is clear that the beamwidth must vary through the structure much as shown in Figure 4.7-4. Note that the separation between lenses is now taken as \(4s\), rather than \(4f\).

![Figure 4-18 The lens waveguide.](image)

We shall calculate the parameters of the Gaussian beam propagating through the guide, assuming this time that the focal spot is not small, so that a far-field approximation cannot be used. In this case, the thin-lens imaging equation 4.3-12 is invalid. However, a suitable modification can be obtained simply by replacing \(u\) and \(v\) with the radii of curvature of two Gaussian beams, one originating from a waist at \(u\), the other converging to a waist at \(v\). If these are denoted \(u'\) and \(v'\), we get:
CHAPTER FOUR

\[ \frac{1}{u'} + \frac{1}{v'} = \frac{1}{f} \quad 4.7-7 \]

For the geometry shown in Figure 4.7-4, \( u' = v' \), so that \( u' = 2f \). However, Equation 4.6-25 implies that the radius of curvature can also be found as:

\[ u' = \frac{u^2 + z_0^2}{u} \quad 4.7-8 \]

Where \( u = 2s \) and \( z_0 \) is the transition distance of the beam. Hence, \( z_0 \) is given by:

\[ z_0 = \sqrt{u' - u} = 2\sqrt{fs - s^2} \quad 4.7-9 \]

For \( z_0 \) to be real, we require \( s < f \). If this is the case, we may invert Equation 4.6-20 to give:

\[ w_0 = \sqrt{2z_0/k_0} \quad 4.7-10 \]

Substituting from Equation 4.7-9, we then get:

\[ w_0 = \left[ \frac{16(fs - s^2)}{k_0^2} \right]^{1/4} = \left[ \frac{4\lambda_0^2 (fs - s^2) / \pi 2}{k_0^2} \right]^{1/4} \quad 4.7-11 \]

Equations 4.7-9 and 4.7-11 therefore imply that the parameters \( z_0 \) and \( w_0 \) both depend strongly on the separation of the lenses.

DESIGN EXAMPLE

For a lens waveguide formed by a sequence of lenses of focal length 100 mm, and light of wavelength 0.633 \( \mu m \), we obtain \( w_0 = 0.14 \) mm for \( s = 50 \) mm and \( w_0 = 53.3 \) \( \mu m \) for \( s = 99.5 \) mm. Note that as \( s \) tends to \( f \), \( w_0 \) and \( z_0 \) both tend to zero, and the solution reverts to the geometrical optics approximation discussed previously.

The lens waveguide can support a number of other modes of propagation, besides the simple Gaussian beam described above. Unfortunately, it is hard to make such a guide work well in practise. For example, the positioning of the lenses is very critical, since any lateral misalignment causes the beam to wander off-axis; equally, the air/glass interfaces cause considerable scatter and loss. Luckily, there are considerably better types of guide, as we shall see later. However, some of these (parabolic index fibres) can also support a Gaussian transverse field distribution.

4.8 MIRRORS AND RESONATORS

We now consider the use of an alternative optical component, the mirror. This often consists of a smooth, polished surface coated with metal. It is common experience that light does not penetrate such a surface to any significant extent, but is entirely reflected instead. Similarly, it is well known that the angle of reflection is equal and opposite to the angle of incidence, as shown in Figure 4.8-1 (both these assumptions will be fully justified in Chapter 5).

Mirrors clearly have many applications, but for the time being we shall consider just two of them. The first example is the Fabry-Perot cavity, which is a resonant structure formed from two plane, parallel mirrors. When filled with an optical gain medium, the device can oscillate, and thus forms the basis of the laser. The second is the spherical mirror. This can be used to transform one spherical wave front into another, and so (just like a lens) it can perform imaging operations. It turns out that a laser cavity can be greatly improved by the use of at least one spherical end-mirror, since this confines the beam inside the cavity.
more effectively. When this is done, the laser output is generally a Gaussian beam, which is one reason why such beams are so common in the first place.

Figure 4.8-1 Light incident on a polished, metallized surface.

THE FABRY-PEROT CAVITY

Figure 4.8-2 shows a Fabry-Perot cavity, consisting of two plane, parallel mirrors with reflectivities $R_1$ and $R_2$, separated by a distance $L$.

We start by considering the behaviour of the device when the two mirrors are perfect (so that $R_1 = R_2 = 1$) and the cavity is filled with a dielectric of refractive index $n$. We first assume that there is a plane wave propagating up and down the cavity, following the zig-zag path shown. We now consider the resulting field amplitude at a particular point $A$. Suppose the wave is initially travelling in the $+z$ direction, and that its amplitude is given by $E_A$. By the time it has completed one round-trip of the cavity the new field $E_A'$ is found as:

$$E_A' = E_A \exp(-j2k_0 n L)$$

Clearly, the combined field travelling in the $+z$ direction at $A$ is found as a summation of many such contributions. These will add up in-phase if:

$$2k_0 n L = 2\nu\pi$$

Where $\nu$ is an integer. If Equation 4.8-2 does not hold, on the other hand, the field will consist of many contributions with different phase, which will add up to zero on average. Rearranging Equation 4.8-2 slightly, we find that the condition for a non-zero field inside the cavity is that:

$$L = \nu \left(\frac{\lambda_0}{2n}\right)$$
This is known as the **longitudinal resonance condition**, and it implies that only a certain set of field patterns can exist inside the cavity. Another way to calculate it is to assume that the total field travelling in both directions must everywhere be given by the sum of one combined field travelling in the +z direction and a similar field travelling in the -z direction. Thus we could write:

\[
E(z) = E_+ \exp(-jk_0nz) + E_- \exp(+jk_0nz)
\]

Now, if the mirrors are perfect, light cannot penetrate them, and the boundary conditions to be satisfied are \(E = 0\) at \(z = 0\) and at \(z = L\) (again, these assumptions will be reviewed in Chapter 5). We can satisfy the former condition by taking \(E_- = -E_+\), which allows us to write Equation 4.8-4 as:

\[
E(z) = E_0 \sin(k_0nz)
\]

Clearly, the latter condition is satisfied when \(k_0nL = \nu\pi\), which is effectively the same as the longitudinal resonance condition previously found in Equation 4.8-3. Note, however, that this time we have also calculated the actual field distribution inside the cavity, which must be a standing-wave pattern as shown in Figure 4.8-3. Each pattern of this type is called a **longitudinal mode**.

![Figure 4.8-3 Standing wave patterns inside the cavity.](image)

**THE LASER**

We now consider the case when there are losses, so that \(R_1, R_2 < 1\). This could arise if power is simply dissipated in the mirrors, or if we have arranged one of the mirrors to be only partially reflecting. In this case, the amplitude of the wave trapped inside the cavity will swiftly fall to zero, as it suffers repeated reflections. If the cavity is filled with an **optical gain medium**, however, these losses may be made up, and the wave can actually be amplified as it passes back and forth.

If the gain medium provides a gain \(g\), then a forward-travelling wave will grow as defined by the exponential term \(\exp(+gz)\). Including this gain, and the mirror loss, we can now find the field \(E_A'\) in terms of the field \(E_A\) as:

\[
E_A' = E_A R_1 R_2 \exp(+2gL) \exp(-j2k_0nL)
\]

Clearly, the longitudinal resonance condition is exactly as before. However, we note that \(|E_A'| > |E_A|\), provided:

\[
R_1 R_2 \exp(+2gL) > 1
\]
In this case, the wave will continually grow in amplitude as it travels up and down the cavity, so the total field will become infinite (or, at least, extremely large, before gain saturation sets in). The cavity then acts as an oscillator, known as a laser, and (as with nearly all such devices) the oscillation process can start simply from noise. Equation 4.8-7 is therefore the condition for laser operation.

We can tap off some of the stored energy by making one of the mirrors partially reflecting. In this case, the output beam will generally contain a mixture of a number of different wavelengths, or lines, each corresponding to a particular longitudinal mode. We can calculate the frequencies concerned by first rearranging Equation 4.8-3 to give the wavelength of the $\nu^{th}$ mode as:

$$\lambda_{\nu} = \frac{2L}{n}$$

4.8-8

Since frequency and wavelength are related to the velocity of light by $c = \frac{\lambda_{\nu}}{\nu}$, we can find the frequency of the $\nu^{th}$ mode as:

$$f_{\nu} = \frac{c \nu}{2Ln}$$

4.8-9

So the frequency separation between adjacent modes is given by:

$$\Delta f = \frac{c}{2Ln}$$

4.8-10

We can now plot the frequency content of the laser output (known as the emission spectrum) as shown in Figure 4.8-4. For a 1 metre long cavity with $n \approx 1$, the mode spacing is $\Delta f \approx 150$ MHz. In practise, the exact number of modes that are observed is limited by the existence of optical gain only over a finite frequency range, or gain bandwidth. This result is indicated by the dashed-line envelope in Figure 4.8-4. However, the number of longitudinal modes experiencing sufficient gain to lase may still be large. The Fabry-Perot laser is therefore a multimode device, unless special efforts are made to select a particular line using additional optical components.

![Figure 4.8-4 Emission spectrum of a Fabry-Perot laser.](image)

THE SPHERICAL MIRROR RESONATOR

We have previously found the field distribution inside the cavity as a standing wave pattern. Of course, this only represents a longitudinal variation, and we now consider the field distribution in a transverse plane and its relation to the construction and operation of the laser. Now, the optical gain will generally be provided by electrical pumping. The power needed for this will be proportional to the volume of active material, which can be reduced by making the cavity cross-section small. To achieve adequate round-trip gain, however, the
cavity length \( L \) must be quite large. These specifications are conflicting, since they require a narrow, bounded beam to travel for long distances. Clearly, the beam will spread laterally en route, as we have previously shown, so energy will leak sideways out of the cavity. Once more, therefore, we need a way of confining a bounded beam.

One answer is provided by the \textit{spherical mirror}. If correctly located, this can exactly retroreflect a spherical wave, as shown in Figure 4.8-5. Here the radius of curvature of the mirror is \( r \), and the origin of the spherical wave is located on-axis, at a distance \( r \) from the mirror.

![Figure 4.8-5 Retroreflection of a spherical wave by a spherical mirror.](image)

Now, remember that the phase-front of a Gaussian beam is spherical. This type of beam can therefore also be reflected back on itself using a spherical mirror with the correct curvature. If two mirrors are used, an optical cavity is formed. This allows the beam to travel back and forth between the mirrors as shown in Figure 4.8-6, loosing very little energy out of the sides of the cavity.

![Figure 4.8-6 The symmetric spherical mirror resonator.](image)

This arrangement is termed a \textit{symmetric spherical mirror resonator}, and the Gaussian beam is one of its possible \textit{transverse oscillation modes}. It turns out that there are other transverse modes, but the Gaussian is the lowest-order one, with most of its energy concentrated near the axis. We can calculate the parameters of such a beam, using analysis...
very similar to that previously employed for the lens waveguide. Assuming the cavity length is $L$, the radius of curvature of the beam at each mirror can be found from Equation 4.6-25 as:

$$r = \frac{[(L/2)^2 + z_0^2]}{(L/2)}$$  

4.8-11

Where (once again) $z_0$ is the transition length of the beam. For the beam to be correctly retroreflected, the radius of curvature of the beam and the mirror must match, so that $z_0$ is given by:

$$z_0 = \sqrt{[(L/2) (r - L/2)]}$$  

4.8-12

As with the lens waveguide, only a certain combination of parameters will give a real value of $z_0$. In this case, we require $L/2 < r$. If this is the case, the width $w_0$ of the beam can be found as:

$$w_0 = \frac{4 \left[ r L/2 - (L/2)^2 \right]}{k_0^2} = \frac{\lambda_0^2 \left[ r L/2 - (L/2)^2 \right]}{\pi^2}$$  

4.8-13

The beamwidth is therefore dependent on the cavity length and on the mirror curvature. For a cavity of length $L = 0.5$ m, fitted with end mirrors of radius 0.5 m, and a wavelength of 0.633 µm, we obtain $w_0 = 0.225$ mm. This corresponds to a beam diameter of 0.45 mm.

Naturally, there are other possible ways of arranging a resonator of this type. One common method is to use one spherical and one plane mirror, in an asymmetric configuration. The beam waist now occurs at the plane mirror; if this is made partially reflecting, an output can be obtained from the cavity. This arrangement is used in many gas laser types, including the ubiquitous helium-neon laser, shown in Figure 4.8-7. The optical gain is provided by a gas plasma tube, which contains an electrically excited gas mixture. The ends of the tube are sloped at the Brewster angle; as we shall see in Chapter 5, this allows one polarization to have greatly reduced reflection loss, through the Brewster effect. Consequently, it has a lower lasing threshold and the output of the laser is linearly polarized - this provides a convenient way of generating polarized light. An aperture is also introduced into the cavity. This allows the desired Gaussian beam to pass with little loss, but heavily attenuates higher-order transverse modes, which have more energy in off-axis regions of the beam.

![Figure 4.8-7 Schematic of a typical HeNe laser.](image)

Note that an entirely different solution to the problem of confining a narrow beam inside a region of optical gain is provided by a waveguide structure. This solution is adopted in all semiconductor lasers, and will be discussed in detail in Chapter 12.
PROBLEMS

4.1 Find the focal length of a converging glass lens, having (a) two spherical surfaces, each of radius 200 mm, and (b) one plane surface, and one spherical surface of radius 400 mm.
[(a) 200 mm; (b) 800 mm]

4.2 The lenses in Question 4.1 are used in turn to form an image of a source positioned at a point 400 mm away. Find the image position in each case. What is the difference between the image types?
[(a) 400 mm away, on the opposite side of the lens; (b) 800 mm away, on the same side]

4.3 A source is positioned at point S, a distance δ off axis, u away from a converging lens of focal length f. Find the position of the image. Hence show that the lens will form a plane image of a plane, two-dimensional object, and find the magnification involved.

4.4 A sequence of identical lenses, each of focal length f, are separated by alternating distances d₁ and d₂. Find a relation between d₁ and d₂ that must be satisfied if the sequence is to act as a lens waveguide.

4.5 The figure below shows a plane monochromatic wave incident normally on an opaque screen, which has a central rectangular aperture of sides a and b in the x- and y-directions, respectively. Calculate and sketch the intensity distribution of the resulting diffraction pattern in the far field.

4.6 The figure below shows a plane monochromatic wave incident normally on a periodic structure known as an amplitude grating. This is a screen of locally-varying transmission, defined by the transparency function \( \tau = \tau_0 + \Delta \tau \cos(Ky) \), where \( K = \frac{2\pi}{\Lambda} \) is the spatial frequency of the periodic modulation in the transmission. What is the output of the grating in the far field?
4.7 Show that an exact two-dimensional Fourier transform can be carried out between the input distribution $A_m$ and the output field $A_{out}$ using the arrangement below.

4.8 (a) Find the half-angle of divergence of a Gaussian beam of wavelength 0.633 µm and diameter 1 mm. (b) How far from the beam waist would you expect to find the transition point between plane and spherical wave behaviour?
\[(a) 4 \times 10^{-4} \text{ rad}; \text{ (b) 1.24 m}\]

4.9 A gas laser is to be constructed with an optical cavity of length 250 mm and end-mirrors of amplitude reflectivity $R_1 = 1.0$ and $R_2 = 0.9$. Determine the threshold gain required from the lasing medium. Estimate the frequency difference between adjacent longitudinal modes.
\[0.211 \text{ m}^{-1}; \text{ 600 MHz}\]

4.10 The optical cavity of a helium-neon laser (for which $\lambda_0 = 0.633 \mu$m) is to be constructed from one spherical mirror of radius 2 m and one plane, partially-reflecting mirror. Describe steps that could be taken to ensure the output is a Gaussian beam. If the two mirrors are 1 m apart, what will be the diameter of the beam?
\[0.9 \text{ mm}\]
SUGGESTIONS FOR FURTHER READING

Hecht E. "Optics" 2nd Edn., Addison-Wesley Publishing Co., Reading MA (1987);
   Chapters 5, 10 & 11.
Marcuse D. "Light transmission optics" Van Nostrand Reinhold Co., New York (1972);
   Chapters 4, 5 & 6.
Steward E.G. "Fourier optics", Ellis Horwood Ltd., Chichester (1983)
CHAPTER
FIVE

REFLECTION AND REFRACTION AT A SINGLE INTERFACE

5.1 THE BEHAVIOUR OF LIGHT AT A DIELECTRIC INTERFACE

The previous Chapters have been concerned with the propagation of waves in empty space or in uniform, homogeneous media. It is now time to introduce more complicated geometries, building towards those appropriate to optical waveguides. We begin with a qualitative discussion of the phenomena that occur when light strikes a dielectric interface. Figure 5.1-1 shows a plane wave incident at an angle $\theta_1$ on two semi-infinite media, with refractive indices $n_1$ and $n_2$, respectively. These are chosen such that $n_1 > n_2$.

Experience suggests that for near-normal incidence (small $\theta_1$) the incident beam will give rise to a plane reflected beam in Medium 1, together with a transmitted beam in Medium 2. The reflected beam travels at an angle equal and opposite to that made by the incident wave with the interface normal. The angle of the transmitted beam is defined by Snell's law (named after Willebrord Snell, 1591-1626). This was originally an experimental observation, and requires that:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$$  \hspace{1cm} 5.1-1

However, we note that for $\theta_1 = \sin^{-1}(n_2/n_1)$, we get $\theta_2 = \pi/2$. In this case, the transmitted wave travels parallel to the interface, as shown in Figure 5.1-2.

This angle of incidence is called the critical angle, $\theta_c$. For an interface between glass and air (when $n_1 = 1.5$ and $n_2 = 1.0$), the critical angle is $\theta_c = 41.8^\circ$. However, $\theta_c$ can never reach $\pi/2$ if we exchange the values of $n_1$ and $n_2$, so there is no critical angle when the wave is
incident from the low-index side. For $\theta_1 > \theta_c$, however, there is no real solution for $\theta_2$, which implies that a propagating transmitted wave cannot arise at all. This is known as total internal reflection, and is illustrated in Figure 5.1-3.

![Figure 5.1-3 Total internal reflection.](image)

It is easy to show that when total internal reflection occurs, no power crosses the interface. Total internal reflection therefore presents a mechanism for confining a field in one region of space. We shall now proceed to a more rigorous analysis of the problem. The first point to note is that the presence of two waves (one reflected, one transmitted) in Medium 1 implies that there must be a more complicated field distribution in this region. We will therefore start by examining the way plane wave solutions to Maxwell's equations can be combined. Since the equations are linear, this is very simple.

### 5.2 SUPERPOSITION OF FIELDS

Perfectly monochromatic fields can produce a range of strikingly beautiful effects, which fall under the general heading of interference phenomena. The simplest of these occurs when two plane waves of the same frequency cross at an angle, as shown in Figure 5.2-1.

![Figure 5.2-1 Interference between two plane waves.](image)

We assume that both waves are polarised in the y-direction, travel in free space, and have equal amplitude $E_y$, so their time-independent electric fields are given by:

\[
E_1 = j E_y \exp[-jk_0(z \cos(\theta) + x \sin(\theta))]
\]

5.2-1

for the upward-travelling wave, and:

\[
E_2 = j E_y \exp[-jk_0(z \cos(\theta) - x \sin(\theta))]
\]

5.2-2

for the downward one. Because Maxwell's equations are themselves linear, the total field can be written as a linear superposition of these two fields, in the form:

\[
E = j E_y \{ \exp[-jk_0(z \cos(\theta) + x \sin(\theta))] + \exp[-jk_0(z \cos(\theta) - x \sin(\theta))] \} 
\]
\[ = j \ 2E_y \exp[-jk_0z \cos(\theta)] \cos[k_0x \sin(\theta)] \]  

5.2-3

Here, we have combined the x-dependent parts of the two fields to yield a real, cosinusoidal amplitude distribution. We can write Equation 5.2-3 in the alternative form:

\[ E = j \ E(x) \exp(-j\beta z) \]  

5.2-4

where:

\[ E(x) = 2E_y \cos[(k_0x \sin(\theta))] \quad \text{and} \quad \beta = k_0 \cos(\theta) \]  

5.2-5

The result can therefore be regarded as a field with non-uniform amplitude \( E(x) \), travelling in the z-direction. However, its propagation constant is reduced from the normal value by the factor \( \cos(\theta) \).

**DESIGN EXAMPLE**

As an exercise, we shall now calculate the time average of the power carried by the combined field above, per unit area in the z-direction. This is representative of the way the field would be detected by the eye, or by photographic film oriented in a plane normal to the z-axis. Once again, we need to evaluate the quantity \( P_z = \frac{1}{2} \Re [E \times H^*] \cdot k \). Now, the time independent electric field is given by Equation 5.2-3. The time independent magnetic field, on the other hand, is best calculated from the curl relationship between \( E \) and \( H \):

\[ \nabla \times E = -j\omega \mu_0 \ H \]  

5.2-6

Since \( E \) has just a y-component, it is easy to show that \( H \) has only x- and z-components, given by:

\[ H_x = 2E_y (-j/\beta) \exp(-j\beta z) \cos[k_0x \sin(\theta)] \]
\[ H_z = 2E_y (-j k_0 \sin(\theta)/\mu_0 \omega) \exp(-j\beta z) \sin[k_0x \sin(\theta)] \]  

5.2-7

Substituting into the expression for \( P_z \), we get the following non-uniform distribution:

\[ P_z = -E_y H_x^* = \frac{4E_y^2}{(\beta/2\mu_0 \omega)} \cos^2[k_0x \sin(\theta)] \]  

5.2-8

This is known as an interference pattern. Notice that it has no dependence on z, but varies periodically with x. It would therefore be visible as a pattern of straight fringes, oriented parallel to the z-axis, as shown in Figure 5.2-2.

The spatial frequency of the pattern is \( K = k_0 \sin(\theta) \), so the separation between fringes is \( \Lambda = 2\pi/K \). Notice that \( \Lambda \) decreases as the angle between the two beams increases, and is of the same order as the wavelength of the light. Interference patterns of this type are particularly important in many branches of optics. However, they do not arise if the two beams are orthogonally polarized, or if the light used is not coherent.
5.3 BOUNDARY MATCHING

There is one further complication to the interface problem. When electromagnetic waves meet changes in dielectric constant, **boundary conditions** must be satisfied. We can find the necessary conditions by applying Maxwell's equations at a boundary discontinuity. Figure 5.3-1 shows a suitable junction between two different semi-infinite media, Medium 1 (where the time-dependent electric field is $E_1$), and Medium 2 (where it is $E_2$).

First, recall from Faraday's law that:

$$\int_{L} E \cdot dL = -\int \frac{\partial B}{\partial t} \cdot da$$

We shall now evaluate the left-hand side of Equation 5.3-1 by performing a line integration round the rectangular dashed loop shown in Figure 5.3-1. This crosses the boundary, and has sides of length $L_1$ and $L_2$. Here $L_2 \ll L_1$, and $L_1$ is small enough for the fields to be approximately uniform along each of the long edges of the loop. The right hand side is of course a surface integral, over the enclosed area.

We now let $L_2$ tend to zero. As this happens, the enclosed area tends to zero, so the right-hand side of Equation 5.3-1 must tend to zero as well. In this case, we must have:
CHAPTER
FIVE

\[ \int_0^L \mathbf{E} \cdot d\mathbf{L} = 0 \]

5.3-2

Now at this point, the integral in Equation 5.3-2 is well approximated by:

\[ \int_0^L \mathbf{E} \cdot d\mathbf{L} = (\mathbf{E}_{t1} - \mathbf{E}_{t2}) L \]

5.3-3

Where \( \mathbf{E}_{t1} \) and \( \mathbf{E}_{t2} \) are the components of \( \mathbf{E}_1 \) and \( \mathbf{E}_2 \) tangential to the boundary. This implies that:

\[ \mathbf{E}_{t1} - \mathbf{E}_{t2} = 0 \]

5.3-4

In other words, the tangential components of \( \mathbf{E} \) must match across the boundary. This is the first of our boundary conditions. We can also write this without explicitly mentioning \( \mathbf{E}_{t1} \) and \( \mathbf{E}_{t2} \), as:

\[ \mathbf{n} \times (\mathbf{E}_2 - \mathbf{E}_1) = 0 \]

5.3-5

Where the vector \( \mathbf{n} \) is a new vector, the local normal to the boundary.

We can do the same sort of thing for \( \mathbf{H}, \mathbf{D} \) and \( \mathbf{B} \). The complete set of boundary conditions is then:

\[ \mathbf{n} \times (\mathbf{E}_2 - \mathbf{E}_1) = 0 \]
\[ \mathbf{n} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{K} \]
\[ \mathbf{n} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = \rho_s \]
\[ \mathbf{n} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = 0 \]

5.3-6

Here \( \mathbf{K} \) is the surface current (which can be ignored in dielectrics) and \( \rho_s \) is the surface charge density (again, zero in dielectrics). Generally, not all of Equations 5.3-6 are needed to solve any particular problem; in fact, picking the most useful subset of conditions is something of an art in itself.

Similar equations can be obtained for the single-frequency, time-independent representation, simply by replacing \( \mathbf{E} \) by \( \mathbf{E} \) and so on. For dielectric media, we therefore obtain:

\[ \mathbf{n} \times (\mathbf{E}_2 - \mathbf{E}_1) = 0 \]
\[ \mathbf{n} \times (\mathbf{H}_2 - \mathbf{H}_1) = 0 \]
\[ \mathbf{n} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = 0 \]
\[ \mathbf{n} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = 0 \]

5.3-7

5.4 ELECTROMAGNETIC TREATMENT OF THE INTERFACE PROBLEM

We must now calculate what happens at the interface more exactly. We assume that the governing equation is the time-independent vector wave equation. For a uniform dielectric, this takes the form:

\[ \nabla^2 \mathbf{E} + n^2 k_0^2 \mathbf{E} = 0 \]

5.4-1
We can use Equation 5.4-1 to solve the interface problem, by first finding solutions for each region as if it were infinite, and then matching them at the boundary. The geometry is shown in Figure 5.4-1. The junction is assumed to lie between two infinite sheets of dielectric, with indices $n_1$ and $n_2$. These are lying in the $x-z$ plane, with the interface at $x = 0$, and the aim is to calculate the field distribution with a plane wave incident from Medium 1 at an angle $\theta_1$.

Figure 5.4-1  Geometry for treatment of the interface problem.

**TE INCIDENCE**

If we assume as before that the incident wave is polarized in the $y$-direction, the electric field is entirely parallel to the interface and only has a $y$-component. This is known as transverse electric or TE incidence. Without justification (apart from common sense) we assume that the transmitted and reflected waves are similarly polarized, so we can work with $E_y$ throughout. The vector wave equation therefore reduces to a single scalar equation. Furthermore, because everything is uniform in the $y$-direction, $\partial/\partial y$ must be zero for all field quantities. The relevant equation is therefore:

$$\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial z^2} + n_2^2 k_0^2 E_y = 0$$  

5.4-2

1. REFLECTION AND REFRACTION

We start by assuming that total internal reflection does not occur, so we need only consider the case of reflection and refraction. The equations we must solve in each medium are:

$$\frac{\partial^2 E_{y1}}{\partial x^2} + \frac{\partial^2 E_{y1}}{\partial z^2} + n_1^2 k_0^2 E_{y1} = 0 \quad \text{in Medium 1}$$

$$\frac{\partial^2 E_{y2}}{\partial x^2} + \frac{\partial^2 E_{y2}}{\partial z^2} + n_2^2 k_0^2 E_{y2} = 0 \quad \text{in Medium 2}$$  

5.4-3

We now assume that the field in each region is a combination of plane waves, the solutions we would find if the media were infinite and separate. A suitable guess is to choose incident and reflected waves in Medium 1, and just a transmitted wave in Medium 2, so that:

$$E_{y1} = E_i \exp[-jk_n(z \sin(\theta_1) - x \cos(\theta_1))] + E_R \exp[-jk_n(z \sin(\theta_1') + x \cos(\theta_1'))]$$

$$E_{y2} = E_T \exp[-jk_n(z \sin(\theta_2) - x \cos(\theta_2))]$$  

5.4-4

Here we have assumed for generality that the reflected wave travels upwards at an angle $\theta_1'$, which may be different from $\theta_1$. Now, there are a number of unknowns in Equation 5.4-4. Given that we know $\theta_1$ and $E_i$, we must find $\theta_1'$, $\theta_2$, $E_R$ and $E_T$. We can find them from the boundary conditions in Equation 5.3-7. In principle, we could use any of these conditions, but here we choose to start with:

R.R.A.Syms and J.R.Cozens  Optical Guided Waves and Devices  6
\[ n_x (E_2 - E_1) = 0 \quad \text{on} \quad x = 0 \]

This requires that the tangential components of \( E \) must match on the boundary. Since \( E \) is wholly tangential anyway, we must have \( E_{y1} = E_{y2} \) on \( x = 0 \), or:

\[
E_i \exp[-jk_n n_1 z \sin(\theta_1)] + E_r \exp[-jk_n n_1 z \sin(\theta_1')] = E_t \exp[-jk_n n_2 z \sin(\theta_2)]
\]

Because it contains exponential terms that vary with \( z \), this equation can only be satisfied for all \( z \) if:

\[
n_1 \sin(\theta_1) = n_1 \sin(\theta_1')
\]

\[
= n_2 \sin(\theta_2)
\]

If this is the case, the \( z \)-variation is removed, and Equation 5.4-6 can be satisfied by a correct choice of the constants \( E_r \) and \( E_t \). Equation 5.4-7 requires firstly that \( \theta_1' = \theta_1 \), so the angle of the reflected beam is indeed equal to that of the incident beam (verifying our initial intuitive guess). However, it also shows that Snell's law appears automatically in the electromagnetic analysis, which is highly satisfactory. Equation 5.4-6 then reduces to:

\[
E_i + E_r = E_t
\]

Performing the boundary matching for the electric field, we therefore find that the wave amplitudes must be continuous at the interface, which is physically reasonable.

The second boundary condition we choose is that:

\[ n_x (H_2 - H_1) = 0 \quad \text{on} \quad x = 0 \]

This implies that the tangential components of \( H \) must also match at \( x = 0 \). We can find the necessary components of \( H \) by using the curl relation between \( E \) and \( H \) given previously in Equation 5.2-6. Since \( E_x = E_z = 0 \), we have:

\[
H_x = (-j/\omega \mu_0) \frac{\partial E_y}{\partial z}; \quad H_y = 0; \quad H_z = (+j/\omega \mu_0) \frac{\partial E_y}{\partial x}
\]

The only tangential component of \( H \) is \( H_z \), so the second boundary condition will be satisfied if:

\[
\frac{\partial E_{y1}}{\partial x} = \frac{\partial E_{y2}}{\partial x} \quad \text{on} \quad x = 0.
\]

Using this condition, we find that:

\[
jk_n n_1 \cos(\theta_1) E_i \exp[-jk_n n_1 z \sin(\theta_1)] - jk_n n_1 \cos(\theta_1) E_r \exp[-jk_n n_1 z \sin(\theta_1)]
\]

\[
= jk_n n_2 \cos(\theta_2) E_t \exp[-jk_n n_2 z \sin(\theta_2)]
\]

Following the same argument as before, we can remove all the exponentials to get:

\[
n_1 \cos(\theta_1) E_i - n_1 \cos(\theta_1) E_r = n_2 \cos(\theta_2) E_t
\]

Equations 5.4-8 and 5.4-13 now represent two simultaneous equations, with two unknowns. By a fairly simple rearrangement, we can therefore find \( E_r \) in terms of \( E_i \). Normally the result is expressed in terms of a reflection coefficient \( \Gamma_E \), which can be shown to reduce to:
\[ \Gamma_E = \frac{E_R}{E_i} = \frac{[n_1 \cos(\theta_1) - n_2 \cos(\theta_2)]}{[n_1 \cos(\theta_1) + n_2 \cos(\theta_2)]} \]

5.4-14

We can also find the transmission coefficient \( T_E \), which is given by:

\[ T_E = \frac{E_T}{E_i} = \frac{2n_1 \cos(\theta_1)}{[n_1 \cos(\theta_1) + n_2 \cos(\theta_2)]} \]

5.4-15

These expressions satisfy physical intuition. For example, if the two media have the same refractive index, the reflection coefficient is zero, and the transmission coefficient is unity. Surprisingly, the reflection coefficient can be large, even at normal incidence. For a glass/air interface (where \( n_1 = 1.5 \) and \( n_2 = 1 \)), we find that \( \Gamma_E = \frac{(1.5 - 1)}{(1.5 + 1)} = 0.2 \).

THE OTHER BOUNDARY CONDITIONS

At this point we pause, and consider the fact that there were two more boundary conditions we could have used. What about them? The third condition is that:

\[ \mathbf{n} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = 0, \quad \text{or} \quad \mathbf{n} \cdot (\sqrt{n_2} \mathbf{E}_2 - \sqrt{n_1} \mathbf{E}_1) = 0 \quad \text{on} \quad x = 0 \]

5.4-16

Since \( \mathbf{E} \) has no component normal to the boundary, all this equation says is \( 0 = 0 \). The third condition is therefore not too helpful, and we can safely ignore it. The fourth is that:

\[ \mathbf{n} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = 0, \quad \text{or} \quad \mathbf{n} \cdot (\mathbf{H}_2 - \mathbf{H}_1) = 0 \quad \text{on} \quad x = 0 \]

5.4-17

Since the only normal component of \( \mathbf{H} \) is \( \mathbf{H}_z \), this implies that:

\[ H_{z2} = H_{z1} \quad \text{on} \quad x = 0 \]

5.4-18

Bearing in mind the definition of \( \mathbf{H} \) given in Equation 5.4-10, we find that:

\[ \frac{\partial E_{y2}}{\partial z} = \frac{\partial E_{y1}}{\partial z} \quad \text{on} \quad x = 0. \]

5.4-19

Doing the necessary differentiation, we get:

\[ \begin{align*}
-jk_0 n_1 \sin(\theta_1) & \ E_\| \ exp[-jk_0 n_1 z \sin(\theta_1)] - jk_0 n_1 \sin(\theta_1) \ E_\| \ exp[-jk_0 n_1 z \sin(\theta_1)] \\
= -jk_0 n_1 \sin(\theta_1) & \ E_\| \ exp[-jk_0 n_1 z \sin(\theta_1)]
\end{align*} \]

5.4-20

If we now use Snell’s law again to remove the exponentials, this reduces to Equation 5.4-8. The two other boundary conditions therefore introduce no new features, so our original choice was a good one.

TM INCIDENCE

The analysis can easily be repeated for the case when the magnetic field is oriented in the y-direction. This is known as transverse magnetic or TM incidence. We shall not work through all the mathematics, which is somewhat repetitive, but merely quote the results. These are slightly different; for the reflection coefficient, we get:

\[ \Gamma_H = \frac{E_R}{E_i} = \frac{[n_2 \cos(\theta_1) - n_1 \cos(\theta_2)]}{[n_2 \cos(\theta_1) + n_1 \cos(\theta_2)]} \]

5.4-21

and for the transmission coefficient, we find that:

\[ T_H = \frac{E_T}{E_i} = \frac{2n_1 \cos(\theta_1)}{[n_2 \cos(\theta_1) + n_1 \cos(\theta_2)]} \]
Collected together, $\Gamma_E$, $\Gamma_T$, $\Gamma_H$ and $T_H$ are often known as the Fresnel coefficients. Most interestingly, $\Gamma_H$ becomes zero at a particular angle of incidence $\theta_1$. This occurs when:

$$n_2 \cos(\theta_1) = n_1 \cos(\theta_2)$$

We can find a suitable angle $\theta_1$ for Equation 5.4-23 to hold, as follows. Squaring both sides of the equation, and using Snell's Law, we obtain:

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

Or:

$$n_2^2 \cos^2(\theta_1) + \left(\frac{n_1^4}{n_2^4}\right) \sin^2(\theta_1) = n_1^2$$

The $\cos^2$ term above can now be eliminated by combining it with the $\sin^2$ term. If this is done, we get:

$$n_2^2 \left[\cos^2(\theta_1) + \sin^2(\theta_1)\right] + \left[(n_1^4 - n_2^4)/n_2^4\right] \sin^2(\theta_1) = n_1^2$$

so that:

$$\left[(n_1^4 - n_2^4)/n_2^4\right] \sin^2(\theta_1) = n_1^2 - n_2^2$$

We therefore find that:

$$\sin^2(\theta_1) = \frac{n_2^2}{n_1^2 + n_2^2}$$

So the relevant angle is given by:

$$\theta_1 = \sin^{-1}\left[\frac{n_2}{\sqrt{(n_1^2 + n_2^2)}}\right]$$

This is often written in the alternative form:

$$\theta_1 = \tan^{-1}\left(\frac{n_2}{n_1}\right)$$

The reduction of the reflection coefficient to zero for TM incidence is called the Brewster effect, after David Brewster (1781 - 1868), and is often exploited in the design of polarizing components. Similarly, the angle defined in Equation 5.4-30 is known as the the Brewster angle $\theta_B$. This can be calculated for an glass/air interface (for which $n_1 = 1.5$ and $n_2 = 1.0$) as $\theta_B \approx 33.7^\circ$. Note that the Brewster effect does not occur for TE incidence.

Figure 5.4-2 shows one way in which the Brewster effect may be used to make a polarizer. A number of identical plates are piled up, and oriented so that an incident, unpolarized beam strikes the stack at the Brewster angle. Any TE polarization component in the beam will be diminished by reflection at each interface, while the TM component will pass through unaffected. After many such reflections, the output beam will therefore be largely TM polarized. This is known as the pile-of-plates polarizer. Though simple, it has been entirely superseded by dichroic polarizers.
Figure 5.4-2  The pile of plates polarizer.

Figure 5.4-3 shows a much more effective use of the Brewster effect, inside the resonant cavity of a He-Ne laser. We have previously described the required optical layout in Chapter 4, and concentrate here on the provision of optical gain by an electrically excited gas plasma, which is contained inside a glass tube. In lasers that emit unpolarized light, the plasma tube is simply capped with end plates that are orthogonal to the bore axis. In polarized lasers, the tube is sealed with windows slanted at the Brewster angle (known, unsurprisingly, as Brewster windows). Now, TE polarized light must suffer some reflection loss at these windows. However, because of the Brewster effect, TM polarized light suffers no loss. Consequently, the lasing threshold for TM light is much lower than for TE, and the device will lase preferentially on a TM-polarized mode.

2. TOTAL INTERNAL REFLECTION

We now consider the case when $\theta_1 > \theta_c$, i.e. when the input beam is incident at an angle greater than the critical angle. In this regime, total internal reflection occurs, so that a propagating wave cannot arise in Medium 2. How do we find the solution now? Well, there is nothing wrong with the analysis we have used so far, but we cannot satisfy the Snell's law equation, $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$. Or rather, we cannot satisfy it with a real value of $\theta_2$. It turns out that we can satisfy it with a complex value instead. This might seem a bizarre idea, but there is no need to dwell on the physical implications involved; it is simplest just to consider this as a mathematical trick. The important thing is that we can now calculate $\cos(\theta_2)$ for $\theta_1 > \theta_c$ as:

$$\cos(\theta_2) = \pm j \sqrt{(n_1/n_2)^2 \sin^2(\theta_1) - 1}$$

5.4-31
This identity will allow us to repeat the previous calculation under the new conditions.

**TE INCIDENCE**

Once again, we will begin with the case when the electric field vector is parallel to the interface. Returning to the solution in Medium 2, 

$$E_{y2} = E_T \exp[-jk_0n_2(z\sin(\theta_2) - x\cos(\theta_2))]$$

and inserting the value for the cosine given in Equation 5.4-31, we find:

$$E_{y2} = E_T \exp \left[\pm k_0n_2x \sqrt{(n_1/n_2)^2 \sin^2(\theta_1) - 1}\right] \exp[-jk_0n_1z\sin(\theta_1)].$$

5.4-32

The solution no longer has the form of a plane wave propagating at an angle $\theta_2$, because its $x$-dependence is not a complex exponential. Instead, it is a wave whose amplitude distribution varies exponentially in the $x$-direction, and that propagates along the interface in the $z$-direction. This is called an **evanescent wave**. Clearly, the amplitude must decay to zero for large negative $x$, so we will keep only the positive sign in Equation 5.4-32, which corresponds to the negative sign in Equation 5.4-31.

What does the complete field distribution look like now? Well, we now know that:

$$E_{y1} = E_I \left\{ \exp(+jk_0n_1x\cos(\theta_1)) + \Gamma_E \exp(-jk_0n_1x\cos(\theta_1)) \right\} \exp(-jk_0n_1z\sin(\theta_1))$$

$$E_{y2} = E_I T_E \exp \left[+k_0n_2x \sqrt{(n_1/n_2)^2 \sin^2(\theta_1) - 1}\right] \exp[-jk_0n_1z\sin(\theta_1)].$$

5.4-33

Where the transmission and reflection coefficients are given by:

$$\Gamma_E = \left[ n_1 \cos(\theta_1) + jn_2 \sqrt{(n_1/n_2)^2 \sin^2(\theta_1) - 1} \right] / \left[ n_1 \cos(\theta_1) - jn_2 \sqrt{(n_1/n_2)^2 \sin^2(\theta_1) - 1} \right]$$

5.4-34

and:

$$T_E = 2n_1 \cos(\theta_1) / \left[ n_1 \cos(\theta_1) - jn_2 \sqrt{(n_1/n_2)^2 \sin^2(\theta_1) - 1} \right]$$

5.4-35

These expressions are rather long-winded. However, we can simplify them by defining three new parameters $\beta$, $\kappa$, and $\gamma$ as:

$$\beta = k_0n_1 \sin(\theta_1)$$

$$\kappa = \sqrt{(n_1^2k_0^2 - \beta^2)}$$

$$\gamma = \sqrt{\beta^2 - n_2^2k_0^2}$$

5.4-36

in which case Equation 5.4-33 becomes:

$$E_{y1} = E_I \left\{ \exp(+j\kappa x) + \Gamma_E \exp(-j\kappa x) \right\} \exp(-j\beta z)$$

$$E_{y2} = E_I T_E \exp (+j\gamma x) \exp(-j\beta z)$$

5.4-37

And Equations 5.4-34 and 5.4-35 simplify to:

$$\Gamma_E = (\kappa + j\gamma) / (\kappa - j\gamma)$$

$$T_E = 2\kappa / (\kappa - j\gamma)$$

5.4-38

The reflection coefficient $\Gamma_E$ is therefore now a complex number, in the form $\Gamma_E = z / z^*$, where $z$ is also a complex number and $*$ denotes the operation of complex conjugation. It is easy to show that the **power reflection coefficient** $|\Gamma_E|^2$ is unity, since:

$$|\Gamma_E|^2 = \Gamma_E \Gamma_E^* = (z/z^*)(z^*/z) = 1$$
This implies that all of the incident power is reflected, which is consistent with our earlier qualitative discussion of total internal reflection. It might be argued that a non-zero transmission coefficient is at odds with a unity power reflection coefficient. The evanescent wave will then have a finite amplitude, and must therefore carry some power. However, it is easy to show that it does not carry any power normal to the interface, but only parallel to it. Power conservation is therefore not violated after all.

Since $\Gamma_E$ is a complex number with modulus unity, it can be written in complex exponential form, as:

$$\Gamma_E = \exp(j2\phi)$$

where the phase term $\phi$ is given by:

$$\tan(\phi) = \gamma/\kappa$$

Similarly, the transmission coefficient $T_E$ can be written as:

$$T_E = 2 \cos(\phi) \exp(j\phi)$$

We can obtain entirely analogous results for the TM transmission and reflection coefficients when total internal reflection occurs. The phase shift is different, however; this time we get $\tan(\phi) = (n_1^2/n_2^2) \gamma/\kappa$.

**DESIGN EXAMPLE**

We can combine the solutions obtained so far, and plot the power reflectivity $|\Gamma|^2$ versus angle over the complete range of incidence. Figure 5.4-4 shows the curves for TE and TM incidence, and the particular example of a glass/air interface (where $n_1 = 1.5$ and $n_2 = 1$).
Naturally, the reflectivities are the same for normal incidence, when $|\Gamma|^2 = 0.2^2 = 0.04$. In each case, the reflectivity is unity after the critical angle, $\theta_1 = 41.8^\circ$. Notice, however, that it falls to zero, for TM incidence alone, at the Brewster angle ($33.7^\circ$).

**THE TRANSVERSE FIELD DISTRIBUTION NEAR THE INTERFACE**

Knowing the transmission and reflection coefficients, we can now rewrite the field distributions in the two media as:

$$E_{y1} = 2E_i \exp(j\phi) \cos(\kappa x - \phi) \exp(-j\beta z) = E \cos(\kappa x - \phi) \exp(-j\beta z)$$

5.4-43

And:

$$E_{y2} = 2E_i \exp(j\phi) \cos(\phi) \exp(\gamma x) \exp(-j\beta z) = E \cos(\phi) \exp(\gamma x) \exp(-j\beta z)$$

5.4-44

Both fields again have the form $E_y = E(x) \exp(-j\beta z)$, so the pattern corresponds to a non-uniform wave travelling in the $z$-direction, parallel to the interface. More specifically, it is a standing wave pattern in Medium 1, and a decaying or evanescent wave in Medium 2. Near the interface, the transverse field function $E(x)$ is therefore as shown in Figure 5.4-5. Notice that the transverse field is continuous at the boundary, as is its first derivative. This follows from application of the boundary matching conditions. It is also worth noting that as $\beta$ tends to $n_2k_0$, $\gamma$ tends to zero, implying that, as the angle of incidence approaches the critical angle, the decay of the evanescent field gets slower. The confinement of the field is therefore less good, and it extends further and further into Medium 2. If we wish to avoid this, the incidence angle should be chosen to lie suitably far away from the critical angle.

![Figure 5.4-5](image)

The results are qualitatively similar for TM incidence; we still get standing waves in Medium 1 and an evanescent wave in Medium 2. However, the details at the boundary are slightly different. The main change is that the transverse electric field distribution is no longer continuous.
THE REFLECTIVITY OF METALS

We conclude with a brief discussion of the reflectivity of metals, which we assumed without much justification to be high in Chapter 4. The analysis developed above will still be valid, provided merely that we correct the expression for refractive index. In Chapter 3, we found that (neglecting collision damping losses) a metal may be described at optical frequencies as a medium with a large, negative value of $\varepsilon'$. Consequently, the refractive index of a metal will be almost purely imaginary, given by $n \approx jn'$, where $n' = \sqrt{|\varepsilon'|}$.

At normal incidence, for example, the reflection coefficient at the interface between a dielectric (of refractive index $n_1$) and a metal (of index $jn_2'$) may then be found from Equation 5.4-14 as:

$$\Gamma_E \approx \frac{(n_1 - jn_2')}{(n_1 + jn_2')}$$  

5.4-45

Once again, $\Gamma_E$ is a complex number, of the form $\Gamma_E = z / z^*$. We therefore conclude without further ado that the power reflectivity will be 100%, justifying our previous assumptions. In practise, collision damping losses reduce this figure somewhat, but (as we know from our bathrooms) even the most basic metal layer will act as a good reflector. Furthermore, if $n_2 \gg n_1$ (as is the case for good conductors), the reflection coefficient $\Gamma_E$ is $\approx -1$. This means the total electric field at the interface (i.e. the sum of the incident and reflected waves) must be close to zero, providing a simple justification for the heuristic boundary condition adopted in Chapter 4.

5.5 MODAL TREATMENT OF THE DIELECTRIC INTERFACE PROBLEM

We now consider an easier route to the solution to the interface problem, which will prove a useful tool for the analysis of waveguides in Chapter 6. Since the solution has the form $E_y(x, z) = E(x) \exp(-j\beta z)$ in both media, the thing to do is to assume this at the outset. Substituting this solution directly into the wave equation 5.4-2, we get:

$$\frac{d^2E}{dx^2} + (n_2^2k_0^2 - \beta^2) E = 0$$

5.5-1

This type of equation is known as a waveguide equation, and it links the transverse field $E(x)$ with the propagation constant $\beta$. In this case, it is a standard second-order differential equation, of the form:

$$\frac{d^2E}{dx^2} + C^2E = 0$$

5.5-2

Note that its solutions are sines and cosines if $C^2 > 0$, and exponentials if $C^2 < 0$.

For our particular geometry, the waveguide equations we must solve are:

$$\frac{d^2E_1}{dx^2} + (n_1^2k_0^2 - \beta^2) E_1 = 0 \quad \text{in Medium 1}$$
$$\frac{d^2E_2}{dx^2} + (n_2^2k_0^2 - \beta^2) E_2 = 0 \quad \text{in Medium 2}$$

5.5-3

Here $E_1$ and $E_2$ are the transverse field distributions in the two media. Knowing (as we do) the exact solution, the sensible thing is to choose a trial solution in the same form. We therefore take:

$$E_1 = E \cos(\kappa x - \phi)$$
$$E_2 = E' \exp(\gamma x)$$

R.R.A. Symms and J.R. Cozens  Optical Guided Waves and Devices 14
Where $\kappa$ and $\gamma$ are as in Equation 5.4-36 and $E$, $E'$ and $\phi$ are unknown constants. Once again, we can determine their values by boundary matching. Now, the first of our original boundary conditions were that $E_{y1} = E_{y2}$ at the interface; for the transverse fields, this implies that $E_1 = E_2$ on $x = 0$, so:

$$E' = E \cos(\phi)$$ \hspace{1cm} 5.5-4

Similarly, the second condition was that $\partial E_{y1}/\partial x = \partial E_{y2}/\partial x$ at the interface; for the transverse fields, this implies that $dE_1/dx = dE_2/dx$ on $x = 0$. Performing the necessary differentiation, we get:

$$-\kappa E \sin(-\phi) = \gamma E'$$ \hspace{1cm} 5.5-5

Dividing Equation 5.5-6 by Equation 5.5-5, we then obtain:

$$\tan(\phi) = \frac{\gamma}{\kappa}$$ \hspace{1cm} 5.5-6

Since this is identical to Equation 5.4-41, we have obtained exactly the same results as before, without invoking the Fresnel coefficients at all. A solution of this type is called a modal solution.

### 5.6 SURFACE PLASMA WAVES

We have become accustomed to solutions to interface problems that involve either propagating waves on both sides of the interface, or propagating waves on one side and evanescent ones on the other. It is now reasonable to ask whether other possibilities exist. Can we, for example, have a solution that is evanescent on both sides of an interface? It turns out that this is indeed possible, and is of engineering significance.

The geometry involved is the interface between a metal (most often, silver or gold) and a dielectric, as shown in Figure 5.6-1. Because the wave is evanescent on both sides, it is effectively confined to the interface, and is therefore a surface wave. Its correct title is a surface plasma wave or surface plasmon, since propagation in the metal layer involves plasma oscillations of the free electrons inside the metal, but this is not important to our electromagnetic analysis. What is important, however, is that surface plasma waves can be shown to exist only for TM polarization.

![Figure 5.6-1](image)

**Figure 5.6-1** Geometry for surface plasmon waves.
We shall tackle the problem using the modal approach above. First, however, we recall once again that at optical frequencies a metal can be adequately modelled as a dielectric, but with a relative permittivity that is large and negative. The methods we have used so far will therefore still prove suitable. Now, it is sensible to perform the analysis in terms of the magnetic field, since this is always oriented in the same direction. In this case, the relevant time-independent wave equation is:

\[ \nabla^2 H + \varepsilon \varepsilon_r k_0^2 H = 0 \]  \hspace{1cm} 5.6-1

Assuming that the magnetic field is parallel to the y-axis, a suitable modal solution has the form:

\[ H = H(x) \exp(-j\beta z) \]  \hspace{1cm} 5.6-2

Where \( H(x) \) is the transverse variation of the magnetic field, and \( \beta \) is the propagation constant. Substituting this solution into Equation 5.6-1, we obtain the following waveguide equation:

\[ \frac{d^2 H}{dx^2} + (\varepsilon \varepsilon_r k_0^2 - \beta^2) H = 0 \]  \hspace{1cm} 5.6-3

To solve the plasmon problem, we simply assume a relative dielectric constant of \( \varepsilon_{r1} \) in Medium 1 (the metal layer), and \( \varepsilon_{r2} \) in Medium 2 (the dielectric). The equations we must solve are therefore:

\[ \frac{d^2 H_1}{dx^2} + (\varepsilon_{r1} k_0^2 - \beta^2) H_1 = 0 \hspace{1cm} \text{in Medium 1} \]
\[ \frac{d^2 H_2}{dx^2} + (\varepsilon_{r2} k_0^2 - \beta^2) H_2 = 0 \hspace{1cm} \text{in Medium 2} \]  \hspace{1cm} 5.6-4

Assuming exponential decay of the transverse magnetic field on either side of the interface, the solutions in each region can then be specified as:

\[ H_1(x) = H \exp(-\gamma_1 x) \]
\[ H_2(x) = H' \exp(\gamma_2 x) \]  \hspace{1cm} 5.6-5

Where \( H \) and \( H' \) are arbitrary constants, and \( \gamma_1 \) and \( \gamma_2 \) are given by:

\[ \gamma_1 = \sqrt{(\beta^2 - \varepsilon_{r1} k_0^2)} \]
\[ \gamma_2 = \sqrt{(\beta^2 - \varepsilon_{r2} k_0^2)} \]  \hspace{1cm} 5.6-6

As before, the boundary conditions are that the tangential components of the magnetic and electric fields must match at the interface, which lies at \( x = 0 \). Matching the magnetic field (which is wholly tangential) we see simply that \( H' = H \). To match the electric fields, we must first evaluate \( E \). This is best done using the curl relation between \( H \) and \( E \):

\[ \nabla \times H = j\omega \varepsilon E \]  \hspace{1cm} 5.6-7

Given that \( H_x = H_z = 0 \), we can then find the individual components of \( E \) as:

\[ E_x = (+j/\omega\varepsilon) \frac{\partial H_y}{\partial z} \hspace{0.5cm} E_y = 0 \hspace{0.5cm} E_z = (-j/\omega\varepsilon) \frac{\partial H_y}{\partial x} \]  \hspace{1cm} 5.6-8

Since the only tangential component of the electric field is \( E_x \), the second condition will be satisfied if:
\[
\frac{1}{\varepsilon_{r1}} \frac{\partial H_y}{\partial x} = \frac{1}{\varepsilon_{r2}} \frac{\partial H_y}{\partial x} \quad \text{on} \quad x = 0
\]

For the modal solution we have assumed, this condition requires:

\[
\frac{1}{\varepsilon_{r1}} \frac{dH_1}{dx} = \frac{1}{\varepsilon_{r2}} \frac{dH_2}{dx} \quad \text{on} \quad x = 0
\]

Performing the necessary differentiation, we then find that:

\[-\gamma_1/\varepsilon_{r1} = \gamma_2/\varepsilon_{r2}\]

Although Equation 5.6-11 looks insoluble (given that \(\gamma_1\) and \(\gamma_2\) are positive quantities), it can be satisfied in this particular case, because \(\varepsilon_{r1}\) is negative. Squaring both sides, we obtain:

\[
\frac{(\beta^2 - \varepsilon_{r1} k_0^2)}{\varepsilon_{r1}^2} = \frac{(\beta^2 - \varepsilon_{r2} k_0^2)}{\varepsilon_{r2}^2}
\]

After some rearrangement, we may then extract the propagation constant as:

\[
\beta = k_0 \sqrt{\frac{\varepsilon_{r1} \varepsilon_{r2}}{\varepsilon_{r1} + \varepsilon_{r2}}}
\]

Knowing \(\varepsilon_{r1}\) and \(\varepsilon_{r2}\), we may therefore calculate \(\beta\). The transverse magnetic field distribution is then much as shown in Figure 5.6-1. The field decays exponentially on both sides of the interface, but the decay constant is different in each medium. Because of the sign of \(\varepsilon_{r1}\), Equation 5.6-6 shows that the rate of decay in the metal is much larger than that in the dielectric. As a result, the light hardly penetrates the metal, and the major part of the field extends into the dielectric.

The analysis above is an extremely simple demonstration that a single interface may actually 'guide' an optical wave. However, to be more realistic, one major modification must be made to our model. As we saw in Chapter 3, the dielectric constant of a metal is actually complex (because of collision damping), so \(\varepsilon_{r1}\) really has the form \(\varepsilon_{r1} = \varepsilon_{r1}' - j \varepsilon_{r1}''\). If this expression is used in Equation 5.6-13, it is found that the propagation constant is also complex, so the plasmon mode decays as it propagates. In real materials, this is very significant, and plasmon modes have an extremely short range.

**DESIGN EXAMPLE**

We shall now calculate the range of a surface plasmon in a typical material, silver. In this case, the data in Figure 3.4-1 shows that \(\varepsilon_{r1}' \approx -16.9\) and \(\varepsilon_{r1}'' \approx 0.55\), at \(\lambda_0 = 0.633 \, \mu m\).

Assuming that the surrounding dielectric is air, for which \(\varepsilon = 1\), \(\beta\) is given by:

\[
\beta = \frac{2\pi}{0.633 \times 10^{-6}} \times \sqrt{(-16.9 - j 0.55) / (-15.9 - j 0.55)} = 10.2 \times 10^6 - j 1.05 \times 10^4
\]

Now, the actual form of the plasmon wave is given by Equation 5.6-2. Splitting the propagation constant into real and imaginary parts, we can write:

\[
\beta = \beta_r - j\beta_i
\]

With this notation, Equation 5.6-2 then becomes:

\[
H = H(x) \exp(-j\beta_i z) \exp(-\beta_r z) j
\]
This shows that the mode amplitude will decay to $1/e$ of its initial value in a distance $z_e$, given by:

$$z_e = 1/\beta_i$$

We may take $z_e$ to be representative of the range of the plasmon mode. Inserting the relevant value from Equation 5.6-14, we then obtain a range of $z_e \approx 100 \mu m$, a very small figure. In practice, other effects reduce the range still further, making a surface plasmon wave useless for communication. Nonetheless, one highly important application has been found for optical plasmons: sensing.

**SENSING WITH SURFACE PLASMON WAVES**

We have already mentioned that because of the imbalance between $\gamma_1$ and $\gamma_2$, the vast proportion of the field extends into the dielectric. As a result, it is possible to show that the value of $\beta_r$ is highly sensitive to the exact nature of the dielectric. If this medium is made in the form of a sensing layer, with a dielectric constant that can be altered by some parameter of interest, a variation in that parameter will then result in a detectable change in the plasmon propagation constant.

The experimental apparatus needed to measure this change is extremely simple. It is known as the **prism coupler**, and will be described in detail in Chapter 7. For the present, we merely note that many sensor materials have already been tested for suitability. The most promising application area is biosensing, where it is hoped that surface plasmon devices will provide a range of extremely cheap, disposable medical sensors (for example, for measuring enzyme levels).
PROBLEMS

5.1. (a) Find the critical angle for an interface between a high-index glass \((n_1 = 1.7)\) and air \((n_2 = 1)\). (b) What is the corresponding Brewster angle, if the incident beam travels in the glass?
\[(a) 36.03^\circ; \ (b) 30.46^\circ\]

5.2. The critical angle for an interface between two media is \(30^\circ\). Find the transmission coefficient, for normal incidence from the high-index side of the interface.
\[1.333\]

5.3. Find the power reflectivity \(\Gamma_2\) and transmissivity \(T_2\), for TE incidence at \(\theta_1 = 30^\circ\) on an interface between air \((n_1 = 1)\) and glass \((n_2 = 1.5)\).
\[0.0577; 0.577\]

5.4. The figures below show the definitions of the reflection and transmission coefficients at a dielectric interface. In the left-hand figure, incidence is from Medium 1, and the coefficients are \(\Gamma\) and \(T\), respectively. In the right-hand figure, incidence is from Medium 2, and the corresponding values are \(\Gamma'\) and \(T'\). Using only the fact that light rays are reversible, show that (a) \(\Gamma = -\Gamma'\), and (b) \(\Gamma^2 + TT' = 1\).

5.5. Verify the results of Question 5.4 analytically, using the Fresnel coefficients.

5.6. The figure below shows a three-layer dielectric stack. Media 1 and 2 are semi-infinite, with refractive indices \(n_1\) and \(n_2\), respectively; Medium 3 has thickness \(d\), and index \(n_3\). The effect of the stack on a plane optical wave may be analysed by summing the contributions from all possible paths between the input and the output, using the transmission and reflection coefficients to calculate the path amplitudes and taking account of any optical phase changes.

For example, the reflectivity may be found by summing all paths starting and ending in Medium 1. Path 1 simply involves reflection at the first interface. The amplitude for this is \(\Gamma_{13}\), where \(\Gamma_{13}\) is the reflection coefficient at the boundary between Media 1 and 3. Path 2 involves transmission through the first interface, reflection at the second, and transmission through the first interface again; the amplitude for this is \(T_{13}\Gamma_{32}T_{31}\exp(-j2k_0n_3d)\), where \(T_{13}\) and \(T_{31}\) are transmission coefficients at the first interface, and \(\Gamma_{32}\) is the reflection coefficient at the second.

Write down the amplitude for Path 3. Generalise your results to form an infinite series for the reflectivity of the stack, and sum the series. Hence show that there will be no reflection at all, if the refractive index of Medium 3 is \(n_3 = \sqrt{n_1n_2}\) and if its thickness \(d\) represents a quarter of a wavelength in the medium. What application might this have?
5.7. Derive the Fresnel coefficients for TM incidence on a dielectric interface.

5.8. The figure below shows a plane wave incident from air on a glass slab of finite thickness. Show that there is no reflection from either interface, for TM incidence at the Brewster angle $\theta_B$.

5.9. Show (a) analytically, and (b) numerically, using data from Question 5.3, that the Fresnel coefficients for TE incidence satisfy the relation:

$$\Gamma^2 + T^2 \left( \frac{n_2}{n_1} \right) \left[ \frac{\cos(\theta_1)}{\cos(\theta_2)} \right] = 1.$$

5.10. Interpret the result of Question 5.9 in terms of power conservation.
SUGGESTIONS FOR FURTHER READING

Kretschmann E. "The determination of the optical constants of metals by the excitation of surface plasmons" Z. Phys. 214, 313-324 (1971)
6.1 GUIDED WAVES IN A METAL GUIDE

It is now time to consider more effective guiding structures for electromagnetic waves. To introduce the basic ideas, we shall start with a very simple geometry, the metal-walled guide. Though this is not actually used at optical frequencies, it contains nearly all the features of more useful guides and can be analysed with little mathematics. The metal-walled guide consists of two plane mirrors held parallel to one another in free space, as shown in Figure 6.1-1. From our knowledge of mirrors, we would expect plane waves to be reflected and re-reflected at each interface, and hence bounce up and down between the mirrors. This zig-zag pathway effectively results in travel only in the z-direction, so the field is guided down the z-axis.

We must begin by working out what kind of field can exist in the guide. First, we recall that the time-independent field due to a y-polarized plane wave travelling at an angle $\theta$ to the z-axis (either upwards or downwards) is:

$$E_y = E \exp\{\pm jk_0 [z \sin(\theta) \pm x \cos(\theta)]\}$$

6.1-1

Since we can guess that the field inside the guide is a combination of an upward- and a downward-travelling wave, we will assume the solution:

$$E_y = E_+ \exp\{-jk_0 [z \sin(\theta) + x \cos(\theta)]\} + E_- \exp\{-jk_0 [z \sin(\theta) - x \cos(\theta)]\}$$

6.1-2

where $E_+$ and $E_-$ are unknown constants. Now, we must satisfy the usual set of boundary conditions at any interfaces. If we adopt the approximations used in Chapters 3, 4 and 5 for a good conductor, these reduce to a requirement that the electric field vanishes at the metal walls (so that $E_+ = 0$ at $x = 0$ and $x = h$). We can satisfy the first condition if $E_- = -E_+$, i.e. if the solution has the form:

$$E_y = E \sin[k_0 x \cos(\theta)] \exp\{-jk_0 z \sin(\theta)\}$$

6.1-3

The solution is therefore a wave with a sinusoidal amplitude envelope that travels in the z-direction. It is called a guided mode, because it has its energy confined within the guide walls. Note that the field can once again be written in the standard form:

$$E_y = E(x) \exp(-j\beta z)$$

6.1-4
where \( E(x) \) is the **transverse field distribution**, and \( \beta \) is the **propagation constant**.

Clearly, in this case, \( \beta \) is given by:

\[
\beta = k_0 \sin \theta
\]

6.1-5

so the smaller the ray angle inside the guide, the smaller the propagation constant. The exact value of \( \theta \) (and thus of \( \beta \)) is fixed by the second boundary condition, namely:

\[
\sin[k_0 h \cos(\theta)] = 0
\]

6.1-6

Now, Equation 6.1-6 is satisfied whenever:

\[
2k_0 h \cos(\theta) = 2\nu\pi \quad \text{(where } \nu = 1, 2, \ldots \text{ etc.)}
\]

6.1-7

This is called the **eigenvalue equation** of the guide. It may be viewed as a condition for transverse resonance, since it implies that the round-trip phase accumulated by bouncing up and down between the walls must be a whole number of multiples of \( 2\pi \).

Each solution corresponds to a particular guided mode, defined by the **mode index** \( \nu \). However, in general, only a fixed number of modes can be supported by the guide. For example, if \( \nu \pi/k_0 h > 1 \), there is no solution, so the condition for a guided mode of order \( \nu \) to exist is that:

\[
h > \nu\pi/k_0
\]

6.1-8

Consequently, if the guide width \( h \) is too small, no modes are supported all. This occurs when:

\[
h < \pi/k_0 \quad \text{or} \quad h < \lambda_0/2
\]

6.1-9

This puts a lower limit on the useable value of \( h \) of half the optical wavelength. If \( h \) is slightly bigger, just one mode is supported, and the guide is then called **single-moded**. If \( h \) is bigger still, a finite number of modes can propagate, so the guide is **multi-moded**. Any mode that cannot propagate is described as being **cut off**. The higher the order of the mode, the smaller the value of \( \theta \), and, exactly at cut-off, the ray angle is zero, so at this point the rays just bounce up and down between the metal walls, making no progress down the guide. Each guided mode has its own particular transverse field distribution. The field patterns of the first two modes are shown in Figure 6.1-2. Note that the lowest-order mode has no sign reversals, while the second-order mode has one, and so on.

**Figure 6.1-2** Transverse field distributions of the two lowest order modes in a metal-walled guide.
The metal-walled guide is very effective at microwave frequencies, when the conductivity of the walls is high, and the reflectivity is correspondingly good. However, at optical frequencies, $\sigma$ is lower, causing unacceptable propagation loss. Fortunately, an alternative guiding mechanism is available, based on total internal reflection at a dielectric interface.

### 6.2 GUIDED WAVES IN A SLAB DIELECTRIC WAVEGUIDE

Figure 6.2-1 shows more useful structure, a dielectric waveguide. This consists of three layers of dielectric: layer 1 (which has thickness $h$, and refractive index $n_1$), and layers 2 and 3 (which are both semi-infinite, and which have indices $n_2$ and $n_3$ respectively). We shall assume that $n_1 > n_2$ and $n_1 > n_3$, so that total internal reflection can occur at each interface. Often, layer 1 is referred to as the guiding layer, while layers 2 and 3 are the substrate and the cover layer. A guide of this type can be formed simply by depositing a high-index guiding layer onto a polished substrate - the cover can then often be air. Because of this common geometry, it is usual to describe the structure as an asymmetric slab guide, and take $n_1 > n_2 > n_3$.

![Figure 6.2-1 A slab dielectric waveguide.](image)

We can represent the refractive index variation using a discontinuous function $n(x)$, as shown in Figure 6.2-2a. Other, smoother profiles can result from using different fabrication methods; for example, Figure 6.2-2b shows the graded-index profile obtained when a guide is made by diffusion of metal atoms into a crystalline substrate. In this case, the index distribution below the crystal surface is a Gaussian function (as will be shown in Chapter 13), given by:

$$n(x) = n_2 + (n_1 - n_2) \exp\left[-(x - h)^2/L_D^2\right] \quad \text{for } x < h$$

where $L_D$ is a constant related to the diffusion conditions, the diffusion length.

![Figure 6.2-2 Refractive index profiles for asymmetric 1-D waveguides.](image)
However, because many of these methods are also used to fabricate more complicated channel guide structures, we shall postpone their detailed discussion, and stay with the simple slab guide for the time being.

Through total internal reflection, waves may bounce to and fro between the guide walls much as before. However, the form of solution used above - combinations of plane waves - rapidly becomes too complicated when used with dielectric guides. Instead, we shall base the analysis on the modal solution introduced in Chapter 5, and merely look for solutions moving at constant speed in the z-direction. For simplicity, we assume y-polarization once again, so that the mode is Transverse Electric, or TE. In each layer, the scalar wave equation we must solve is therefore given by:

$$\nabla^2 E_{yi}(x, z) + n_i^2 k_0^2 E_{yi}(x, z) = 0 \quad (i = 1, 2, 3)$$  \hspace{1cm} (6.2-2)

while the solutions can be written in the form:

$$E_{yi}(x, z) = E_i(x) \exp(-j\beta z) \quad (i = 1, 2, 3)$$  \hspace{1cm} (6.2-3)

The waveguide equation (which links the transverse field $E_i$ with the propagation constant $\beta$) is then given in each layer by:

$$d^2 E_i/dx^2 + [n_i^2 k_0^2 - \beta^2] E_i = 0$$  \hspace{1cm} (6.2-4)

Since we are interested in fields that are confined within the guide (which will be standing waves inside the guiding layer, and evanescent fields outside) we will assume the following trial solutions:

- In layer 1: $E_1 = E \cos(\kappa x - \phi)$
- In layer 2: $E_2 = E' \exp(\gamma x)$
- In layer 3: $E_3 = E'' \exp[-\delta(x - h)]$

Where the constants $\kappa$, $\gamma$ and $\delta$ are as given in Chapter 5, namely:

$$\kappa = \sqrt{(n_1^2 k_0^2 - \beta^2)}$$
$$\gamma = \sqrt{(\beta^2 - n_2^2 k_0^2)}$$
$$\delta = \sqrt{(\beta^2 - n_3^2 k_0^2)}$$  \hspace{1cm} (6.2-5)

Once more, we must satisfy the boundary conditions, which require continuity of $E_i(x)$ and its gradient $dE_i/dx$ at each interface. Matching fields at $x = 0$ gives the answer we found in Chapter 5, namely:

$$E' = E \cos(\phi)$$  \hspace{1cm} (6.2-7)

while matching the field gradients gives:

$$\gamma E' = -\kappa E \sin(\phi)$$  \hspace{1cm} (6.2-8)

Dividing Equations 6.2-7 and 6.2-8, we may eliminate the field amplitudes from the problem, and obtain a closed-form expression for $\phi$:

$$\tan(\phi) = \gamma/\kappa, \text{ or } \phi = \tan^{-1}(\gamma/\kappa)$$  \hspace{1cm} (6.2-9)
Similarly, matching fields at \( x = h \) gives:

\[ E'' = E \cos(\kappa h - \phi) \]

while matching field gradients gives:

\[-\delta E'' = -\kappa E \sin(\kappa h - \phi)\]

Dividing Equations 6.2-10 and 6.2-11, we therefore obtain:

\[ \tan(\kappa h - \phi) = \delta/\kappa \]

Now, self-consistent solutions are only possible when Equations 6.2-7 - 6.2-12 are all satisfied simultaneously. The most important point is that the trigonometric relations in Equations 6.2-9 and 6.2-12 are both satisfied. Once this is done, the field amplitudes \( E' \) and \( E'' \) can be found in terms of \( E \), simply by substituting values for \( \phi \) and \( \kappa \) into Equations 6.2-7 and 6.2-10.

We can reduce the two main equations to a single one, as follows. Using the standard trigonometrical identity:

\[ \tan(A - B) = (\tan A - \tan B) / (1 + \tan A \tan B) \]

we can convert the left-hand side of Equation 6.2-12 to:

\[ \tan(\kappa h - \phi) = [\tan(\kappa h) - \tan \phi] / [1 + \tan(\kappa h) \tan \phi] \]

Substituting the value of \( \tan \phi \) from Equation 6.2-9, we then obtain:

\[ [\tan(\kappa h) - \gamma/\kappa] / [1 + \tan(\kappa h) \gamma/\kappa] = \delta/\kappa \]

Finally, after some rearrangement, we get:

\[ \tan(\kappa h) = \kappa[\gamma + \delta] / [\kappa^2 - \gamma \delta] \]

This is the eigenvalue equation for the dielectric guide. Once again, it can be shown that only certain values of \( \beta \) can satisfy it, so this guide will also only support a discrete set of guided modes. Since the parameters \( \kappa, \gamma \) and \( \delta \) are all functions of the propagation constant \( \beta \), the eigenvalue equation is obviously a function of \( \beta \) as well. However, it is a transcendental equation (which means the solution cannot be written in closed form), so the \( \beta \)-values must be found numerically. This is not difficult, but before we examine the solutions themselves, we will examine the connection between the eigenvalue equation and the transverse resonance condition mentioned previously.

THE TRANSVERSE RESONANCE CONDITION

We first recall the definition of the parameter \( \phi \), from our discussion of total internal reflection in Chapter 5. \( \phi \) first appeared when the reflection coefficient \( \Gamma_i \) was written in the form:

\[ \Gamma_i = \exp(j2\phi) \]
Since this is a complex exponential, -2φ can be interpreted as a phase-shift between the incident and reflected waves (note the negative sign here). We can therefore define two new parameters -2φ_{12} and -2φ_{13} for our three-layer guide. These are the shifts that occur on reflection from the interfaces between layers 1 and 2, and layers 1 and 3, respectively. φ_{12} can be identified with our previous parameter φ, so we may write tan(φ_{12}) = γ/κ. By analogy, we can also put tan(φ_{13}) = δ/κ. Equation 6.2-12 can therefore be written as:

\[ \tan(k h - \phi_{12}) = \tan(\phi_{13}) \]  

6.2-18

Taking the inverse tangent of both sides of Equation 6.2-18, allowing for the periodic nature of the tan function (which repeats every π radians), and multiplying both sides by a factor of two, we get:

\[ 2k h - 2\phi_{12} - 2\phi_{13} = 2\nu \pi \]  

6.2-19

Equation 6.2-19 is now a transverse resonance condition, much like Equation 6.1-7. However, there are some differences. Firstly, the term kh has replaced k_0 h sin(θ). This is not a real change, since it is easy to show that κ = k_0 n_1 sin(θ), the new term therefore merely reflects the fact that the guiding layer is a medium of refractive index n_1. Secondly, the phase terms -2φ_{12} and -2φ_{13} have been introduced. The overall interpretation of Equation 6.2-19 is therefore that the total phase accumulated in bouncing between the guide walls, including the phase changes experienced on reflection, must be a whole number of multiples of 2π. The astute reader will be able to explain why this extra feature (the phase changes due to reflection) did not apparently show up in our analysis of the metal-walled guide.

THE TRANSVERSE FIELD PATTERNS

Once the values of β are found, the transverse field patterns can be drawn. It is easiest to do this for a symmetric guide, which can only support modes with symmetric or antisymmetric field patterns. In this case, it can be shown that the eigenvalue equation 6.2-16 reduces to:

\[ \tan(k h/2) = \gamma/\kappa \]  

6.2-20

for all modes with symmetric fields, and:

\[ \tan(k h/2) = -\kappa/\gamma \]  

6.2-21

for all antisymmetric modes. If Equations 6.2-20 and 6.2-21 are solved numerically, the following results are obtained. At low optical frequencies (when λ_0 is large and k_0 is low), the guide is single-moded and only symmetric patterns of the type shown in Figure 6.2-3 can be supported.

Note that the field varies cosinusoidally inside the guide, and exponentially outside, as expected. At extremely low frequencies, a significant amount of power propagates in the evanescent field. A mode of this type is described as poorly confined. As the frequency rises, however, the field concentrates more towards the centre of the guide. We shall see later that this improved confinement of the mode is advantageous, since it reduces the propagation loss that occurs when the waveguide is formed into a bend. As the frequency rises further, the guide becomes two-moded and a second, antisymmetric field solution becomes possible. This is shown in Figure 6.2-4.
CUTOFF CONDITIONS

At higher frequencies still, even more patterns are possible. We can work out how many modes can propagate at any given frequency as follows. Particular modes cease to be guided when the ray angle inside the guide tends to the critical angle (note that this is different from the metal-walled guide, where $\theta_{\text{cut-off}} = 0^\circ$). Since $\beta = n_1 k_0 \sin \theta$, and Snell's law requires $n_1 \sin \theta = n_2$ at $\theta = \theta_c$, this implies that $\beta \to n_2 k_0$ at cutoff. At the same time, we can show that $\gamma \to 0$, so the confinement of the field tends to zero at this point. For symmetric modes, the cutoff condition then reduces to:

$$\tan (\kappa h/2)_{\text{c.o.}} = 0$$  \hspace{1cm} 6.2-22

and hence:

$$(\kappa h/2)_{\text{c.o.}} = 0, \pi, 2\pi \ldots$$  \hspace{1cm} 6.2-23

Similarly, for antisymmetric modes, the cutoff condition is that

$$(\kappa h/2)_{\text{c.o.}} = \pi/2, 3\pi/2 \ldots$$  \hspace{1cm} 6.2-24

So in general we must have:
\[(\kappa h/2)_{c.o.} = \nu \pi/2\]

where \(\nu\) is again the mode number. This is the cutoff condition for all the modes in a symmetric slab guide. However, it can be expressed more conveniently with a little further manipulation, as we now show. Since \(\kappa = \sqrt{(n_1^2 k_0^2 - \beta^2)}\), then at cutoff we must have:

\[\kappa_{c.o.} = k_0 \sqrt{(n_1^2 - n_2^2)}\]

so that:

\[(\kappa h/2)_{c.o.} = (k_0 h/2) \sqrt{(n_1^2 - n_2^2)}\]

This parameter is often called the 'V' value of the guide; it is a dimensionless number, which can be used to characterise a guide - generally, a guide will be heavily multimoded if \(V \gg 1\). Combining Equations 6.2-25 and 6.2-27, we get:

\[(k_0 h/2) \sqrt{(n_1^2 - n_2^2)} = \nu \pi/2\]

This now represents the cutoff condition for all possible guided modes in a guide of width \(h\), constructed from dielectric layers of refractive indices \(n_1\) and \(n_2\). We can use it as follows. For the lowest order mode (corresponding to \(\nu = 0\)), the cutoff condition is satisfied when:

\[(k_0 h/2) \sqrt{(n_1^2 - n_2^2)} = 0\]

Since \(h\), \(n_1\) and \(n_2\) are all finite, it follows that \(k_0\) must be zero to satisfy Equation 6.2-29. Effectively, therefore, there is no cutoff for the lowest order mode; this is a particular property of the symmetric slab guide. Similarly, for the second lowest order mode (\(\nu = 1\)) cutoff occurs when:

\[(k_0 h/2) \sqrt{(n_1^2 - n_2^2)} = \pi/2\]

and so on. We can also use the cutoff condition to work out the dimensions needed for a guide to be single-moded, given the guide indices and the optical wavelength. In this case, we require the second-order mode to be just cut off. Rearranging Equation 6.2-30 slightly, we then find that:

\[h < \lambda_0/2\sqrt{(n_1^2 - n_2^2)}\]

There is therefore a strict upper limit on the allowable width of the guide.

**DESIGN EXAMPLE**

We can now put in some example numbers to illustrate typical dimensions of single-mode guides. We choose \(\lambda_0 = 0.633\ \mu m\), corresponding to the He-Ne laser wavelength, and guide indices of \(n_1 \approx n_2 \approx 1.5\); these are typical for different types of glass. With the relatively large index difference of \(n_1 - n_2 = 0.01\), we then get \(h < 1.8\ \mu m\). For the same parameters, but with the smaller index difference of \(n_1 - n_2 = 0.001\), we get \(h < 5.8\ \mu m\). Single-mode guides are clearly small in practical materials!
CHAPTER SIX

PHASE VELOCITY, GROUP VELOCITY AND WAVEGUIDE DISPERSION

Naturally enough, we will be concerned with the information-carrying capacity of waveguides. This will prove especially important for optical fibres, which we will discuss in Chapter 8. Now, our previous analysis of plane waves (in Chapter 2) showed that dispersion occurs whenever \( v_{ph} \) (or \( \omega/k \)) varies with frequency. The analogous quantity can be found for the \( \nu \)th guided mode, simply by inserting the relevant value for the propagation constant. In this case, we get:

\[
v_{ph\nu} = \frac{\omega}{\beta_{\nu}}
\]

6.2-32

It is also often convenient to define an effective index \( n_{eff\nu} \) for the mode, such that the phase velocity is given by:

\[
v_{ph\nu} = \frac{c}{n_{eff\nu}}
\]

6.2-33

A moment's thought then shows that the effective index and the propagation constant are related by:

\[
\beta_{\nu} = \frac{2\pi n_{eff\nu}}{\lambda_0}
\]

6.2-34

For a given mode, the effective index may therefore be interpreted as the refractive index of an equivalent bulk medium, which would give identical values for the phase velocity and propagation constant to those obtained in the guide.

Once more, the dispersion characteristics of modes can be represented by a graph of \( \omega \) against \( \beta \). For the symmetric slab guide, it is as shown in Figure 6.2-5. Here the lowest order mode has been labelled the TE\(_0\) mode, the second lowest the TE\(_1\) mode, and so on.

![Dispersion diagram for the symmetric slab guide.](image)

Figure 6.2-5  Dispersion diagram for the symmetric slab guide.

Note that each characteristic lies between two lines, which have slopes of \( c/n_1 \) and \( c/n_2 \), respectively, so that the phase velocity of any mode cannot be less than \( c/n_1 \) or greater than \( c/n_2 \). The former value is approached if the mode is well-confined, when most of the field is travelling inside the guide (layer 1, which has refractive index \( n_1 \)). The latter is approached as the mode tends towards cut-off, when most of the field is travelling outside the guide (in
a medium of index \( n_2 \). Consequently, the effective index of all guided modes must lie between \( n_1 \) and \( n_2 \).

Since the phase velocity clearly depends on frequency (and also on the mode number), the guide must be dispersive. For each mode, we may define a group velocity \( v_{gv} \) as:

\[
v_{gv} = \frac{d\omega}{d\beta}
\]

6.2-35

In general, the group velocity will also depend on frequency and mode number. Note that this effect (known as waveguide dispersion) is distinct from the material dispersion discussed in Chapters 2 and 3, which for simplicity was ignored in Figure 6.2-5; it forms an extra contribution to the overall dispersion, and one which may be extremely large. Often, the two components of waveguide dispersion are referred to separately, as inter- and intra-modal dispersion, respectively. The former refers to the variation in \( v_{gv} \) between the different modes, the latter to changes in a particular mode's group velocity with frequency. We shall return to both aspects in Chapter 8.

### 6.3 OTHER TYPES OF MODE

The modal solutions found so far are, of course, descriptive of light confined inside the guide. However, further solutions must exist which account for light propagating outside the guide. These are known as radiation modes, and we now describe the way they are calculated. Remember that we solved Equation 6.2-4 to find the TE modes of a slab guide. This is, of course, a general second-order differential equation, which can be written in the form:

\[
\frac{d^2E}{dx^2} + C_i^2 E_i = 0
\]

6.3-1

In our previous analysis, the form of the solution was exponential or sinusoidal, depending on the sign of the term \( C_i^2 = (n_i^2 k_0^2 - \beta^2) \). If we consider all possible values of \( \beta \), it turns out that a wider range of solutions can be found. If we again take \( n_1 > n_2 > n_3 \), which is often the case in real guides, the complete set can be represented as a diagram in \( \beta \)-space as shown in Figure 6.3-1. The essential features of the diagram are that:

1) For \( \beta > k_0 n_1 \), the solutions are exponential in all three layers. Since this implies infinite field amplitudes at large distances from the guide (which are physically unrealistic) we will ignore these solutions for the time being.

2) For \( k_0 n_1 > \beta > k_0 n_2 \), there are a discrete number of bound or guided modes, which are the solutions already found. These vary cosinusoidally inside the guide core, and decay exponentially outside the guide.

3) For \( k_0 n_2 > \beta > k_0 n_3 \), the solutions vary exponentially in the cover (layer 3), and cosinusoidally in both the guide (layer 1) and substrate (layer 2). Since these fully penetrate the substrate region, they are called substrate modes. Any value of \( \beta \) is allowed, between the two limits given above, so the set forms a continuum.

4) For \( k_0 n_3 > \beta \), solutions vary cosinusoidally in all three layers. These particular field patterns are known as radiation modes. Once again, any value of \( \beta \) is allowed in the range above, so the set forms another continuum.
The introduction of radiation modes suggests that we should alter our previous dispersion diagram to account for them. Equally, we should allow for the possibility of backward-travelling modes (which have negative values of $\beta$). Both modifications are simple to carry out, and the complete set of modes for a symmetric slab guide is then as shown in the modified $\omega - \beta$ diagram of Figure 6.3-2. We will find this representation useful in the solution of coupled mode problems in Chapter 10.
THE MAGNETIC FIELD PATTERNS

Having found the complete set of all possible solutions for the electric field, it is simple to obtain the corresponding magnetic field patterns. This gives a self-consistent, vectorial solution for the entire guided electromagnetic field. First, recall that the TE solution for $E$ is a $y$-polarized guided mode, written as $E = j E(x) \exp(-j\beta z)$. Now, from the curl relation between $E$ and $H$, we can show that the associated magnetic field has two components, given by:

$$H_z = (\frac{1}{j \omega \mu_0}) \frac{\partial}{\partial x} E_y = (-1/j \omega \mu_0) \frac{\partial E_y}{\partial x} = (-1/j \omega \mu_0) \frac{d}{dx} E \exp(-j\beta z)$$

6.3-2

and:

$$H_y = \frac{\beta}{\omega \mu_0} E(x) \exp(-j\beta z)$$

6.3-3

Unlike the electric field, the magnetic field therefore has a longitudinal component (i.e. a component in the direction of propagation) as well as a transverse one.

THE VECTORIAL REPRESENTATION OF MODES

It is often convenient to represent both the fields in terms of vectorial transverse fields. Thus, for the $\nu^{th}$ mode, we would write:

$$E = E_{\nu} \exp(-j\beta \nu z) \quad \text{and} \quad H = H_{\nu} \exp(-j\beta \nu z).$$

6.3-4

where $E_{\nu}$ and $H_{\nu}$ are vectorial descriptions of the transverse fields. This notation will prove useful in some of the later proofs in this section. We may find expressions for $E_{\nu}$ and $H_{\nu}$ for TE modes by comparing Equation 6.3-4 with the results above. If this is done, we get:

$$E_{\nu} = j E_{\nu} \quad \text{and} \quad H_{\nu} = i \left( \frac{\beta}{\omega \mu_0} \right) E_{\nu} + k \left( \frac{1}{j \omega \mu_0} \right) \frac{d}{dx} E_{\nu} \exp(-j\beta \nu z)$$

6.3-5

We can see from Equation 6.3-5 that (unlike the transverse electric field, which is real), the transverse magnetic field is complex, since $H_{\nu z}$ is imaginary.

TM MODES

A similar analysis can be performed for the case when the magnetic field lies in the $y$-direction. The field patterns are then known as Transverse Magnetic, or TM modes, and are complementary to the TE solutions already found. The results are qualitatively similar, but this time the eigenvalue equation for guided modes takes the form:

$$\tan(\kappa h) = \kappa (\gamma' + \delta') / (\kappa^2 - \gamma \delta')$$

6.3-6

where:

$$\gamma' = \gamma \left( n_2^2/n_1^2 \right) \quad \text{and} \quad \delta' = \delta \left( n_2^2/n_1^2 \right)$$

6.3-7

Broadly speaking, the behaviour of TM modes is similar to that of TE modes. However, detailed analysis shows that the field profiles no longer have the smooth, continuous form found earlier. Similarly, the detailed shape of the dispersion diagram is slightly different, because of the refractive index ratios that appear in Equation 6.3-7. However, if the index differences are small (so that the ratios approach unity) the two are practically identical.
6.4 THE WEAK-GUIDANCE APPROXIMATION

The vectorial analysis used so far is very complicated, and a simpler approach is often desirable. This is especially true for the two-dimensional refractive index distributions that occur in channel guides or optical fibres. Luckily, it turns out that if the index difference \( \Delta n \) forming the guide is small, a scalar approximation can be used instead. Since this is often the case in practical guides, the approximation is a very useful one. It is based on the observation that, for a small value of \( \Delta n \), the critical angle \( \theta_c \) at a single interface is very large. Hence, for a guided mode, the ray directions inside the guide are almost parallel to the axis of propagation, as shown in Figure 6.4-1.

Since the total field inside the guide may be considered as a summation of similar bouncing waves, the combined E- and H-vectors must be almost exactly orthogonal to each other, and to the axis of propagation. This will be the case anyway for our TE mode, assuming that \( \partial E/\partial x \) is small in Equation 6.3-3 (the physical justification being that the field inside a weak guide must be a fringe pattern of large periodicity). The resulting modes are known as TEM modes (implying that both the electric and the magnetic field components are approximately transverse).

![Figure 6.4-1 Field directions for total internal reflection at a low-\( \Delta n \) interface.](image)

As it turns out, the following analysis can also be used to treat two-dimensional guides of arbitrary cross-section, provided the index changes involved are small. They may also be gradual changes, rather than the discrete steps we have encountered so far. To illustrate this, let us consider a guide formed by a 2-D index distribution \( n(x, y) \), lying in the z-direction as shown in Figure 6.4-2.

![Figure 6.4-2 A two-dimensional waveguide.](image)

If we pick one field component to work with (say \( E \)), then this can point in either the x- or the y-direction. For example, let us suppose that it points in the y-direction. The scalar amplitude \( E_y(x, y, z) \) then satisfies (at least, approximately) the three-dimensional scalar wave equation:
We now assume a modal solution in the usual form, namely as a product of a transverse electric field distribution and an exponential propagation term. This time, the transverse field must also be two-dimensional, so we take:

$$E_y(x, y, z) = E(x, y) \exp(-j \beta z)$$

where $E(x, y)$ is the transverse electric field. Substituting into Equation 6.4-1, we obtain the following waveguide equation linking $E$ with $\beta$:

$$\nabla^2 E(x, y) + [n^2(x, y)k_0^2 - \beta^2] E(x, y) = 0$$

Here the subscripts on the Laplacian indicate that differentiation is to be performed with respect to $x$ and $y$ only. Though Equation 6.4-3 is very similar to our previous waveguide equation, it is also valid for 2-D guides. Although several standard techniques exist for solving this type of equation, we will postpone any attempt at a solution until Chapter 8. For now, we simply note that a complete set of two-dimensional modes - both guided and radiation modes - can exist. However, we can assume directly that the transverse magnetic field points in the x-direction, and that its amplitude can be found from that of the transverse electric field as:

$$H(x, y) = -(\beta/\omega \mu_0) E(x, y)$$

The two transverse fields are therefore related by a simple scaling factor.

**6.5 THE ORTHOGONALITY OF MODES**

Since the modes above form a complete set of solutions to Maxwell’s equations for a waveguide geometry, we must be able to describe an arbitrary field by a weighted summation of all types of mode. Before we detail how this is done, we must examine some further general properties of the solutions. The most useful one is that the modes are orthogonal to each other. This means that an integral over the whole guide cross-section, of a product of the transverse field of two modes, is zero. This can be proved using the power conservation theorem, as we now show.

The power propagating in a guide can be found by integrating the normal component of the time-averaged Poynting vector over the guide cross-section. For a guide lying in z-direction, we obtain:

$$P_z = 1/2 \text{Re} \left[ \int \int_A (E \times H^*) \cdot k \, dx \, dy \right]$$

Since power is conserved in a lossless system, this implies that $dP_z/dz = 0$, or:

$$d/dz [1/4 \int \int_A (E \times H^* + E^* \times H) \cdot k \, dx \, dy] = 0$$

where we have used the identity $\text{Re}(z) = 1/2 (z + z^*)$. Let us now assume that any field can be written as a sum of modes. For simplicity, we restrict the expansion to forward-travelling guided modes (although generally we should include backward-travelling and radiation modes as well). Suppose that the $\mu^\alpha$ mode has transverse fields $E_{\mu}^\alpha$ and $H_{\mu}^\alpha$, and propagation constant $\beta_{\mu}^\alpha$. We can then write:
Here the terms $a_{\mu}$ are expansion coefficients, which describe the amplitude of each mode contributing to the total field. Substituting these expressions into Equation 6.5-2, we get:

$$E = \sum_{\mu} a_{\mu} E_{\mu} \exp(-j\beta_{\mu} z) \quad \text{and} \quad H = \sum_{\mu} a_{\mu} H_{\mu} \exp(-j\beta_{\mu} z) \quad (6.5-3)$$

Or:

$$\sum_{\mu,\nu} a_{\mu} a_{\nu}^* \exp[j(\beta_{\nu} - \beta_{\mu}) z] \int_{\Lambda} [E_{\mu}^* H_{\nu} + E_{\nu}^* H_{\mu}] \cdot k \, dx \, dy = 0 \quad (6.5-4)$$

Now, the coefficients $a_{\mu}$ and $a_{\nu}^*$ are arbitrary. Equally, the propagation constants of different modes are unequal, so that $\beta_{\mu} \neq \beta_{\nu}$. The only way Equation 6.5-5 can be satisfied for all $z$ is therefore if:

$$\int_{\Lambda} (E_{\mu}^* H_{\nu} + E_{\nu}^* H_{\mu}) \cdot k \, dx \, dy = 0 \quad (6.5-5)$$

This called the orthogonality condition. It is also possible to show (by including backward-travelling modes as well) that each of the two parts of the integral are themselves zero, so that:

$$\int_{\Lambda} (E_{\mu}^* H_{\nu} + E_{\nu}^* H_{\mu}) \cdot k \, dx \, dy = 0 \quad (6.5-6)$$

We will now show how this condition is used in practice, using a number of examples.

**DESIGN EXAMPLE # 1**

We begin with the case of TE modes in a slab guide, for which there is clearly no need to integrate in the $y$-direction. The orthogonality condition therefore reduces to:

$$\int_{x} (E_{\mu}^* H_{\nu}^*) \cdot k \, dx = 0 \quad (6.5-7)$$

Or, in terms of the relevant vectorial components:

$$\int_{x} -E_{\mu y} H_{\nu x}^* \, dx = 0 \quad (6.5-8)$$

Substituting for $H_{\nu x}$, and assuming that $E_{\nu y}$ is real, we then obtain:

$$\int_{x} (E_{\mu y} E_{\nu y}) \, dx = 0 \quad (6.5-9)$$

In other words, the integral of the product of two different modes, when taken over the cross-section of the guide, is zero. This is confirmed by Figure 6.2-4, which shows the field distributions for the lowest-order symmetric and antisymmetric modes in a symmetric guide. Clearly, the integral of the product of these two functions must be zero.
DESIGN EXAMPLE # 2

We can obtain a similar result for TEM modes in a weak two-dimensional graded-index guide, by using the relation between the transverse electric and magnetic fields given in Equation 6.4-4. Substituting into Equation 6.5-7, and making no assumption that the fields are real, we get:

\[ \iint_A (E_{\mu y} E_{\nu y}^*) \, dx \, dy = 0 \]  \hspace{1cm} 6.5-11

In both cases described above, the orthogonality condition is often written in shorthand as:

\[ \langle E_{\mu y} , E_{\nu y} \rangle = 0 \]  \hspace{1cm} 6.5-12

The bracket notation employed here is often used in quantum mechanics. The operation of multiplying two functions together and integrating over a range is called an inner product.

6.6 THE POWER CARRIED BY A MODE

We can use the orthogonality relation to define the power carried down the guide by each mode. From the previous section, we know that the total power flow \( P_z \) in the z-direction is given by:

\[
P_z = \frac{1}{4} \iint_A \sum_{\mu,\nu} \{ a_\mu a_\nu^* \exp[j(\beta_\nu - \beta_\mu)z] E_\mu \times H_\nu^* + a_\nu^* a_\mu \exp[j(\beta_\nu - \beta_\mu)z] E_\nu^* \times H_\mu \} \cdot \mathbf{k} \, dx \, dy
\]  \hspace{1cm} 6.6-1

However, from the orthogonality relation, we also know that many of the terms in Equation 6.6-1 must be zero. Removing these gives:

\[
P_z = \sum_\mu a_\mu a_\mu^* \left[ \frac{1}{4} \iint_A \{ E_\mu \times H_\mu^* + E_\mu^* \times H_\mu \} \cdot \mathbf{k} \, dx \, dy \right]
\]  \hspace{1cm} 6.6-2

We can now simplify this expression by defining a modal power coefficient \( P_\mu \), such that:

\[
P_\mu = \frac{1}{4} \iint_A \{ E_\mu \times H_\mu^* + E_\mu^* \times H_\mu \} \cdot \mathbf{k} \, dx \, dy
\]  \hspace{1cm} 6.6-3

In which case, we obtain:

\[
P_z = \sum_\mu a_\mu a_\mu^* P_\mu
\]  \hspace{1cm} 6.6-4

This is a very simple result. It implies that the power carried by forward-travelling modes is found as the sum of the modulus-squares of the mode amplitudes, weighted by the coefficients \( P_\mu \).

DESIGN EXAMPLE

Calculation of the modal power coefficients is quite simple. For TE modes in a slab guide, for example, we can define \( P_\mu \) per unit width in the y-direction as:

\[
P_\mu = \frac{1}{4} \int_x \{ E_\mu \times H_\mu^* + E_\mu^* \times H_\mu \} \cdot \mathbf{k} \, dx
\]  \hspace{1cm} 6.6-5
Inserting the relevant vectorial components, and substituting for $H_{\mu x}$ using Equation 6.3-5, we get:

$$P_{\mu} = \left(\frac{\beta}{4\omega_{\mu_0}}\right) \int_x (E_{\mu y}^*E_{\mu y} + E_{\mu y}E_{\mu y}^*) \, dx$$

6.6-6

It is obvious that each component in the integral above contributes an equal amount to the total. Furthermore, the transverse fields $E_{\mu y}$ are real in this case. We therefore obtain:

$$P_{\mu} = \left(\frac{\beta}{2\omega_{\mu_0}}\right) \int_x E_{\mu y}^2 \, dx$$

6.6-7

Note that this can of course be written in inner product notation, as $P_{\mu} = \left(\frac{\beta}{2\omega_{\mu_0}}\right) \langle E_{\mu y}, E_{\mu y} \rangle$.

6.7 THE EXPANSION OF ARBITRARY FIELDS IN TERMS OF MODES

The orthogonality condition can be also used to show how arbitrary transverse fields can be written in terms of the modal solutions, by finding the expansion coefficients themselves. Suppose we wish to expand a forward-travelling field $E(x, y, 0)$ at $z = 0$ in this way. We first put:

$$E(x, y, 0) = \sum_{\mu} a_{\mu} E_{\mu}(x, y)$$

6.7-1

If we now take vector products of both sides of Equation 6.7-1 with the complex conjugate of the transverse magnetic field of one mode (say, the $\nu^{th}$), and integrate over the cross-section, we get:

$$\iint_A [E \times H_{\nu}^*] \cdot k \, dx \, dy = \iint_A \sum_{\mu} a_{\mu} [E_{\mu} \times H_{\nu}^*] \cdot k \, dx \, dy$$

6.7-2

Now, since the field is forward travelling, we could also have put:

$$H = \sum_{\mu} a_{\mu} H_{\mu}$$

6.7-3

Performing a similar operation, but this time using the complex conjugate of the transverse electric field of the $\nu^{th}$ mode, we get:

$$\iint_A [E_{\nu}^* \times H] \cdot k \, dx \, dy = \iint_A \sum_{\mu} a_{\mu} [E_{\nu}^* \times H_{\mu}] \cdot k \, dx \, dy$$

6.7-4

Adding Equations 6.7-2 and 6.7-4 together then gives:

$$\iint_A [E \times H_{\nu}^* + E_{\nu}^* \times H] \cdot k \, dx \, dy = \iint_A \sum_{\mu} a_{\mu} [E_{\mu} \times H_{\nu}^* + E_{\nu}^* \times H_{\mu}] \cdot k \, dx \, dy$$

6.7-5

Now, from the orthogonality condition, we know that the right-hand side of Equation 6.7-5 must be zero, unless $\mu = \nu$. Hence, we can rearrange it to extract the modal coefficient $a_{\nu}$ as:

$$a_{\nu} = \frac{\iint_A [E \times H_{\nu}^* + E_{\nu}^* \times H] \cdot k \, dx \, dy}{\iint_A [E_{\nu} \times H_{\nu}^* + E_{\nu}^* \times H_{\nu}] \cdot k \, dx \, dy}$$

6.7-6

The coefficients can therefore be found by evaluating Equation 6.7-6 for each mode in turn. Once more, we shall illustrate this with an example.
DESIGN EXAMPLE

To expand a TE field $E_y(x)$ in terms of the TE modes of a slab guide, we follow the usual procedure of dropping the integration with respect to the $y$-variable. Substituting the relevant vectorial components into Equation 6.7-6, we get:

$$a_\nu = \frac{\int_x E_y E_{\nu y} \, dx}{\int_x E_{\nu y}^2 \, dx}$$

Note that the integral in the denominator of Equation 6.7-7 really just represents a normalisation factor. The integral in the numerator is more important; it is often called an overlap integral, because it represents the 'overlap' of the two fields $E_y$ and $E_{\nu y}$. This is really a measure of how similar the total field is to the $\nu$th mode, rather like a correlation. Clearly, if the two are actually identical, we obtain $a_\nu = 1$ (and $a_\mu = 0$ for $\nu \neq \mu$). This is common sense. Finally, we note that Equation 6.7-7 can be written in inner product notation, as $a_\nu = <E_y, E_{\nu y}> / <E_{\nu y}, E_{\nu y}>$.

6.8 APPLICATION OF THE OVERLAP PRINCIPLE

Overlap integrals are particularly useful in calculating the coupling efficiency in end-fire couplers, when an external field is used to excite a guided mode. They can also be used to explain what happens when a mode passes through a waveguide discontinuity or a taper. We shall postpone a description of end-fire coupling until Chapter 8, concentrating for now on discontinuities and tapers.

DISCONTINUITIES

Figure 6.8-1 shows a number of different waveguide discontinuities. In each case, the substrate is continuous, but there is some difference in the guiding layer on either side of the junction.

![Figure 6.8-1](image)

This could involve a change in layer thickness, or the addition of an overlay of a different material. Both can be achieved easily using conventional planar processing techniques. Alternatively, the orientation of the guide can change. This requires machining of the substrate surface to the desired topology before deposition of the guiding layer. A junction can also be formed between two entirely different guides, with no common substrate. In this case, the process is called butt coupling.
In both cases, Equation 6.7-6 can be used to calculate the effect of the discontinuity. For example, let us assume in a TEM model that the set of modes on the input side of the discontinuity have transverse field distributions $E_{\nu y}$. On the other side, the corresponding fields are $E_{\nu y'}$. We will now show how to find the efficiency with which an input mode (for example, the lowest-order mode $E_{0y}$) is coupled across the junction. We may take the input field at the junction as $a_0 E_{0y}$, where $a_0$ is the input mode amplitude. The power $P_{in}$ carried by this field is then:

$$P_{in} = \left( \frac{\beta}{2 \omega \mu_0} \right) |a_0|^2 < E_{0y} , E_{0y} >$$  \hspace{1cm} (6.8-1)

After the junction, the portion of the total field carried by the lowest-order mode is now $a_0'E_{0y}'$, where $a_0'$ is the new mode amplitude. This can be found as:

$$a_0' = <a_0 , E_{0y} , E_{0y}'> / < E_{0y} , E_{0y}'>$$  \hspace{1cm} (6.8-2)

The power $P_{out}$ carried by this field is (by analogy with Equation 6.8-1):

$$P_{out} = \left( \frac{\beta}{2 \omega \mu_0} \right) |a_0'|^2 < E_{0y}' , E_{0y}' >$$  \hspace{1cm} (6.8-3)

Substituting for $a_0'$, this reduces to:

$$P_{out} = \left( \frac{\beta}{2 \omega \mu_0} \right) |a_0|^2 |< E_{0y} , E_{0y}'>|^2 / < E_{0y} , E_{0y}'>$$  \hspace{1cm} (6.8-4)

If we now define the coupling efficiency $\eta$ as the ratio $P_{out}/P_{in}$, we obtain:

$$\eta = \frac{|< E_{0y} , E_{0y}'>|^2}{< E_{0y} , E_{0y}'> < E_{0y}' , E_{0y}' >}$$  \hspace{1cm} (6.8-5)

Clearly, $\eta$ is a dimensionless number, which tends to unity as $E_{0y}$ tends to $E_{0y}'$. Although this is what one would intuitively expect, the implication is that the input mode will cross a discontinuity without much in the way of conversion to other modes, merely provided it is small enough. We can also deduce that the most effective waveguide joint will be one between two guides with matched fields. Though we have not included radiation modes in the calculation, this would be required in a rigorous model; it would then be found that some power is lost to radiation at any discontinuity.

**TAPERS**

We can extend this principle to discover what happens in a tapered waveguide. The upper diagram in Figure 6.8-2 shows a typical slow taper, which involves a gradual change in the guide cross-section. The lower one shows a discrete approximation to this shape, consisting of a series of steps. Once again, we might wish to discover how a mode of a particular order travels through the taper. If we can avoid any mode conversion, we might then be able to design the structure to enlarge the cross-section of a particular mode in a slow and controlled manner. The taper would then act as a beam expander, for use as a matching section between two different guides.

The theory we need to analyse the problem is called the method of local normal modes. Using a TEM model, the strategy is to proceed as follows. We start by arranging that the steps in our discrete approximation are at regular intervals $\Delta z$ apart. We then consider the field at three different points. The first is point 1, just to the left of the discontinuity at $z = 0$ in Figure 6.8-2. Here it is assumed that the local normal modes have transverse field functions $E_{\mu y}$ and propagation constants $\beta_\mu$. 

R.R.A.Syms and J.R.Cozens  
Optical Guided Waves and Devices  
19
If we expand the field $E_y$ in terms of these modes, we can write:

$$E_y = \sum_{\mu} a_{\mu} E_{\mu y}$$  \hspace{1cm} 6.8-6

The second place of interest is point 2, just to the right of the discontinuity at $z = 0$. In this region, the local transverse fields and propagation constants have the slightly different values of $E_{\mu y}'$ and $\beta_{\mu}'$. If we expand the same field in terms of these modes instead, we can put:

$$E_y = \sum_{\mu} b_{\mu} E_{\mu y}'$$  \hspace{1cm} 6.8-7

Here the terms $b_{\mu}$ are an entirely different set of mode amplitudes, but we can relate them to our original $a$-values using the overlap principle. If this is done, we get:

$$b_{\mu} = \sum_{\nu} a_{\nu} \langle E_{\nu y}, E_{\mu y}' \rangle / \langle E_{\mu y}', E_{\mu y}' \rangle$$  \hspace{1cm} 6.8-8

This relation between the $b$-values and the $a$-values effectively tells us how the mode amplitudes change simply in crossing the step at $z = 0$. Now, the next place of interest is point 3. This is just to the left of the discontinuity at $z = \Delta z$, and we call field here $E_y'$. This can be found by propagating the field at point 2 over the distance $\Delta z$ between points 2 and 3. All that happens is that the phase of each mode changes slightly en route, so we must have:

$$E_y' = \sum_{\mu} b_{\mu} E_{\mu y}' \exp(-j\beta_{\mu}'\Delta z)$$  \hspace{1cm} 6.8-9

Now, our aim is to see how the mode amplitudes change with distance. We should, therefore, have expressed the field at point 3 in a form comparable with that used at point 1, i.e. as:

$$E_y' = \sum_{\mu} a_{\mu}' E_{\mu y}'$$  \hspace{1cm} 6.8-10

Here the terms $a_{\mu}'$ represent the amplitudes of the local normal modes at the start of the next discontinuity at $z = \Delta z$. These new amplitudes now can be extracted as follows. First, we equate 6.8-10 and 6.8-9, which gives:

$$a_{\mu}' = b_{\mu} \exp(-j\beta_{\mu}'\Delta z)$$

Figure 6.8-2  Continuous and discrete models of a waveguide taper.
Then we substitute for $b_\mu$ using Equation 6.8-8, to get:

$$a_\mu' = \{ \sum_\nu a_\nu < E_\nu y, E_\nu y' > / < E_\nu y', E_\nu y' > \} \exp(-j\beta_\mu' \Delta z)$$

At this point, we note that the quantity $a_\mu' - a_\mu$ is actually the change in $a_\mu$ that has occurred over the distance $\Delta z$. We might call this $\Delta a_\mu$. Using Equation 6.8-12, we may write it as:

$$\Delta a_\mu = \{ a_\mu < E_\mu y, E_\mu y' > / < E_\mu y', E_\mu y' > \} \exp(-j\beta_\mu' \Delta z) - 1$$

We now make a number of approximations, which will be valid provided the distance $\Delta z$ is short enough and the discontinuity is small. We start by assuming that the transverse modal field changes only a small amount in the distance $\Delta z$, so we can put $E_\mu y' = E_\mu y + \Delta E_\mu y$. To reasonable accuracy, we can then say that:

$$< E_\mu y, E_\mu y' > / < E_\mu y', E_\mu y' > \approx 1$$

Using the orthogonality principle, we may also show that:

$$< E_\nu y, E_\mu y' > / < E_\mu y', E_\mu y' > \approx < E_\nu y, \Delta E_\mu y > / < E_\mu y, E_\mu y >$$

We now assume that the propagation constant also changes only a small amount in the distance $\Delta z$, so we can put $\beta_\mu' = \beta_\mu + \Delta \beta_\mu$. To first-order approximation, we can then write:

$$\exp(-j\beta_\mu' \Delta z) \approx (1 - j\beta_\mu \Delta z) \approx 1$$

However, we note that a slightly different approximation must be used for the other exponential term in Equation 6.8-13, because it is the difference between two similar quantities. We therefore write:

$$\exp(-j\beta_\mu' \Delta z) - 1 = -j\beta_\mu \Delta z$$

When all these values are substituted into Equation 6.8-13, it reduces to:

$$\Delta a_\mu = -j\beta_\mu \Delta z a_\mu + \sum_{\nu, \nu \neq \mu} a_\nu < E_\nu y, \Delta E_\mu y > / < E_\mu y', E_\mu y >$$

Dividing both sides by $\Delta z$, and letting $\Delta z$ tend to zero, we then get:

$$\frac{da_\mu}{dz} = -j\beta_\mu a_\mu + \sum_{\nu, \nu \neq \mu} a_\nu < E_\nu y, \frac{\partial E_\mu y}{\partial z} > / < E_\mu y', E_\mu y >$$

Where $a_\mu$ and $\beta_\mu$ are functions of $z$, and $E_\mu y$ is a function of $x$, $y$ and $z$. Equation 6.8-19 is now a differential equation, which describes the way the amplitudes of the local normal modes change with distance. In fact, there is one such equation for each value of $\mu$ (i.e., for each mode of interest) and all the equations must be solved simultaneously for a given set of boundary conditions. Because each mode amplitude appears in every equation, they are known as coupled mode equations.

In principle, the equations can be solved as follows. The local normal modes must first be found at all points along the taper, so that the functions $E_\mu y(x, y, z)$ and $\beta_\mu(z)$ are known. The equations are then integrated numerically. However, this process is exceptionally tedious, so we will not perform it here; instead, we will simply draw some broad
conclusions. Firstly, we note that the solution for a particular mode must be $a_n = A_n \exp(-j\beta_n z)$ (where $A_n$ is a constant) if the summation in Equation 6.8-19 is zero. Under these conditions, the mode amplitude does not change, only its phase. To design a conversion-free taper, we must simply arrange for it to be slow enough, so that:

$$< E_{vy}, \frac{\partial E_{\nu y}}{\partial z} > \rightarrow 0 \quad 6.8-20$$

Even if the taper does not satisfy these conditions, we can extract a significant result from the mathematics. We have already seen that the transverse fields of a symmetric slab guide are either symmetric or antisymmetric patterns. If we taper such a guide symmetrically, the same must be true for the local normal modes. More importantly, we can also say that $\frac{\partial E_{vy}}{\partial z}$ must be symmetric for a symmetric mode, and antisymmetric otherwise. It therefore follows that $< E_{vy}, \frac{\partial E_{vy}}{\partial z} > = 0$ if one of the modes is symmetric and the other antisymmetric. Consequently, such a taper cannot convert a symmetric mode to an antisymmetric one, or vice versa. We will find this rule useful in Chapter 9.
CHAPTER SIX

PROBLEMS

6.1. Sketch the dispersion diagram for guided modes in a metal-walled waveguide.

6.2. Find the cutoff frequency of a planar waveguide, formed from a slab of dielectric with metallized walls. The slab has thickness 0.5 µm and refractive index 1.5. [2 x 10^{14} Hz]

6.3. Show that the eigenvalue equation for symmetric modes in a symmetric slab dielectric waveguide is given by: \( \tan(\kappa h/2) = \gamma/\kappa \), where \( \gamma, \kappa \) and \( h \) have their usual meanings.

6.4. A symmetric slab guide is to be used at 1.5 µm wavelength. What is the V-value of the guide, if its thickness is 6 µm, and the refractive indices of the layers are defined by \( n_1 \approx n_2 \approx 1.5 \), and (a) \( n_1 - n_2 = 0.05 \), and (b) \( n_1 - n_2 = 0.005 \)? How many modes are supported in each case? [(a) 4.866; (b) 1.539; 4 modes; 1 mode]

6.5. Show that the eigenvalue equation for TM modes in an asymmetric slab dielectric waveguide is given by: \( \tan(\kappa h) = \kappa(\gamma' + \delta') / (\kappa^2 - \gamma'\delta') \), where \( \gamma, \kappa \) and \( h \) have their usual meanings, and \( \gamma' \) and \( \delta' \) are given by: \( \gamma' = \gamma(n_1^2 / n_2^2) \) and \( \delta' = \delta(n_1^2 / n_3^2) \).

6.6. Sketch the variation of the transverse magnetic field for the lowest-order TM mode of a symmetric slab dielectric waveguide (a) far from cutoff, and (b) near to cutoff.

6.7. A symmetric slab guide has refractive indices \( n_1 = 1.505, n_2 = 1.5 \). If the guide thickness is 1 µm, estimate the effective index of the lowest-order TE mode of the guide at \( \lambda_0 = 0.633 \) µm. [1.50125]

6.8. Show by direct integration that two symmetric TE modes of a symmetric slab dielectric waveguide are orthogonal to one another, if the modes are of different order.

6.9. Two identical planar waveguides are butt-coupled together, as shown below. Assuming that the transverse variation of the electric field in the left-hand guide is \( E_y(x) = E_0 \exp(-\alpha x^2) \), calculate and sketch the dependence of the coupling efficiency on any accidental lateral misalignment \( \delta \).

6.10. (i) Figure I below shows a discontinuity between two different, symmetric slab dielectric waveguides. Describe qualitatively what you expect to occur when the lowest-order guided mode is incident from the left, assuming that the discontinuity is (a) small, and (b) large.
(ii) The two guides are now joined by a taper section, as shown in Figure ii). Describe what will occur under similar circumstances, if the taper is (a) fast and (b) slow.
CHAPTER SIX

SUGGESTIONS FOR FURTHER READING

Lee D.L. "Electromagnetic principles of integrated optics" John Wiley and Sons, New York (1986); Chapter 4.
7.1 OVERVIEW OF PLANAR WAVEGUIDE COMPONENTS

We will now consider the range of devices that comprise the system of planar waveguide integrated optics. Since these operate on sheet beams (i.e., beams that are confined in one direction only), they are analogous to the devices used to manipulate free-space beams. Consequently, the major building blocks are refractive components (generally, lenses) and diffractive components (known as gratings). However, we must also consider the important question of input and output couplers, needed to connect the guided wave system to the outside world.

Before proceeding further, we must first ask what information is needed to understand the behaviour of an optical device. The general answer is that we will be happy if we can predict the directions and amplitudes of the output beams, for any given input. Some geometries are relatively simple, and yield this information quite readily - for example, in Chapter 5 we found the complete solution to the single interface problem. Others are less tractable, and force us to accept a partial solution. Many integrated optic devices fall into this 'difficult' category. The beam directions are easy to predict, but often the amplitudes can only be found after considerable mathematics. We will therefore concentrate on a simple method of finding the beam directions, based on the principle of phase matching.

7.2 PHASE MATCHING AT A SINGLE INTERFACE

We begin by revisiting the interface problem mentioned above. Figure 7.2-1 shows reflection and refraction at the interface between two semi-infinite media. Here a beam of wavelength $\lambda_0$ is incident at an angle $\theta_1$, and the two media have refractive indices $n_1$ and $n_2$ respectively. We assume that $n_1$ is greater than $n_2$, but that $\theta_1$ is less than the critical angle so that a propagating transmitted wave arises.

![Figure 7.2-1 Phase-matching at a single interface – transmission and reflection.](image)

The procedure in any phase-matching problem is the same. We start by identifying all the propagation constants of interest. Here there are only two, with values $k_1$ in Layer 1 and $k_2$ in Layer 2, given by:
CHAPTER
SEVEN

\[ k_1 = \frac{2\pi n_1}{\lambda_0} = k_0 n_1 \quad \text{and} \quad k_2 = k_0 n_2 \]

We then consider the input wave. Assuming TE incidence, the electric field of the input beam only has a y-component. We know from past experience that this can be written in the form:

\[ E_{yi} = E_i \exp\{-jk_0 n_1 [z \sin(\theta_i) - x \cos(\theta_i)]\} \]

Now, however, we note that it could be written in shorthand form, as:

\[ E_{yi} = E_i \exp(-jk \cdot r) \]

Here \( r \) is the radius vector, while \( k \) (known as the propagation vector) has components:

\[ k_i = -k_1 \cos(\theta_i) i + k_1 \sin(\theta_i) k \]

It is easy to demonstrate equivalence between the two forms, simply by evaluating the scalar product term in Equation 7.2-3.

Now, examining Equation 7.2-4, we see that the propagation vector is oriented parallel to the direction of travel of the wave, and has modulus:

\[ |k_i| = \sqrt{(k_{ix}^2 + k_{iz}^2)} = k_1 \]

Since we are only really interested in beam directions, this vector contains all the information we need - in fact, it corresponds roughly to the ray introduced in our earlier discussion of bulk optical devices. What properties does it have? Well, if one end of the vector is fixed, the locus of all possible positions of the other end in the x-z plane must be a circle, of radius \( k_i \). Figure 7.2-2 shows two equally valid representations of this circle. In three dimensions, the locus is a sphere. For the range of possible input beams in our interface problem, the locus is adequately represented by the upper half-circle in Figure 7.2-1. This can be drawn with a radius proportional to \( k_1 \) (or to \( n_1 \)).

Similarly, we can represent the reflected beam by the vector \( k_r \), and the transmitted beam by \( k_t \). The moduli of these two vectors are \( |k_r| = k_1 \) and \( |k_t| = k_2 \), so the locus of the free end of \( k_r \) is also defined by the upper half circle in Figure 7.2-1, while that of \( k_t \) may be taken as the lower half-circle. Note that the radius of this second circle is proportional to \( k_2 \) (and thus to \( n_2 \)), and is therefore smaller than the upper one, since \( n_2 < n_1 \). We will now use these vectors to interpret our previous results in a new way. Firstly, we know that the incident and reflected beams make equal angles with the interface normal, so the z-components of their wave vectors must be the same. We may therefore put:

\[ k_{iz} = k_{iz} \]

Similarly, we know that the angle \( \theta_2 \) of the transmitted wave is governed by Snell's law, \( n_1 \sin(\theta_i) = n_2 \sin(\theta_2) \). This requires that the z-component of the transmitted wave vector is also the same, so:

\[ k_{iz} = k_{iz} \]

Together, these equations imply that the z-components of all three vectors are identical. This is consistent with our earlier result that all parts of the field travel at the same speed in the z-direction. However, it has given us a new way to find the beam directions.
We first draw circles of the correct relative radii, together with the propagation vector of the incident beam. We then add in the reflected beam vector, and compute the transmitted beam vector from the intersection with the lower circle of the interface normal through the end of the reflected wave vector. Of course, $\theta_1$ may be greater than the critical angle. In this case, total internal reflection occurs, and the dashed-line construction yields no intersection at all with the lower circle (as in Figure 7.2-3). Our interpretation of this event is that the field in Layer 2 cannot now consist of a propagating wave; instead, it must be a boundary wave of some kind, travelling parallel to the interface.

From previous experience, we recognise it to be an evanescent wave, with propagation vector $\vec{\beta}$. The direction of $\vec{\beta}$ must be parallel to the interface, and its magnitude is given by:

$$|\vec{\beta}| = k_{iz}$$

The propagation vector $\vec{\beta}$ is therefore defined by the revised dashed-line construction in Figure 7.2-3.
Chapter Seven

Figures 7.2-1 and 7.2-3 are both illustrations of the general principle of phase-matching, which requires that the components of all propagation vectors parallel to an interface must match at that interface. We shall now see how this can be used to analyse more complicated device geometries.

7.3 The FTIR Beamsplitter

We begin with the cube beamsplitter, which is also often known as the frustrated total-internal-reflection (or FTIR) beamsplitter. This normally consists of two symmetric 90° glass prisms, separated by a small, accurately-defined air gap. Figure 7.3-1 shows the device in use. Here, a beam input through the left-hand face of the upper prism is divided into two components: a reflected wave, emerging from the right-hand face of the upper prism, and a transmitted wave from the corresponding face of the lower prism. It turns out (as we shall see later) that the power splitting ratio between the two beams can be set by varying the gap width. Why should this behaviour occur?

Well, the central region of the device can be considered as a stacked structure, consisting of three separate layers. The two outer layers correspond to the prisms, and have refractive index \( n_1 \), while the inner layer (the gap) has refractive index \( n_2 \). Since the device is a multilayer, it would seem reasonable that its operation can be explained in terms of the phase-matching principle. In fact, Figure 7.3-2 shows the required phase-matching diagram, and we will now explain its construction.

As before, we start by considering the fate of the incident wave. If we take \( n_1 = 1.5 \) and \( n_2 = 1 \) (i.e., glass prisms, separated by an air gap) then the critical angle at the upper prism/gap interface is \( \theta_c = \sin^{-1}(1/1.5) = 41.8° \). If the input beam travels at the typical angle of \( \theta_1 = 45° \), it will suffer total internal reflection. The upper part of the phase-matching diagram is therefore constructed exactly as for Figure 7.2-3; the incident and reflected wave-vectors are first drawn in, and the dashed line construction is used to show that an evanescent wave exists beneath the surface of the first prism.

To construct the lower part of the diagram, we must consider the field in the gap more carefully. If the gap is large, it is reasonable to expect that the evanescent field below the upper prism will decay quickly enough for it to be unaffected by the presence of the lower interface. Typically, this occurs if the gap is wider than a few \( \mu \text{m} \). In this case, the physical situation is exactly as for simple total internal reflection, and no power crosses either interface. There is then no transmitted beam.
If the gap is sufficiently small, however, a new phenomenon occurs. Remember that the field solutions for evanescent waves may be of two types, which either decay or grow away from an interface. Normal evanescent waves consist entirely of the former type, and we have discarded the latter as ‘physically unrealistic’ in the past, because they imply infinite field amplitudes at infinite distances from the interface. It turns out, however, that in order to satisfy the boundary conditions at the two interfaces here (which are only a finite distance apart), the field in the gap region must be constructed as a linear sum of the two solutions. It then consists of a rather more complicated boundary wave, which can have significant amplitude at the lower interface.

![Figure 7.3-2 Phase-matching in the FTIR beamsplitter.](image)

The waves generated by this hybrid field can be found by phase matching at both interfaces. We have already covered the procedure for the upper one. At the lower one, there is only one possibility - a propagating transmitted beam, travelling downwards at the angle $\theta_1$. This is found by a similar construction, matching the $z$-component of the propagation vector of the wave in Layer 3 with that of the boundary wave at the lower interface. The amplitude of the transmitted wave must clearly tend to unity as the gap tends to zero, because both interfaces will then disappear. In fact, it can be set to any desired value between unity and zero by choosing the gap correctly. Since the device is lossless, the power in the reflected beam must decrease as the transmitted beam increases. Generally, the gap is adjusted to equalised the two, whereupon the device acts as a 50 : 50 beamsplitter.

Rather surprisingly, therefore, power can be transferred across a gap region by a non-propagating, evanescent field, provided the gap is small enough. The process is often known as **optical tunnelling**. It explains the name of the FTIR splitter - the tunnelling effectively frustrates the total internal reflection that would normally occur at the upper interface. The gap required for significant tunnelling depends primarily on the decay rate of the evanescent fields. If $\theta_1$ is much larger than the critical angle, the fields decay very rapidly, and a very
small gap is needed. As $\theta_1$ approaches $\theta_c$, the fields decay more slowly, and the gap may be increased. We will now show how optical tunnelling can be exploited for the excitation of the guided modes of a planar waveguide.

### 7.4 THE PRISM COUPLER

Planar guided modes can be excited by an external free-space beam using a **prism coupler**, shown in Figure 7.4-1. This consists of a prism of suitable refractive index, which is separated from the waveguide by a small gap. Typically, the prism index should be somewhat higher than that of the substrate, and the gap should be of the order of 1 $\mu$m. This can be achieved by clamping the prism and guide together - dust particles then serve as spacers.

![Figure 7.4-1 The prism coupler.](image)

The input beam is again assumed to be a plane wave, which enters the prism from free-space at an angle $\theta_1$. If $\theta_1$ is large enough, this beam is total-externally-reflected at the prism/air-gap interface. Once again, therefore, the field beneath the prism is an exponentially-decaying evanescent wave, travelling parallel to the interface. The propagation constant of this field is given by:

$$\beta = k_1 \sin(\theta_1)$$

7.4-1

This field can be used to excite a guided mode in the waveguide underneath, if $\beta$ is matched to the propagation constant of the mode, as we now demonstrate.

The basic mechanism is illustrated by the phase-matching diagram shown in Figure 7.4-2. Here, the upper part of the diagram shows the excitation of a boundary wave beneath the prism. If the gap is very large, nothing further occurs. However, if it is small enough, the field solution in the gap must be modified as described for the FTIR splitter. This implies that there is again the possibility of power transfer across the two interfaces by optical tunnelling. If conditions are correct, this can give rise to a travelling wave in the waveguide layer. The propagation vector $k_g$ of this wave is found using a similar construction - the intersection of the dashed line with a circle of radius proportional to $n_3$ gives $k_g$.

By choosing the angle of incidence correctly, the angle $\theta_2$ can be made to correspond exactly with the direction of a downward-travelling, propagating wave in a zig-zag ray model of a guided mode. This wave will therefore be total-externally-reflected at the lower interface, since there is no intersection of the dashed line with the lowest circle (which has radius
proportional to $n_4$). Naturally, such a reflection must give rise to an upward-travelling wave in the guide layer (with propagation vector $k_g'$), and an evanescent boundary wave in the substrate.

![Figure 7.4-2 Phase-matching in the prism coupler.](image)

We have now identified all the waves in Figure 7.4-2. Note, however, that the conditions for the constructions above to be valid are quite critical. The prism index must be higher than that of the gap material, so that total internal reflection does indeed occur at the upper interface. However, it must also be higher than that of the guide itself, so that the ray angles $\theta_2$ and $\theta'_2$ (which are close to 90° in a weak-index guide) are obtained at a more convenient external angle $\theta_1$ (say, $30^\circ < \theta_1 < 60^\circ$).

Figure 7.4-3 shows how the whole of the transverse field pattern might look at this point. Near the waveguide, the pattern looks very similar to that of an ordinary guided mode. In the substrate, it must be evanescent, while in the guide layer itself it must be a standing-wave pattern, composed of the sum of two zig-zag waves.

Certainly, there are differences in the gap region, where the field is a hybrid boundary wave, and in the prism itself, where it is again a standing wave. However, if the gap is not too close
to zero, it is not inconceivable that the 'desired' portion of the field could be made to detach itself from the rest and propagate forward as a guided mode.

Figure 7.4-3 The transverse field distribution in the prism coupler.

We can achieve this by removing the external excitation at a suitable value of $z$. Could we then be said to have launched a guided mode? The short answer is yes, but there are several provisos.

Firstly, to make a useful input coupler, the prism should be relatively small compared with the substrate length, so that the guided beam propagates out from under the prism to other devices. In this case, the input beam can no longer be a plane wave, but must be a bounded beam. The amplitude of the guided mode must therefore build up from zero at the left-hand end of the coupler, to some value at the right. Secondly, the power transfer process is bi-directional - a prism coupler can just as easily be used in reverse, to couple a guided mode out into free space. If we are not careful, therefore, the guided mode will start coupling back out of the coupler at its right-hand end.

Fortunately, it has been found that a halved prism coupler can convert a bounded beam into a guided mode with good efficiency. This provides a sharp termination to the coupling region, at a point when the guided mode has built up to a reasonable amplitude (Figure 7.4-4). If the beam reflected from the prism is now measured, it will be found to contain rather less than 100% of the input power. This is because a substantial fraction of the power has now been coupled into the guided mode, and the required energy must clearly be provided by the input beam.
So far, our discussion has concentrated on the situation that occurs when the input beam is incident at the correct angle $\theta_1 = \theta_p$, when it is phase-matched to the guided mode. What happens at other angles? If $\theta_1$ is substantially different from $\theta_p$, it turns out that the field amplitude inside the guide is drastically reduced from that shown in Figure 7.4-3. This implies that the guided mode is not excited nearly so well, and the prism reflectivity returns to close to 100%. For incidence angles near to $\theta_p$, the reflectivity falls, following the typical resonance curve shown in Figure 7.4-5.

Both the width and the depth of the resonance depend on several factors - the precise variation of the air-gap, the length of the prism, the shape of the input beam, and absorption losses within the system. However, measurement of the reflectivity minimum can be used to determine the phase-matching angle $\theta_p$, and hence the propagation constant $\beta$ of the guided mode (from Equation 7.4-1). Note that angles are generally measured external to the prism, and therefore correction must be made for refraction at its input face. Consequently, the prism angles and refractive index must be known accurately.

A prism coupler can also provide other information. For example, the number of minima in the reflectivity curve yields the number of modes supported by the guide. Knowledge of the
effective indices of all the modes may even be used to reconstruct the refractive index profile
of the waveguide, using an inverse technique beyond the scope of this book. Waveguide
propagation loss can also be established, using two prisms. One acts as an input coupler,
while the other (a distance \( z \) away from the first) is an output coupler. The power transferred
through both prisms via the waveguide is then measured for different values of \( z \). Any
variation must clearly be due to propagation loss, since this depends on \( z \). The prism
coupler is therefore an extremely versatile characterisation tool.

**DESIGN EXAMPLE**

It is desired to characterise a weak planar guide (\( \Delta n \approx 0.01 \)) formed on a substrate of
refractive index 1.5. How should the coupling prism be chosen? Well, the effective index of
the guide must lie in the range 1.5 < \( n_{\text{eff}} < 1.51 \). The phase matching condition (Equation
7.4-1) can be written in the form:

\[
 n_{\text{eff}} = n_1 \sin(\theta_1) 
\]

For \( \theta_1 \approx 60^o \), we obtain \( n_1 \approx 1.7 \). A prism of high-index glass would therefore prove
suitable. It would be convenient to make this as a right-angle prism, with hypotenuse angles
of 60° and 30°.

**7.5 PHASE MATCHING FOR GUIDED MODES**

Until now, we have only considered one-dimensional propagation of a mode, down the z-
axis. However, a planar guided mode can travel in any direction in the y - z plane. To
proceed further, we need a more general description. This is provided by a vectorial notation
similar to that introduced at the beginning of this Chapter. Assuming that the mode has a
transverse field distribution \( E(x) \) and propagation constant \( \beta \), a suitable scalar solution
might have the form:

\[
 E(x, y, z) = E(x) \exp(-j\beta \cdot r) \]

where:

\[
 |\beta| = \frac{2\pi n_{\text{eff}}}{\lambda_0} 
\]

Equation 7.5-1 still describes a mode with an x-dependent transverse field. However, its
phase variation is defined by the vector \( \beta \), which is assumed to lie in the y - z plane. The
orientation of \( \beta \) gives the wave direction, while its magnitude describes the effective index of
the mode. The locus of all possible propagation vectors in the y - z plane is, once again, a
circle, this time of radius \( \beta \). However, if the guide supports more than one bound mode (say,
n of them), the loci of all possible modes of guided propagation are a set of concentric
circles, of radii \( \beta_1, \beta_2, ..., \beta_n \). Figure 7.5-1 shows the circles for a three-moded guide, drawn
with radii proportional to \( n_{\text{eff}1}, n_{\text{eff}2} \) and \( n_{\text{eff}3} \).

Just as in bulk optics, a guided mode can be reflected and refracted at a discontinuity.
However, instead of the discontinuity being between two homogeneous media, it will
generally be between two different types of waveguide. As discussed in Chapter 6, these
could have different guide layer thicknesses, or the same thickness but different refractive
indices, or different overlays, or be inclined at an angle to each other. Consequently, the
propagation vectors in each guide may have different magnitudes, or lie in different planes.
Solution of the general boundary-matching problem at an interface is very complicated. To find a full solution, it is often necessary to assume the presence of both the full set of bound and radiation modes on either side of the interface, including both polarizations. This implies that mode conversion, polarization conversion and scattering into radiation modes can all occur at interfaces. The possible propagation directions are then governed by the phase matching principle.

Here, we will consider a simplified example involving a coplanar discontinuity in a three-moded guide. The interface is parallel to the y-axis, and we will restrict our attention to the bound modes. Figure 7.5-2 shows the phase-matching diagram, for the case when the lowest-order mode (which has effective index $n_{\text{eff}1}$) is incident obliquely from the left.

The diagram is constructed by first drawing circles of radii proportional to $n_{\text{eff}1}$, $n_{\text{eff}2}$ and $n_{\text{eff}3}$ (using the relevant values on either side of the discontinuity). The propagation vector of the
input wave (labelled I) is then drawn in at the correct angle, and with length proportional to \( n_{\text{eff}} \). The component of this mode that is reflected must travel at an equal and opposite angle to the interface normal, allowing the vector \( R \) to be added. All other possible propagation vectors can then be found from the intersection of the dashed line - the normal to the interface passing through the end of \( R \) - with the two sets of circles. In this example, there are six possible vectors, corresponding to three transmitted waves (collectively labelled T) and three reflected waves (labelled R). The input beam therefore suffers reflection, refraction and mode conversion.

If the discontinuity is small, however, most of these effects can be neglected. From Chapter 6, we know that little mode conversion occurs at weak interfaces. The most important of the waves described above will therefore be the reflected component \( R \) and the transmitted component \( T \) that are of the same order as the input mode. A reasonable approximation to Figure 7.5-2 is therefore provided by the reduced diagram of Figure 7.5-3, which retains only the waves I, \( R \) and \( T \).

Figure 7.5-3  Approximate phase matching diagram for the interface between two guides.

Furthermore, unless the angle of incidence is very oblique, we might expect the amplitude to \( T \) to be much larger than that of \( R \), so that the main effect occurring is one of refraction of the input mode. We might guess that this will be governed by a modified Snell’s law equation, of the form:

\[
 n_{\text{eff}_1} \sin(\theta_1) = n_{\text{eff}_2} \sin(\theta_2) 
\]

where \( n_{\text{eff}} \) is the effective index of the lowest-order mode in waveguide 1, and \( \theta_1 \) is its angle of travel, and \( n_{\text{eff}} \) and \( \theta_2 \) are the corresponding values in waveguide 2. To illustrate this, Figure 7.5-4 shows the refraction of a real guided beam at an interface between two guides of different thickness.

To summarise, we may assume that if the interface is sufficiently weak, the input field will travel across it without losing too much power. Its transverse field distribution will not alter much, but its direction of propagation will change through refraction. This means that refractive optical components may operate by slight local modification of an otherwise uniform planar guide.
7.6 REFRACTIVE OPTICAL COMPONENTS

Refractive guided wave components can be subdivided into a number of classes. The simplest work through the sudden change in direction that takes place at an abrupt discontinuity, as described above. This allows guided wave lenses to be constructed; these are analogous to bulk optical lenses (except that in general they may only convert one cylindrical guided beam into another). For example, a simple plano-convex lens might be made by the deposition of high-index overlay material in a region bounded by a parabola and a straight line, as shown in Figure 7.6-1.

![Diagram of a plano-convex overlay lens]

**DESIGN EXAMPLE**

We will now try to estimate the factors governing the performance of overlay lenses. Previously, we obtained the following formula for the focal length of a bulk-optic lens, made of material of refractive index $n_L$ and with two spherical surface of radii $r_1$ and $r_2$:
In this case, the surrounding medium is air (which has unity refractive index). More generally, we might expect that for a surround of index \( n_M \), the formula above modifies to:

\[
1/f = (n_L / n_M - 1) \left( 1/r_1 + 1/r_2 \right)
\]  

7.6-2

Now, if a guided wave lens also works by refraction, we would expect its operation to be controlled by the values of the effective index inside and outside the lens. If these are \( n_{\text{eff}L} \) and \( n_{\text{eff}M} \), the corresponding formula is:

\[
1/f = (n_{\text{eff}L} / n_{\text{eff}M} - 1) \left( 1/r_1 + 1/r_2 \right)
\]  

7.6-3

where \( r_1 \) and \( r_2 \) are now the radii of curvature of the two circular boundaries of the high-index overlay.

We can now directly compare the lens shapes required to perform the same task in bulk and guided wave optics. Suppose we wish to design a plano-convex lens, with a given focal length \( f \). In this case, we may take \( r_1 \) to be infinite, so that \( r_2 \) is given by:

\[
r_2 = (n_L / n_M - 1) f
\]  

(in bulk optics)

\[
r_2 = (n_{\text{eff}L} / n_{\text{eff}M} - 1) f
\]  

(in guided wave optics)

7.6-4

Typically, \( n_L / n_M - 1 \) will be around 1.5 - 1 = 0.5 in the bulk device. What is the corresponding value of \( n_{\text{eff}L} / n_{\text{eff}M} - 1 \)? With a substrate of refractive index \( n_s = 1.5 \), we might expect to be able to form guides with effective indices \( n_{\text{eff}M} \approx 1.6 \) using a high-index guiding layer. Now, the discontinuities forming the lens must be weak, to prevent scattering to radiation modes. This implies that \( n_{\text{eff}L} - n_{\text{eff}M} \) must be much smaller than \( n_{\text{eff}M} - n_s \). An upper limit might be \( n_{\text{eff}L} - n_{\text{eff}M} = 0.02 \). We then find that \( n_{\text{eff}L} / n_{\text{eff}M} - 1 \) is 0.02/1.6 = 0.0125. Substituting these values into Equation 7.6-4, we conclude that the guided wave lens has a radius of curvature 0.5/0.0125 = 40 times smaller than the bulk device.

If we now take the focal length of the guided wave lens to be 5 cm (the maximum that will fit on a chip of reasonable dimensions), we find that \( r_2 = 0.63 \) mm. This implies that the largest possible aperture of the lens (2\( r_2 \)) is only 1.26 mm, and to achieve this the lens must differ in shape quite considerably from the ideal thin-lens assumed in the calculation - in fact, it must be semicircular. Consequently, the lens will have a very low numerical aperture, and will also produce a very aberrated focus.

**FRESNEL, LUNEBERG AND GEODESIC LENSES**

There are several solutions to this problem. Remember that the transfer function of a lens is given by:

\[
\tau_L = \tau_s \exp(jk_0 R^2/2f)
\]  

7.6-5

Where \( \tau_s \) is a constant, and \( R \) is the distance from the axis. The focusing property of the lens can thus be ascribed to the parabolic variation of phase-shift it imparts. In guided wave optics, \( \tau_L \) must have a similar form. Now, any phase shift of 2\( \pi \) radians (where \( \nu \) is an integer) can be disregarded, since all these are equivalent to no shift at all. We can therefore safely remove any parts of the overlay that simply contribute shifts of 2\( \nu \pi \), and still obtain a parabolic phase variation. The modified overlay pattern shown in Figure 7.6-2, where each
step in thickness is a step of $2\pi$ in phase, is therefore equivalent to a plano-convex lens; however, it is much thinner. This is known as a **Fresnel lens**.

Though the Fresnel lens is useful, it does not provide an answer to the underlying problem: the weak change in effective index possible at an abrupt interface leads to a requirement for small radii of curvature of the lens boundaries. The answer is to use larger changes in effective index, which are applied gradually to prevent mode conversion. We cannot analyse such lenses here, but merely state that there are two common types. Both are circularly symmetric, and can therefore nominally focus a plane guided wave with aberrations that are independent of the angle of incidence of the beam. In fact, they are two-dimensional forms of the **Maxwell fish-eye** or **Luneburg** lens in bulk optics.

![Figure 7.6-2 Fresnel overlay lens.](image)

The first is the **overlay Luneberg lens**, shown in Figure 7.6-3. As before, the lens is formed from a high-index overlay, but this time with a graded, circularly-symmetric thickness profile. For some particular overlay profiles, focussing is achieved through continuous refraction of the input wave, and the ray paths inside the lens are therefore curved. The desired overlay profile may be approximated by sputtering material through a circular aperture.

![Figure 7-6-3 Overlay Luneberg lens.](image)

An equivalent lens can also be formed without using an overlay, as follows. Before fabrication of the guide, a circularly-symmetric depression (generally spherical) is precision-turned in the substrate using a diamond lathe. Usually, the edges of the depression are also rounded, to provide a gradual input transition that reduces scattering. The guide is then fabricated, with uniform thickness, to follow the topology of the depression. The slow change in local orientation of the guide then causes continuous
refraction, so that rays follow a geodesic path (the shortest path, in accordance with Fermat’s principle) through the lens, which is consequently known as a **Geodesic lens**.

Figure 7.6-4 shows a geodesic lens. Note that the optical path through the centre of the lens is larger than that at the edges, because of the increased distance that must be travelled in crossing the deepest parts of the depression. This retards the centre of the wavefront, so that a converging output is obtained.

![Geodesic lens diagram](image)

**Figure 7.6-4** Geodesic lens.

Figure 7.6-5 shows ray paths through a real geodesic lens, which are made visible through the use of a fluorescing dye. Clearly, it is impossible to manufacture a diverging lens using the geodesic technique. However, we note that two waveguide lenses, with differing focal lengths, can be used as a telescope beam expander (previously described in Chapter 4). This arrangement is often used to expand a beam coupled directly into the guide from a laser to a more useful cross-section.

![Ray paths image](image)

**Figure 7.6-5** Ray paths in a geodesic lens (photo courtesy Y. Okamura, Osaka University).

### 7.7 GRATINGS

A further important class of components operate by diffraction rather than refraction. These are periodic structures, known as **gratings**. In many ways, they are ideal guided wave components, and they may be analysed using an extension of the phase matching principle. We will show how this is done first in bulk optics, and then in guided wave optics.
OPTICALLY THIN GRATINGS

To investigate the operation of a grating, we need a model. Figure 7.7-1 shows the simplest possible representation of a periodic structure, which is made from an infinite array of point scatterers spaced at equal distances $\Lambda_y$ along the y-axis. We can work out what happens when a light-wave strikes it using qualitative arguments. The wave is infinite, plane, and of free-space wavelength $\lambda_0$; it travels at an angle $\theta_0$ to the z-axis, and the refractive index everywhere is $n$, so the effective wavelength is $\lambda_0/n$.

Allowing each point to scatter isotropically and independently, we might initially expect light to be transmitted through the grating in all directions - in other words, as a spectrum of all possible plane waves. However, this argument completely ignores the periodicity of the structure. Further thought reveals that a plane wave output is actually possible only when the scattered components from any two adjacent points (e.g. A and D) add up exactly in-phase. If this is not the case, the net contribution from all the scatterers will average to zero in the far field of the grating (see, for example, Question 4.6).

![Figure 7.7-1](image)

This type of constructive interference can only occur when the difference between the path lengths AB and CD in Figure 7.7-1 is a whole number of wavelengths. We can calculate the relevant path difference for an output angle $\theta_L$ as:

$$AB - CD = \Lambda_y \left[ \sin(\theta_L) - \sin(\theta_0) \right]$$

For constructive interference, we require $AB - CD = L \lambda_0 / n$, where $L$ is an integer, so that:

$$\sin(\theta_L) = \sin(\theta_0) + L \lambda_0 / n \Lambda_y$$

Equation 7.7-2 implies that constructive interference only occurs at a set of discrete angles $\theta_L$. The action of the grating is thus to split the input beam up into a number of plane waves, travelling in different directions. These are known as diffraction orders; associated with each is an index $L$, and the solution of Equation 7.7-2 gives the direction of the $L^{th}$ order. Since the equation contains $\lambda_0$, the angles depend on the wavelength of the optical beam, so gratings are intrinsically dispersive and will produce a series of output spectra from a polychromatic input. Equation 7.7-2 also shows that the diffraction orders will be more
widely separated the closer the spacing of the scatterers, i.e. as $\Lambda_y$ decreases. To get a useful separation and dispersive power, $\Lambda_y$ should be of the same order as $\lambda_0$.

We cannot estimate the intensities of the individual diffraction orders from this argument, but we would expect to see many of them, of roughly equal intensity. This is often a nuisance, as we may be interested in just one order. Equally, we would not expect the intensities to depend much on the wavelength or the angle. This type of grating is therefore unselective, favouring no particular incidence condition or diffraction order. It is usually known as an \textit{optically thin} grating, and the multiwave diffraction regime is called the \textbf{Raman-Nath} regime, after its original investigators in the field of ultrasonics (Sir C.V. Raman and N.S.N. Nath).

Before proceeding further, we note that the results above may be interpreted geometrically, using the modified phase-matching diagram shown in Figure 7.7-2. This is rather different to the diagrams we have used so far, and deserves a reasonably thorough explanation. First, an important new parameter, known as the \textbf{grating vector}, is defined. For historical reasons, this is labelled $K$. It is oriented in the direction of the periodicity of the grating (i.e. in the $y$-direction in this case), and has modulus:

\begin{equation}
|K| = \frac{2\pi}{\Lambda_y} \tag{7.7-3}
\end{equation}

Next, the usual circular locus for the propagation vectors is drawn, with radius proportional to $k = n\lambda_0$, and the propagation vector of the input wave is added. To avoid confusion with the grating vector, this is labelled $\rho_0$. The propagation vector $\rho_L$ of the $L^{th}$ diffraction order is then found by first adding $L$ times the grating vector to $\rho_0$, and then finding the intersection with the circular locus of the dashed line, drawn normal to the grating boundary through the tip of this new vector.

![Figure 7.7-2 Phase matching diagram for diffraction orders in a grating.](image)

Simple geometry then shows that:

\begin{equation}
\rho_{Ly} = \rho_{0y} + L|K| \tag{7.7-4}
\end{equation}

so that:

\begin{equation}
nk_0 \sin(\theta_L) = nk_0 \sin(\theta_0) + L|K| \tag{7.7-5}
\end{equation}
or:

\[(2\pi n/\lambda_0) \sin(\theta_L) = (2\pi n/\lambda_0) \sin(\theta_0) + 2\pi L/\Lambda_y\]  

Equation 7.7-6

Apart from a slight rearrangement, it should be clear that Equation 7.7-6 is identical to 7.7-2. The modified phase-matching diagram, which is known as a **K-vector diagram**, is therefore a useful way to predict the directions of diffraction orders geometrically. Often, the dashed-line construction is omitted for clarity, being implicitly understood, and the wave vectors are taken simply as:

\[\rho_L' = \rho_0 + LK\]  

Equation 7.7-7

This relation is termed **K-vector closure**. Figure 7.7-3 shows how this modified construction can be used to predict the wave directions for an optically thin grating. In this case, \(\Lambda_y\) is large (so \(|K|\) is small) and the output consists of many diffraction orders, all travelling roughly parallel to the z-axis. The correction in wave direction that results from completing the construction is then only a small one.

![Figure 7.7-3 Reduced phase matching diagram for higher diffraction orders in the Raman-Nath diffraction regime.](image)

**OPTICALLY THICK GRATINGS**

To remove the unwanted diffraction orders, we need to introduce some kind of inherent selectivity into the grating. Figure 7.7-4 shows a suitable structure (known as an **optically thick grating**) which now has a finite thickness, and consists of fringes or extended scatterers instead of points. These are slanted at an angle \(\phi\) to the z-axis and are spaced at distances \(\Lambda\) apart, but the y-component of their spacing is still \(\Lambda_y = \Lambda / \cos(\phi)\).

Again we ask what happens when the grating is illuminated, by the same wave as before. Equation 7.7-2 must still be valid, because the y-period of the grating is the same. However, the scattering is now distributed in the z-direction. This constrains the allowed diffraction angles even more. Ignoring multiple scattering, we argue that constructive interference only occurs when all scattered components add in-phase. This must be true not only for components scattered by different fringes, but also for contributions from all points along the same fringe. Path lengths like EF and HG in Figure 7.7-4 must therefore be equal. This restricts the diffraction angles additionally to:

\[\theta_L = \theta_0 \quad \text{or} \quad \theta_L = 2\phi - \theta_0\]  

Equation 7.7-8
The fringes then act as partial mirrors, allowing only transmission or reflection.

![Figure 7.7-4 An optically thick grating.](image)

Now, if $\theta_L = \theta_0$, the only solution to Equation 7.7-2 is $L = 0$, so the only allowed wave is the 0th diffraction order (the input wave). However, if $\theta_L = 2\phi - \theta_0$, we get:

$$\sin(2\phi - \theta_0) - \sin(\theta_0) = 2 \sin(\phi - \theta_0) \cos(\phi)$$

$$= L\lambda_0 / n\Lambda$$

Or:

$$2\Lambda \sin(\phi - \theta_0) = L\lambda_0 / n$$

7.7-9

This implies that other diffraction orders can exist, but only at certain specific angles. The -1th order, for example, is allowed at the first Bragg angle, which is defined by Bragg's law:

$$2\Lambda \sin(\theta_0 - \phi) = \lambda_0 / n$$

7.7-10

Other waves are allowed at higher Bragg angles, but generally the first diffraction order is the most useful. Combining the two conditions above, we can show that up to two waves can exist at once. For incidence exactly at the Bragg angle, the two permitted waves are the 0th (the input) and -1th orders. Away from this angle, the input wave travels through the grating unaltered. This regime of diffraction is known as the two-wave or Bragg regime, and the modified structure has therefore introduced the desired selectivity into the diffraction process.

In practice, the size of the grating will be limited and conditions 7.7-2 and 7.7-11 relax accordingly. Neither the grating nor the input wave can be infinite, so the wave directions will be less well defined. This affects the resolution of the grating, but provided it is many optical wavelengths wide our arguments should be valid. Finite thickness is more important. Depending on its parameters, a grating may either act like the optically thin grating in Figure 7.7-1, and produce many diffraction orders, or more like the optically thick grating in Figure 7.7-4 and give only two. A reasonable guide is that two-wave behaviour occurs whenever the input wave has to cross many fringe planes before emerging from the grating.
DESIGN EXAMPLE

We can work out the value of $\Lambda$ needed to satisfy Bragg's Law (Equation 7.7-11) for some typical values. Assuming that $\lambda_0 = 0.5145 \, \mu m$ (a green argon-ion laser line) and $n = 1.6$, that the grating is unslanted (so $\phi = 0$) and that the Bragg angle is $\theta_0 = 45^o$, we get $\Lambda \approx 0.23 \, \mu m$. This is an extremely small figure, which implies a grating with a large deflection angle will be hard to make.

THE BANDWIDTH OF OPTICALLY THICK GRATINGS

Even in the two-wave regime, restricted thickness will result in a finite angular and wavelength range over which significant diffraction occurs. The bandwidth can be estimated by assuming that the efficiency of the first diffraction order ($L = -1$) will be zero when there is a whole wavelength difference between contributions scattered from either end of a fringe. We can work out the change in optical wavelength needed for this to occur as follows, assuming an unslanted grating of thickness $d$. The path difference between the two components is:

$$\Phi = d \{ \cos(\theta_0) - \cos(\theta_1) \}$$  \hspace{1cm} 7.7-12

When the Bragg condition is satisfied, we must have $\theta_1 = -\theta_0$, so the path difference is zero. Now, the change in path difference $\Delta \Phi$ accompanying a change in wavelength $\Delta \lambda_0$ is given by:

$$\Delta \Phi = (d \Phi/d\lambda_0) \Delta \lambda_0$$  \hspace{1cm} 7.7-13

We may evaluate Equation 7.7-13 by first expanding it using the chain rule for differentiation, as:

$$\Delta \Phi = (d \Phi/d\theta_1) (d \theta_1/d\lambda_0) \Delta \lambda_0$$  \hspace{1cm} 7.7-14

d$\Phi/d\theta_1$ and $d \theta_1/d\lambda_0$ may then be found by differentiating Equations 7.7-12 and 7.7-2 in turn, to get:

$$d \Phi/d\theta_1 = d \sin(\theta_1) \quad \text{and} \quad d \theta_1/d\lambda_0 = -1 / \{ n \Lambda \cos(\theta_1) \}$$  \hspace{1cm} 7.7-15

Substituting these expressions into Equation 7.7-14, and putting $\theta_1 = -\theta_0$, we then find that:

$$\Delta \Phi = d \Delta \lambda_0 \sin(\theta_0) / \{ n \Lambda \cos(\theta_0) \}$$  \hspace{1cm} 7.7-16

Now, according to our criterion, the efficiency is zero when $\Delta \Phi = \lambda_0/n$, so that:

$$\Delta \lambda_0/\lambda_0 = (\Lambda/d) \cot(\theta_0)$$  \hspace{1cm} 7.7-17

This important result implies that the bandwidth of an optically thick grating is inversely proportional to its thickness, so that the thicker the grating, the smaller the range of wavelengths over which it will diffract light efficiently.

DESIGN EXAMPLE

For the example used before, namely $\lambda_0 = 0.5145 \, \mu m$, $\theta_0 = 45^o$ and $\Lambda = 0.23 \, \mu m$, and a moderate thickness of $d = 10 \, \mu m$, we get $\Delta \lambda \approx 12 \, \text{nm}$. For a thicker grating, with $d = 1 \, \text{mm},$

R.R.A.Syms and J.R.Cozens  Optical Guided Waves and Devices 21
this reduces to $\Delta \lambda = 12 \text{ Å}$. An optically thick grating can therefore act as an extremely narrow-band wavelength filter, and it is equally simple to show that its angular selectivity must be correspondingly high.

**OPTICALLY THICK PHASE GRATINGS**

In addition to high selectivity, we will generally also require high diffraction efficiency. It turns out that an optically thick grating can be very efficient indeed, provided the structure is lossless, and the fringe planes are formed by periodic modulation of the refractive index of the medium. The desired structure is therefore an optically thick phase grating. We cannot calculate the exact response characteristic for such a grating using the information above (though we will see how this is done for one particular geometry in Chapter 10). However, using coupled wave theory, it can be shown that the response characteristic is as in Figure 7.7-5.

![Figure 7.7-5 Filter response of a high efficiency, optically thick grating.](image)

This figure shows how the transmitted and diffracted beam intensities vary, as either the angle of incidence or the wavelength of the input beam is changed. At Bragg incidence, when $\theta_0 = \theta_B$ (or $\lambda_0 = \lambda_B$), the diffracted beam is excited efficiently. There is then a strong depletion of the transmitted beam, which may reach 100% if the grating is correctly designed (as in the example above). As $\theta_0$ (or $\lambda_0$) is changed, the diffracted beam intensity falls. The transmitted intensity shows a corresponding rise, and both curves have a typical ‘filter’ characteristic. It turns out that the expression for the bandwidth calculated above gives surprisingly good agreement with the predictions of coupled wave theory.

**K-VECTOR CLOSURE IN AN OPTICALLY THICK GRATING**

The behaviour of a slanted, optically thick grating can also be predicted, using a modified K-vector closure diagram. To cope with slanted fringes, we simply introduce a more general grating vector $\mathbf{K}$. This is oriented normal to the plane of the fringes, and its magnitude is:

$$|\mathbf{K}| = 2\pi/\Lambda$$

The diagram is then constructed as before. For example, Figure 7.7-6 shows the diagram for a slanted grating, replayed at the Bragg angle. The input wave vector $\mathbf{p}_0$ is first drawn in. The corresponding vectors $\mathbf{p}_1'$ for the diffraction orders are then computed from Equation 7.7-7. When this is done, it is found that the end of one such vector, $\mathbf{p}_1'$, lies exactly on the circle, so that $\mathbf{p}_1' = \mathbf{p}_1$. This particular circumstance only arises exactly at Bragg incidence, and the $-1^{\text{st}}$ order is then said to be phase matched by the grating. This order can then be
excited efficiently, because all the components scattered in the direction concerned will add in-phase. The ends of other propagation vectors (e.g. $\mathbf{p}_{-2}$) lie much further away from the circle. This suggests that the $-2^{th}$ order is far from being phase-matched, and will not be excited efficiently. For other diffraction orders (e.g. the $+1^{th}$), the dashed line construction provides no intersection with the circle at all. These waves cannot exist as propagating waves, and are therefore cut off. Thus, according to the diagram, only the two waves defined by the bold vectors are important, and we may reduce the K-vector construction accordingly. We will do so in the discussion below, which is devoted to the use of gratings in guided wave optics.

![Figure 7.7-6 Reduced phase matching diagram for a slanted, thick grating in the Bragg diffraction regime.](image)

**7.8 GRATINGS IN GUIDED WAVE OPTICS**

A wide variety of methods can be used to construct a phase grating in guided wave optics. All that is required is that the effective index be varied periodically. Figure 7.8-1 shows one example, where the surface of a planar guide has been corrugated, by etching through a patterned mask. The resulting changes in thickness of the guide layer then give rise to the necessary variation in effective index.

![Figure 7.8-1 A corrugated grating.](image)

**TRANSMISSION GRATINGS**

Figure 7.8-2 shows how such a corrugation might be used as a transmission grating. This is assumed to be optically thick, and therefore operating in the two-wave diffraction regime. Figure 7.8-2a shows a plan view of the waveguide; the input beam is incident from the left, and gives rise to two transmitted waves, the $0^{th}$ and $-1^{th}$ diffraction orders, on the right. Figure 7.8-2b shows the K-vector diagram, for incidence at the Bragg angle. Note that
only two propagation vectors are included in the construction, and that the grating vector (which is normal to the horizontal fringes in Figure 7.8-2a) is oriented vertically. This demonstrates an important point. A large angular deflection of the input beam has been achieved, using a relatively small perturbation to the effective index of the guide, because the device operates by diffraction rather than refraction. Since small index changes are the norm in guided wave optics, gratings are ideal waveguide components in this respect.

![Figure 7.8-2 Transmission grating: a) schematic and b) K-vector diagram.](image)

**GRATING LENSES**

In fact, diffractive devices can act as more effective waveguide lenses than the overlay components discussed earlier, because larger beam deflection allows a higher numerical aperture. Figure 7.8-3 shows one example, the **chirped grating lens**. Here a slow variation in the grating period produces a corresponding change in the deflection angle over the lens aperture, which results in an approximate focussing action. A linear chirp (i.e., a linear variation in period) is suitable for focussing an off-axis plane input beam, as shown. More generally, the grating fringes must be curved, and the local fringe orientation and spacing is found by local application of the Bragg condition.

![Figure 7-28 The chirped grating lens.](image)

Grating lenses suffer from two main disadvantages. Firstly, because of their inherent dispersion, the focal spot moves as the optical wavelength is altered, and can become highly aberrated. Secondly, their angular selectivity results in a greatly restricted field of view. However, these can be unimportant (for example, in a telescope beam expander, when a fixed angle and wavelength are used).
CHAPTER
SEVEN

REFLECTION GRATINGS

Figure 7.8-4a shows a plan view of an alternative structure. This is a reflection grating, formed using fringes that are now oriented parallel to the input boundary. The input wave is incident as before, but this time the two diffraction orders emerge from different sides of the grating. The 0th order is transmitted, but the -1th order is now a reflected wave. Figure 7.8-4b shows the K-vector diagram, again for incidence at the Bragg angle. Note that the grating vector is now oriented horizontally, and that its magnitude is rather larger than that in the transmission grating above. Reflection gratings therefore typically require much smaller grating periods.

![Reflection grating schematic and K-vector diagram](image)

Figure 7.8-4  Reflection grating: a) schematic, and b) K-vector diagram.

'TWO-DIMENSIONAL' GRATINGS

It is of course unnecessary to consider grating boundary shapes to be fixed in the form of the simple slab geometries considered so far. Figure 7.8-5 shows a grating whose fringes are oriented at 45o to the edges of a boundary, which has the form of a long, thin rectangle. Here, a relatively narrow input beam can be diffracted as a much wider output beam. The device therefore functions as a compact beam expander; which could replace an expansion telescope. However, the grating strength must be carefully tailored along its length, to ensure a uniform amplitude distribution in the output beam.

![Beam expander grating schematic and K-vector diagram](image)

Figure 7.8-5  Beam expander grating: a) schematic and b) K-vector diagram.
ELECTRO- AND ACOUSTO-OPTIC GRATINGS

Equally, it is not necessary to consider the grating to be a static entity. There are two common forms of electrically-switchable grating, which operate using combinations of the material effects discussed in Chapter 3. The first is the electro-optic grating, shown in Figure 7.8-6a. Here a periodic metal structure, known as an interdigital electrode, is placed on the surface of the guide that is assumed to have been fabricated in an electro-optic material (e.g., LiNbO$_3$ or LiTaO$_3$). With no voltage applied to the electrodes, there is no grating. However, when a static voltage is applied, the ’finger’ voltages alternate in sign. This results in a periodic variation in electric field beneath the electrodes (Figure 7.8-6b), which induces a corresponding variation in index in the guide through the electro-optic effect. The grating may therefore be switched on and off at will.

However, the beam deflection angle (which is determined by the electrode pitch) is fixed. Usually, it is rather small, since the lithographic process used to define the electrodes is limited to a linewidth of $\approx 0.5 - 1 \mu m$, which yields a period of $\Lambda \approx 2 - 4 \mu m$. Despite this, the electrodes may be sufficiently long that the device acts as a volume grating, so the diffraction efficiency can be high. 98% efficiency has been achieved with Ti-indiffused guides on LiTaO$_3$ substrates. The electro-optic grating may therefore act as an effective modulator. Its speed is limited to about 1 GHz by the capacitance of the electrodes.

The second is the acousto-optic grating, shown in Figure 7.8-7. This is a more versatile device, which can steer a beam as well as modulating it. Again, a periodic electrode structure is placed on the surface of the guide, which is assumed this time to have been fabricated in a piezo-electric, acousto-optic material. LiNbO$_3$ is again suitable, but non-piezo electric materials (e.g. Si) can also be used if a layer of piezo-electric material (typically ZnO) is deposited between the electrodes and the substrate.

This electrode structure, which is known as an interdigital transducer, does not itself act as the grating. However, when it is excited by an A.C. radio-frequency voltage, it creates a time-varying, spatially-periodic electric field in the material beneath, which in turn creates a similar pattern of stress through the piezo-electric effect. Now, this type of stress distribution effectively corresponds to a pattern of standing acoustic waves, which can of course be decomposed into the sum of two travelling waves. Consequently, the net result is...
the excitation of two acoustic waves, which emerge from beneath the transducer, travelling in opposite directions. Since the waves propagate near the surface of the material, they are known as **surface acoustic waves** (or SAWs). Typically, only one wave is required, and the other is removed using a surface absorber. The remaining wave creates a moving variation in effective index, through the acousto-optic effect, and it is this variation which acts as the grating. Once again, the grating may be switched on and off, so the device can act as a modulator. This time, speed is mainly limited by excessive acoustic propagation losses at frequencies above \( \approx 1 \text{ GHz} \).

![Figure 7.8-7 The acousto-optic grating.](image)

Most importantly, the grating period may be varied over a limited range (depending on the transducer bandwidth) by changing the RF frequency. This alters the diffraction angle, so acousto-optic gratings may also be used as beam deflectors. Diffraction may again take place in the Bragg regime at high frequencies, but the Raman-Nath regime is often used to minimise the efficiency variations that occur when the frequency is altered. One further important difference from the electro-optic grating is that the diffraction orders are all frequency-shifted (by integer multiples of the acoustic frequency) through the Doppler effect. The acousto-optic grating may therefore also be used as a frequency modulator.

**DESIGN EXAMPLE**

We may estimate the parameters of a guided-wave acousto-optic beam deflector formed on a lithium niobate substrate as follows. For LiNbO\(_3\), the acoustic velocity is \( 6.57 \times 10^3 \text{ m/s} \) and the refractive index is \( n_e = 2.2 \). Hence, at a typical operating frequency of 500 MHz, the acoustic wavelength is \( 6.57 \times 10^3 / 500 \times 10^6 = 1.31 \times 10^{-5} \text{ m} \), or 13.1 \( \mu \text{m} \). The period \( \Lambda \) of the interdigital transducer must have the same value, so assuming a mark-to-space ratio of unity, each finger electrode must be \( 13.1 / 4 = 3.275 \mu \text{m} \) wide, a feature size well within the capacity of photolithography. Assuming an optical wavelength of 0.633 \( \mu \text{m} \) (red light from a He-Ne laser), and normal incidence, the angle of deflection of the first diffraction order will be \( \theta_1 = \sin^{-1}(\lambda_0/n_\Lambda) = \sin^{-1}(0.633 / 2.2 \times 13.1) \approx 1.25^\circ \).

**MODE CONVERSION BY A GRATING**

Finally, gratings are not restricted to interactions involving the phase matching of two modes of the same type. Consider, for example, a single-moded guide. We have seen that this can
support two polarization modes; consequently, even in this case there are two possible loci for propagation vectors. Figure 7.8-8a shows the K-vector diagram illustrating the conversion of a TE-moded input into a TE-moded diffracted beam. Here, both $\rho_0$ and $\rho_{-1}$ lie on the full-line circle, the TE-mode locus. However, an equally valid possibility is the diffraction of a TM-moded input into a TM-moded output. This is shown by the alternative construction, where $\rho_0$ and $\rho_{-1}$ lie on the dashed-line TM-mode locus. Because of the difference in radii of the circles, the Bragg angles for TE-TE and TM-TM diffraction are slightly different. One further possibility exists: a TE-moded input may be diffracted as a TM-moded output. The phase-matching condition for this case is shown in Figure 7.8-8b; once again, the Bragg angle involved is different. In this case, mode-conversion accompanies diffraction. In designing a grating device, therefore, care is needed to ensure that only the desired interaction occurs.

Figure 7.8-8  K-vector diagrams for a) TE-TE and TM-TM diffraction, and b) TE-TM mode conversion.

GRATING COUPLERS

A grating may even be used to phase-match guided and radiation modes, forming an input/output coupler similar to the prism coupler discussed earlier. This device is known as the grating coupler, and its principle of operation is shown in Figure 7.8-9.

Figure 7.8-9  The grating coupler.
Here a plane input beam is incident from free space on a corrugated grating, which is fabricated on the surface of a planar waveguide. Figure 7.8-10 shows the corresponding phase-matching diagram. If the angle of incidence $\theta_1$ is correct, the z-component of the first diffraction order of the grating will equal the z-component of the propagation vector of the guided mode. This occurs when:

$$k_{iz} + |K| = |\beta|$$  \hspace{1cm} (7.8-1)

If this is the case, power will be transferred to the guided beam, much as in the prism coupler.

Although prism and grating couplers are functionally similar, the latter have several important advantages. Firstly, they are flat, rugged, and fully integrated with the waveguide. Secondly, it is possible to design grating couplers with a varying periodicity, which can perform multiple functions - for example, they can couple a guided beam into free space, and focus it at the same time. Figure 7.8-11 shows the **chirped grating coupler**, which contains a grating with linearly varying period.

A guided mode incident on such a grating will be diffracted into a free-space beam, whose direction of propagation depends on the local grating period. In this case, the output will be
a cylindrical wave converging on a line focus above the guide. However, by simultaneously chirping the grating period and curving the grating fringes, an output can be generated that converges onto a point focus. This is important in one of the applications discussed in Chapter 14, the integrated optic disc pickup.
PROBLEMS

7.1. The figure below shows the cross-section of an FTIR beamsplitter, formed from two layers of dielectric of refractive index \( n_1 \), separated by a layer of refractive index \( n_2 \) and thickness \( h \). A plane wave is incident on the upper interface as shown. Assuming TE incidence, and that \( \theta_1 > \sin^{-1}(n_2/n_1) \), write down general expressions for the electric field in all three layers. What boundary conditions are satisfied by these expressions?

![Incident plane wave](image.png)

7.2. Repeat the analysis of Question 7.1, assuming a solution in modal form. Show that, provided \( h \) is large enough, the amplitude of the transmitted wave decreases exponentially with \( h \).

7.3. A high-index glass prism (\( n = 1.7 \)) is used at 1.5 \( \mu\text{m} \) wavelength to characterise a planar guide formed by ion exchange in glass. Assuming that the guide supports three modes, with effective indices of 1.515, 1.530 and 1.545, respectively, sketch the variation with launch angle that you would expect in prism reflectivity. What are the propagation constants of the three modes?

7.4. A step-change in thickness is etched into a planar guide. On the left of the step, the guide is two-moded (with effective indices for the two modes of 1.504 and 1.505, respectively). On the right, it is single-moded, with effective index 1.5025. The lowest-order mode is incident from the left at an angle \( \theta_1 \). Sketch the phase matching diagrams you would expect for (a) \( \theta_1 \approx 0^\circ \), and (b) \( \theta_1 = 87.5^\circ \), and interpret your results.

7.5. Find the most compact arrangement you can for a \( x5 \) telescope beam expander, assuming that it is to be fabricated on a planar waveguide using step-index overlay lenses. The input beam width is 0.5 mm, the effective index of the guide is 1.6, and the maximum change in effective index that can be obtained using the overlay without inducing mode conversion is 0.02.

7.6. Calculate the Bragg angle for an unslanted transmission grating of period \( \Lambda = 1.0 \mu\text{m} \), in a material of refractive index 1.6. The optical wavelength is 1.5 \( \mu\text{m} \). Estimate the spectral bandwidth of the grating, assuming that it is optically thick, with a physical thickness of 50 \( \mu\text{m} \). How many fringe planes are crossed by the input beam, before it emerges from the grating?

\[27.95^\circ; 56.5 \text{ nm}; \approx 27\]

7.7. Show that the angular bandwidth of an unslanted transmission grating of period \( \Lambda \) and thickness \( d \) is given by \( \Delta \theta_0 = \Lambda/d \). Hence calculate the bandwidth of a grating of...
thickness 20 µm in a medium of refractive index 1.6, assuming that the Bragg angle is 30° at \( \lambda_0 = 0.633 \) µm.

[1.13o]

7.8 Show that Bragg's law may be written in terms of the grating vector and the propagation constant as \( 2k \sin(\theta_0 - \phi) = |\mathbf{K}| \). Interpret this geometrically using a phase-matching diagram.

7.9 Sketch the K-vector diagrams you would expect for the following geometries: (a) an unslanted transmission grating of thickness 5 µm and period 5 µm, for normal incidence, and (b) an unslanted reflection grating of thickness 5 µm and period 0.5 µm, for incidence at the Bragg angle. The refractive index of the medium is 1.6, and the optical wavelength is 0.633 µm.

7.10 Discuss the advantages and disadvantages of Bragg grating lenses, as compared with other types of waveguide lens. You should consider size, bandwidth, and ease of fabrication.
SUGGESTIONS FOR FURTHER READING


Ohmachi Y. "Acousto-optic TE0-TM0 mode conversion in a thin film of amorphous tellurium dioxide" Elect. Lett. 9, 539-541 (1973)


8.1 OPTICAL FIBRE TYPES

In this Chapter, we shall consider the properties of optical fibre, by far the most important type of waveguide for the transmission of information at optical frequencies. Optical fibre usually (though not always) has a circular symmetry, and is most often fabricated from very pure silica, or SiO₂. This has a refractive index of \( n = 1.458 \) at \( \lambda = 850 \) nm. Figure 8.1-1 shows the typical geometry of a silica fibre; by selective doping, the centre of the fibre (which is known as the core) is arranged to have a higher refractive index than the surrounding region (the cladding), thus forming a waveguide. Useful dopants include germania (GeO₂) and phosphorus pentoxide (P₂O₅), both of which increase the refractive index of silica, and boric oxide (B₂O₃) and fluorine (F₂), which reduce it. Thus, a typical fibre might consist of a GeO₂:SiO₂ core, with a SiO₂ cladding. Alternatively, a pure SiO₂ core could be used, with a B₂O₃:SiO₂ cladding.

Other low-loss fibre materials that have been investigated include low melting-point silicate glasses (soda-lime silicates, germanosilicates and borosilicates) and halide crystals (TlBr, TlBr, KCl, CsI and AgBr). Most recently, attention has been transferred to the fluoride glasses (mixtures of GdF₃, BaF₂, ZrF₄ and AlF₃), for reasons which will be discussed in the following section. However, whatever the exact composition of the fibre, the cladding is almost always surrounded by a protective plastic jacket, which acts as a strain relief and prevents the ingress of water.

Alternatively, plastic-coated silica fibres (or PCS fibres, which might consist of a natural quartz core clad in silicone resin) and all-plastic fibres (for example, consisting of a polystyrene core with a methyl methacrylate cladding) are often used. Although these are cheap, they suffer from considerably higher propagation loss than silica fibre (\( \approx 10 \) dB/km for PCS fibre, and \( \approx 500 \) dB/km for all-plastic fibre), and are therefore suitable only for transmission over short distances. For all-plastic fibres, the limit is normally tens of metres.

There are many variations on the basic geometry. Some are illustrated in Figure 8.1-2, which shows a number of different cylindrically symmetric fibre cross-sections, and the radial refractive index distribution defining them. The most important distinctions are as follows. Firstly, optical fibre is generally either single-moded or very multi-moded - the number of modes supported by a multi-moded fibre is normally very large (hundreds). Secondly, the core may be defined either by an abrupt increase in the refractive index, or by
a more gradual variation in index (which is typically parabolic). The former is known as step-index fibre, the latter as graded-index fibre.

![Image of fibre types]

The dimensions of these fibre types differ quite considerably. Single-mode fibre has roughly the dimensions of a human hair, with a cladding diameter of the order of 125 µm, and a core diameter of 8 - 12 µm (depending on the operating wavelength). Multi-mode fibre is rather thicker: cladding diameters range from 125 - 400 µm, while the core diameters range from 50 - 200 µm. For field use, fibres are supplied in cable form. Cable styles vary widely, but a typical construction would consist of a plastic tube, strengthened with steel wires or Kevlar to resist tensile strain, which is filled with the fibre in a loose helical lay. A recent development has been the availability of fibre in the form of ribbon cable. This consists of a number (e.g. 8) of single-mode fibres held parallel in a flat ribbon by a plastic matrix. The individual fibres are often colour-coded, and may be separated out into individual strands, allowing multiple-fibre links to be constructed very simply.

### 8.2 LOSS IN SILICA AND FLUORIDE GLASS FIBRE

One of the most important features of optical fibre is its exceptionally low propagation loss, which is normally quoted in units of decibels (dB), according to:

$$\text{loss} = -10 \log_{10}(P_{\text{out}}/P_{\text{in}}) \text{ dB}$$  \hspace{1cm} 8.2-1

Thus, reduction of the power by 50% between input and output corresponds to 3 dB loss. Expressed in this way, loss varies linearly with distance, so fibre may be characterised by attenuation in dB/km. Current fibre performance has been achieved through a considerable research and development effort, which began in the mid-1960s. In fact, loss in silica-based fibre has been reduced quite staggeringly, from more than 1000 dB/km in 1966 to the present value of $\approx 0.2$ dB/km (at $\lambda_0 = 1.55 \mu$m).

There are a number of fundamental mechanisms, which combine to set a lower limit on transmission loss. These are illustrated by Figure 8.2-1, which shows a typical plot of attenuation versus optical wavelength for silica fibre. At short wavelengths ($\lambda_0 < 1.6 \mu$m), attenuation is mainly dominated by scattering loss (rather than by absorption due to...
electronic transitions at ultraviolet wavelengths, as might be expected). The effect itself is known as **Rayleigh scattering**; its origins lie in any small inhomogeneities and imperfections in the structure of the glass forming the fibre - for example, compositional fluctuations on solidification, trapped gas bubbles, dopants and so on. It can be shown that this type of scattering varies as $1/\lambda_0^4$, so it is responsible for the sharp rise in attenuation in Figure 8.2-1 at short wavelengths.

![Figure 8.2-1 Typical attenuation characteristics for silica-based optical fibre.](image)

At longer wavelengths, however, the attenuation is mainly caused by **intrinsic absorption**, arising from the excitation of lattice transitions at near-infrared wavelengths. Minimum attenuation is therefore obtained when the Rayleigh scattering and infrared absorption curves cross, which occurs at around $\lambda_0 = 1.55 \mu m$. If low loss is the main criterion, this is the most desirable operating point. Modern fibres approach the two envelopes of Rayleigh scattering and infrared absorption very closely. However, there may also be significant attenuation near a number of discrete wavelengths, as can be seen in Figure 8.2-1. This type of **extrinsic absorption** can be due either to the dopants themselves, or to additional impurities. The most significant absorption band lies at $\lambda_0 = 1.39 \mu m$, and is caused by the presence of residual hydroxyl (OH) ions, which also give rise to a number of smaller absorption peaks. These ions originate as water contamination, and must be removed by careful dehydration. If this is done, the ultimate limit of low-loss can be achieved almost exactly.

If different materials could be found that have their infrared absorption lines at longer wavelengths, the operating wavelength could be shifted in the same direction. The effect of Rayleigh scattering would then be reduced, and lower minimum loss obtained. This is the thinking behind current research on **fluoride glass fibres**. It is believed that such fibres could have an optimum attenuation as low as 0.001 dB/km, an incredibly small figure. Figure 8.2-2 shows theoretically predicted loss spectra for two different fluoride glass fibres, based on (a) BaF$_2$-GdF$_3$-ZrF$_4$ glass and (b) CaF$_2$-BaF$_2$-YF$_3$-AlF$_3$ glass. These suggest that improved performance should be available at slightly longer wavelengths, in the range $\lambda_0 = 2.5 - 3.5 \mu m$. Although this encouraging prediction has yet to be fulfilled (the current 'best' is only 0.7 dB/km), research is actively being continued.
8.3 STEP-INDEX OPTICAL FIBRES

We shall now consider the propagation of rays and guided modes in optical fibre, beginning with step-index fibre. Functionally, this operates in a similar way to the slab guide introduced in Chapter 6; the fibre is constructed from a core of radius \( a \) and refractive index \( n_1 \), and a cladding of refractive index \( n_2 \) (such that \( n_1 > n_2 \)), as shown in Figure 8.3-1a.

![Figure 8.3-1](image)

Figure 8.3-1 (a) The refractive index distribution of step-index fibre; (b) and (c) propagation of meridional and skew rays.

Figure 8.2-2 Predicted loss spectra for two fluoride glass fibres (based on data by Shibata et al., Elect. Lett. 17, 775-777 (1981))
The guiding mechanism is again total internal reflection at the core/cladding interface. However, consideration of the possible paths that may be followed by rays travelling down the fibre reveals a new feature.

Two entirely different types of ray are now possible. **Meridional rays** follow pathways that can be drawn on a single plane (as shown in Figure 8.3-1b). These are directly analogous to the rays that propagate in planar slab guides. **Skew rays**, on the other hand, do not travel in the exact centre of the fibre, but propagate in an annular region near the outer edge of the core, following off-axis helical paths that effectively circle the fibre axis after a finite number of reflections. These have no such analogy in a slab model, and are best visualised in terms of their projection on the fibre cross-section (Figure 8.3-1c). The existence of such rays implies that the analysis of an optical fibre will be more difficult than a simple extension to cylindrical geometry would suggest.

**NUMERICAL APERTURE**

One of the most important parameters of a fibre is its **numerical aperture**, or N.A. This determines the acceptance cone of the fibre, and the angular spread of radiation it may emit. Despite the distinction between meridional and skew rays outlined above, a reasonable approximation to the NA may be obtained by considering only the former, as shown in Figure 8.3-2. The extent of the acceptance cone will be determined by the maximum angle \( \theta_{\text{emax}} \) that an external ray entering the end-face of the fibre may have, while still giving rise to totally-reflected rays within. This is limited by the onset of cutoff, so the largest angle \( \theta_{1\text{max}} \) allowed inside the fibre must be set by the critical angle, \( \theta_c \).

![Figure 8.3-2](image)

**Figure 8.3-2** Geometry for calculation of the numerical aperture of a step-index fibre,

Simple trigonometry shows that \( \theta_{1\text{max}} = \pi/2 - \theta_c \). The internal and external angles \( \theta_{1\text{max}} \) and \( \theta_{\text{emax}} \) are, of course, related by Snell's law, so assuming that the fibre is surrounded by a medium of refractive index \( n_e \), we must have:

\[
n_e \sin(\theta_{\text{emax}}) = n_1 \sin(\theta_{1\text{max}}) = n_1 \cos(\theta_c)
\]

8.3-1

However, \( \theta_c \) is defined by the fibre core and cladding indices, as \( \theta_c = \sin^{-1}(n_2/n_1) \). After slight re-arrangement, we may then express Equation 8.3-1 as:

\[
n_e \sin(\theta_{\text{emax}}) = \sqrt{n_1^2 - n_2^2}
\]

8.3-2

According to the definition in Chapter 4, this is equal to the numerical aperture of the fibre. A large difference between core and cladding indices will therefore usually result in a high NA. Now, if the external medium is air, we know that \( n_e = 1 \). Using a small-angle
approximation, we may then put \( \sin(\theta_{\text{emax}}) = \theta_{\text{emax}} \). The solid angle \( \Omega \) of the acceptance cone is then:

\[
\Omega = \pi \theta_{\text{emax}}^2 \approx \pi (n_1^2 - n_2^2)
\]

8.3-3

If skew rays are considered as well, it can be shown that both the numerical aperture and the value of \( \Omega \) derived above are under-estimates, but they remain reasonable approximations.

**EXACT MODES OF A STEP-INDEX FIBRE**

Full analysis of a step-index fibre requires the solution of Maxwell’s equations for a cylindrical geometry. Consequently, the fields are constructed from more abstract mathematical functions than were required for the slab guide (actually, Bessel functions), and the modes themselves are more complicated. For many of them, the axial field components \( E_z \) and \( H_z \) are both non-zero, so these modes are known as **hybrid** rather than TE or TM modes. Furthermore, due to the circular cross-section, the transverse fields have an angular variation as well as a radial one, so they require two identifying parameters instead of a single mode number. Because of these complications, we will omit the detailed analysis (to be found in many other texts) and present only the broad conclusions.

The modes of the step-index optical fibre are, in general, classified as \( \text{TE}_{0,\nu}, \text{TM}_{0,\nu}, \text{EH}_{\mu,\nu} \) or \( \text{HE}_{\mu,\nu} \) types. The last two are hybrid fields, which have \( E_z \) and \( H_z \) as the dominant axial component, respectively. The most important mode is the \( \text{HE}_{1,1} \) mode (roughly equivalent to the \( \text{TE}_0 \) and \( \text{TM}_0 \) modes of the slab guide) which cuts off at zero frequency. However, the \( \text{TE}_{0,1} \) and \( \text{TM}_{0,1} \) modes are also significant. These have the next lowest cutoff frequencies, and so their presence or absence may be used to define whether the fibre is multi-moded or not. Now, in Chapter 6 we showed that a parameter \( V = (k_0 h/2) \sqrt{(n_1^2 - n_2^2)} \) could be used to characterise a symmetric slab guide of thickness \( h \) and core and cladding indices \( n_1 \) and \( n_2 \). In the same way, a \( V \)-value may be defined for a step-index fibre of core radius \( a \) and similar indices, in the form:

\[
V = (k_0 a) \sqrt{(n_1^2 - n_2^2)}
\]

8.3-4

It turns out that the \( \text{TE}_{0,1} \) and \( \text{TM}_{0,1} \) modes cut off when \( V = 2.405 \). Thus, if the frequency is such that \( V < 2.405 \), only one mode will propagate and the fibre is single-moded.

Assuming that \( n_1 = n_2 = n \), and \( n_1 - n_2 = \Delta n \) (which will be the case in weakly-guiding fibres) the minimum wavelength for single-mode operation is therefore:

\[
\lambda_0 = (\pi d/2.405) \sqrt{(2\Delta n)}
\]

8.3-5

where \( d \) is the fibre diameter. For \( d = 8 \mu\text{m}, n = 1.5, \) and \( \Delta n = 0.001 \), say, we require \( \lambda_0 > 0.57 \mu\text{m} \).

**LP MODES**

The hybrid modes have a complicated polarization, although the \( \text{HE}_{1,1} \) mode is almost plane-polarized. However, for weak guides, it can be shown that all the other modes can be grouped together in combinations, which are **degenerate** (which means they have the same propagation constant or phase velocity). A linear combination of these modes can then be used to construct a new set of nearly plane-polarized fields, which represent a considerable simplification over the exact ones. These modes are called 'linearly polarized', and are given new labels of the form \( \text{LP}_{\mu,\nu} \). Table 8.3-1 shows the constituents of some low-order LP
modes. The particular case of the $\text{HE}_{1,1}$ mode is the $\text{LP}_{0,1}$ mode; as a rule, the $\text{LP}_{1,\nu}$ mode is constructed from the $\text{HE}_{2,\nu}$, $\text{TM}_{0,\nu}$ and $\text{TE}_{0,\nu}$ modes, and the $\text{LP}_{\mu,\nu}$ mode from the $\text{HE}_{\mu+1,\nu}$ and $\text{EH}_{\mu-1,\nu}$ modes (for $\mu \neq 0$ or 1).

<table>
<thead>
<tr>
<th>LP mode</th>
<th>Exact mode constituents</th>
<th>LP mode</th>
<th>Exact mode constituents</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{LP}_{0,1}$</td>
<td>$\text{HE}_{1,1}$</td>
<td>$\text{LP}_{0,2}$</td>
<td>$\text{HE}_{1,2}$</td>
</tr>
<tr>
<td>$\text{LP}_{1,1}$</td>
<td>$\text{HE}<em>{2,1}$, $\text{TM}</em>{0,1}$ &amp; $\text{TE}_{0,1}$</td>
<td>$\text{LP}_{1,2}$</td>
<td>$\text{HE}<em>{2,2}$, $\text{TM}</em>{0,2}$ &amp; $\text{TE}_{0,2}$</td>
</tr>
<tr>
<td>$\text{LP}_{2,1}$</td>
<td>$\text{HE}<em>{3,1}$ &amp; $\text{EH}</em>{1,1}$</td>
<td>$\text{LP}_{2,2}$</td>
<td>$\text{HE}<em>{3,2}$ &amp; $\text{EH}</em>{1,2}$</td>
</tr>
</tbody>
</table>

Table 8.3-1 Constituents of some low-order LP modes

Figure 8.3-3 shows approximate representations of the intensity distributions of three low-order LP modes. The $\text{LP}_{0,1}$ mode is circularly-symmetric, with a field distribution containing a single peak at the centre of the fibre core. The higher-order modes, on the other hand, have multilobed patterns, with $2\mu$ maxima as measured in the circumferential direction, and $\nu$ maxima in the radial direction.

![Figure 8.3-3 Approximate intensity distributions for the LP_{0,1}, LP_{2,1} and LP_{4,2} modes.](image)

The dispersion diagram for the step-index fibre is of the general form shown in Figure 8.3-4. The main difference in relation to the slab guide is that higher-order modes start to propagate in much greater numbers as soon as the $V$-value exceeds the limit for single-mode operation.

![Figure 8.3-4 General dispersion diagram for step-index optical fibre.](image)
In fact, it can be shown that the number of modes propagating is given approximately by:

\[ N \approx \frac{V^2}{2} \]

8.3-6

provided \( N \) is large. A fibre with a \( V \)-value of 10 would therefore support \( \approx 50 \) modes.

**DESIGN EXAMPLE**

Using the approximate guidelines presented so far, we may compare the main features of single and multimode fibres. With the typical values of \( d = 8 \, \mu m \), \( n_1 = 1.5 \) and \( n_2 = 1.499 \) for the former, and \( d = 50 \, \mu m \), \( n_1 = 1.5 \) and \( n_2 = 1.49 \) for the latter, we obtain the parameters shown in Table 8.3-2 for a wavelength of \( \lambda_0 = 0.85 \, \mu m \). The high numerical aperture of the multimode fibre should be noted, together with the large number of modes that it supports.

<table>
<thead>
<tr>
<th>Fibre</th>
<th>( d ) (( \mu m ))</th>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>NA</th>
<th>V-value</th>
<th>No. of modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>8</td>
<td>1.5</td>
<td>1,499</td>
<td>0.055</td>
<td>1.62</td>
<td>1</td>
</tr>
<tr>
<td>#2</td>
<td>50</td>
<td>1.5</td>
<td>1.49</td>
<td>0.17</td>
<td>31.95</td>
<td>510</td>
</tr>
</tbody>
</table>

Table 8.3-2  Characteristic parameters of typical single- and multi-mode fibres.

**OTHER FIBRE MODES**

Just as in the slab guide, we may expect the guided modes of a fibre to be complemented by radiation modes, and that the full range of modes will exist in two sets, each corresponding to one of two possible polarizations. While this is true, there are two important new features. Firstly, the radiation modes are normally subdivided into fields that propagate inside and outside the cladding. The former are known as cladding modes, and are often a nuisance. They may be removed using a device known as a mode-stripper. In this, the plastic jacket is removed from a short section of the fibre, which is placed on a glass plate in a pool of index-matching liquid. Reflection at the edge of the cladding is then greatly reduced, and most of the unwanted light radiates into the glass plate. Secondly, since the fibre is cylindrically symmetric, the two sets of polarization modes are degenerate. A single-moded fibre therefore actually supports two modes of orthogonal polarization, which propagate at the same speed. As we will show later, this normally results in a random polarization fluctuation.

**8.4 PARABOLIC-INDEX OPTICAL FIBRES**

We now consider parabolic-index fibre. Here, the possible ray trajectories are still more complicated. Because the refractive index distribution is now a continuous function of radius, rays no longer propagate as straight lines, with total internal reflection causing abrupt changes in direction. Instead, their paths are determined by a process of continual refraction. This has the effect of gradually bending the direction of any ray travelling away from the axis, until it turns towards the axis once again. For meridional rays, the result is a sinusoidal trajectory, oscillating on either side of the axis. Figure 8.4-1a shows two different ray paths in the same fibre. As we shall shortly demonstrate, each has almost exactly the same periodicity in the axial direction, although the turning points (e.g. \( z_a \) and \( z_b \)) lie at different radii. This has important consequences for signal transmission.
For skew rays, a similar process of continual refraction occurs. Consequently, the projection of a skew ray trajectory on the fibre cross-section has the appearance of a smooth, rotating orbit, as shown in Figure 8.4-1b. In fact, the ray path can again be shown to lie entirely between two coaxial cylinders. These are known as **caustic surfaces**, and represent the locii of all possible turning points.

Figure 8.4-1  (a) and (b) Trajectories of meridional and skew rays in a parabolic-index fibre.

RAY TRAJECTORIES IN PARABOLIC-INDEX MEDIA

Using some fairly simple scalar analysis, it is possible to calculate the trajectory of a ray in a graded-index guide, as we now show. We start by noting that the exact refractive index profile of the fibre is given by the full line in Figure 8.4-2. However, on the assumption that the light is confined mainly to the region of parabolic variation, we can disregard the index steps at the edge of the core and cladding, and model the index distribution by a continuous parabolic function of the form:

\[ n(r) = n_0 \sqrt{1 - (r/r_0)^2} \]  \(8.4-1\)

where \(n_0\) is the refractive index at the centre of the fibre, and \(r_0\) is a characteristic radius. This profile is shown dashed in Figure 8.4-2; while it implies a negative refractive index when \(r > r_0\), we can assume that \(r_0\) is large enough for this to be unimportant for all practical purposes.

Figure 8.4-2  Exact and approximate refractive index distribution of a parabolic index fibre.

At this point, we shall adopt a one-dimensional model, ignoring the y-variation of the refractive index and constraining the trajectory to lie in the x-z plane. This will allow us to deduce many of the features of a two-dimensional guide in a qualitative way. Later on, we will check for agreement with the more complicated geometry. With this approximation, we may put \(r = x\) and \(r_0 = x_0\) in Equation 8.4-1. Now, the direction of the ray at any point \((x, z)\) may be then defined by a vector \(\mathbf{k}\), such that \(|k|\) is equal to the local value of the propagation constant, \(k_0 n(x)\). Similarly, the local slope of the trajectory may be found as:
dx/dz = k_x/k_z  

where k_x and k_z are the x- and z-components of k. However, if the entire trajectory corresponds to a guided mode, we may also set \( k_z = \beta \), where \( \beta \) is the modal propagation constant. Since \( k_x \) and \( k_z \) are related by \( k_x^2 + k_z^2 = |k|^2 \), we obtain:

\[
k_x = \sqrt{(k_0^2 n_0^2 - \beta^2)}
\]

Substituting from Equation 8.4-1, we then get:

\[
k_x = \sqrt{(k_0^2 n_0^2 - \beta^2) - k_0^2 n_0^2 (x/x_0)^2}
\]

Equation 8.4-2 may be therefore written as:

\[
\beta^2 (dx/dz)^2 = (k_0^2 n_0^2 - \beta^2) - k_0^2 n_0^2 (x/x_0)^2
\]

We might guess that the solution of Equation 8.4-5 is indeed a sinusoidal trajectory, of the general form \( x = A \sin(\beta z + \Psi) \), where \( \Psi \) is an arbitrary phase factor. We shall now try to verify this hypothesis, and simultaneously determine the unknown constants \( A \) and \( B \). Performing the necessary differentiation, substituting into Equation 8.4-5, and rearranging the terms, we get:

\[
(k_0^2 n_0^2 - \beta^2) - (k_0^2 n_0^2 A^2/x_0^2) \sin^2(\beta z + \Psi) - \beta^2 A^2 B^2 \cos^2(\beta z + \Psi) = 0
\]

Now, the only way Equation 8.4-6 can be satisfied for all \( z \) is if the \( \sin^2 \) and \( \cos^2 \) terms can somehow be removed. This can be done if their coefficients are equal (since \( \sin^2(\theta) + \cos^2(\theta) = 1 \)), which requires:

\[
k_0^2 n_0^2 A^2/x_0^2 = \beta^2 A^2 B^2 \quad \text{or} \quad B = k_0 n_0/\beta x_0
\]

This places an immediate restriction on the value of \( B \), the spatial frequency of the periodic oscillation of the trajectory. Assuming that Equation 8.4-7 is indeed satisfied, Equation 8.4-6 reduces to:

\[
(k_0^2 n_0^2 - \beta^2) - (k_0^2 n_0^2 A^2/x_0^2) = 0
\]

so that the amplitude \( A \) of the oscillation is given by:

\[
A = x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2}
\]

Equation 8.4-9 shows that the smaller the value of \( \beta \) (i.e., the higher the order of the mode) the larger the amplitude. Our solution for the ray trajectory is therefore:

\[
x = x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2} \sin((k_0 n_0/\beta x_0)z + \Psi)
\]

Now, for weak guides (when \( \beta \approx k_0 n_0 \) for all modes), \( B \approx 1/x_0 \). With this approximation, we can put:

\[
x = x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2} \sin(z/x_0 + \Psi)
\]
Equation 8.4-11 implies that the spatial frequency of the trajectory is virtually independent of $\beta$, i.e. the same for all modes. It turns out that the above analysis is modified somewhat in the case of a two-dimensional parabolic-index fibre, but only qualitatively so, providing comforting evidence for the shape of the ray paths in Figure 8.4-1. More importantly, it allows the use of a parabolic-index medium as a type of imaging device, known as a **GRIN-rod** lens (where the GRIN acronym stands for gradient-index).

**GRIN-rod LENSES**

A GRIN-rod lens is a short length of glass rod of moderately large diameter (1 - 3 mm), which has the radial index variation of Equation 8.4-1. Formally, it is a section of highly multimoded fibre, but in practice no cladding is used. It is fabricated by ion exchange of thallium- or caesium-doped silica glass, in a bath of a molten salt. Typically, KNO$_3$ is used, at a temperature of 500°C. During the exchange, sodium and either thallium or caesium ions diffuse out of the glass, and are replaced by potassium ions from the melt. The difference in polarizability of K$^+$ and Tl$^+$ ions then causes a reduction in the refractive index. Now, the local concentration of Tl$^+$ at the end of the process is determined by standard diffusion equations. Due to the cylindrically symmetric geometry, the concentration is higher at the centre of the rod than at the periphery, following a parabolic variation in between. The concentration of K$^+$ ions, on the other hand, is approximately uniform. The net result is a refractive index profile corresponding well to the ideal variation. GRIN-rod lenses are produced commercially by the Nippon Sheet Glass Co. of Japan, under the trade name 'SELFOC micro-lens'.

We shall again base the following argument on one-dimensional analysis. Firstly, we can guess that rays in a GRIN-rod lens will also follow periodic trajectories. The period $P$ of the oscillation is defined as the **pitch length**, given by:

$$P = 2\pi x_0$$  \hspace{1cm} 8.4-12

It is normal for the lens length to be chosen as some fraction of a pitch, close to either $P/2$ (a half-pitch lens) or $P/4$ (quarter-pitch). If this is done, a GRIN-rod lens may perform imaging operations, as we now demonstrate. Figure 8.4-3a shows a half-pitch lens, used to form an image of a point source of light located on-axis at $A$.

![Figure 8.4-3](image)

**R.A.A.Syms and J.R.Cozens**  Optical Guided Waves and Devices 11
Naturally, this source will emit rays in all directions, but each will begin at the same point. Assuming that this is defined as $x = 0$, $z = 0$, the trajectories of all rays inside the lens have the form of Equation 8.4-11, with the phase factor $\Psi = 0$. Each sinusoidal trajectory will have a different amplitude, so that the slope of the trajectory at $z = 0$ matches the direction of an input ray. However, after travelling a distance $z = P/2$ through the lens, all rays will have returned to the axis once again, at the point $A'$. Consequently, the lens may be considered to form a real image of $A$ at $A'$. Since an off-axis point object above $A$ can be also shown to form a real point image displaced the same distance below $A'$, the lens forms an inverted image of the entire input plane at the output end-face. Similar reasoning shows that a quarter-pitch lens forms an image of a point source at infinity, thus collimating a point emitter (Figure 8.4-3b).

For practical reasons, GRIN-rod lenses are normally made in lengths somewhat shorter than those above. For example, Figure 8.4-3c shows a 0.23P lens, which will collimate light from a point source slightly displaced from the lens, thus allowing a working tolerance for alignment and so on. This is particularly important, since dispersion of the lens material generally results in a wavelength-dependent value of $P$. Without such a tolerance, operation at a wavelength for which $P$ is shorter than the design value would require the source to be placed inside the lens.

GRIN-rod lenses have several advantages over conventional lenses. Since they are small, they are well-suited to use in micro-optic systems, and are often used to connect optical fibre components. Most importantly, their imaging properties do not stem from refraction at their input and output surfaces (which are planar); instead, redirection of rays is performed by the medium itself. This allows components to be fixed directly to the end faces with index-matching cement, reducing reflection losses and eliminating scratchable or movable interfaces. Consequently, the assembly can be highly rugged. Figure 8.4-4 shows a collection of GRIN-rod lenses.

![GRIN-rod lenses](image)

Figure 8.4-4. GRIN-rod lenses (photograph courtesy D.DeRose, NSG America Inc.).

The ray picture above is just one explanation for the imaging behaviour of a GRIN-rod lens, and an alternative view is provided by the transfer function model introduced in Chapter 4. Remember that the properties of a conventional lens stemmed from the parabolic variation of its optical thickness. The required variation was effected by controlling the thickness of a material with constant refractive index. In a GRIN-rod lens, much the same result is achieved by control of the radial distribution of refractive index, keeping the physical thickness constant. We shall provide one further explanation later on, in terms of guided modes.
THE EIGENVALUE EQUATION FOR A PARABOLIC-INDEX GUIDE

The solution for ray paths given in Equation 8.4-10 represents almost the entire story. However, to obtain a closed form solution, we need to know the allowed values of $\beta$. This requires the solution of an eigenvalue equation. To set up a suitable equation, we shall make use of the transverse resonance condition previously introduced for slab guides in Chapter 6, which required the total phase change accumulated in a round trip between the guide walls to be a whole number of multiples of $2\pi$. Remember that this consisted of two terms: the phase change incurred simply in propagating between the walls, and the phase shift resulting from reflection at the walls.

In the one-dimensional parabolic-index guide, propagation between any two consecutive turning points (e.g. $z_a$ and $z_b$) clearly corresponds to half of one round trip. The first term may be therefore be found by integrating the x-component of the propagation constant over the transverse distance between these points, as:

\[
2 \int_{x_a}^{x_b} k_x \, dx = 2 \int_{z_a}^{z_b} k_x \left( \frac{dx}{dz} \right) \, dz
\]

Equation 8.4-13

However, using Equation 8.4-2, this may be transformed to:

\[
2 \int_{x_a}^{x_b} k_x \, dx = \left( \frac{2}{\beta} \right) \int_{z_a}^{z_b} k_x^2 \, dz
\]

Equation 8.4-14

Substituting our solution for the trajectory (Equation 8.4-10) into Equation 8.4-4, and inserting the result into Equation 8.4-14, we then get:

\[
2 \int_{x_a}^{x_b} k_x \, dx = \left( \frac{2}{\beta} \right) \left( k_0^2 n_0^2 - \beta^2 \right) \int_{z_a}^{z_b} \cos^2 \left( \frac{k_0 n_0}{\beta} x_0 \right) + \Psi \right) \, dz
\]

Equation 8.4-15

Now, since $z_a$ and $z_b$ are separated by half a cycle, we must have $\left( \frac{k_0 n_0}{\beta} x_0 \right) \left( z_b - z_a \right) = \pi$. Integrating over this distance, we get:

\[
2 \int_{x_a}^{x_b} k_x \, dx = x_0 \pi \left( k_0 n_0 - \beta^2/k_0 n_0 \right)
\]

Equation 8.4-16

The second term - the phase change experienced on reflection - is more difficult to deal with. After all, if the ray path is determined by continuous refraction, total internal reflection is not needed to change the direction of the ray. Nonetheless, the sign of $k_x$ does actually change at each of the turning points, which implies that some form of reflection must be occurring there. To assess the consequences, we need to return to Chapter 5, where we found general expressions for the reflection coefficients for TE- and TM-polarized light at a dielectric interface. In each case, it can be shown that when the index step tends to zero and the angle of incidence approaches 90°, the reflection coefficient tends to $\exp(j\pi/2)$. Consequently, there is an additional phase shift of $\phi_a = -\pi/2$ at each of the turning points.

Including all contributions, the transverse resonance condition can therefore be written:

\[
2 \int_{x_a}^{x_b} k_x \, dx + 2\phi_a = 2v\pi \quad (v = 1, 2, \ldots)
\]

Equation 8.4-17

Inserting the values found above, we then get:

\[
x_0 \pi \left( k_0 n_0 - \beta^2/k_0 n_0 \right) - \pi = 2v\pi
\]

Equation 8.4-18

which yields the following allowed values for the propagation constant:
\[ \beta_\nu = k_0 n_0 \sqrt{1 - \frac{2(2\nu + 1)}{k_0 n_0 x_0}} \]  
8.4-19

If the second term in the square root above is small (which implies weak guidance, and therefore paraxial ray paths), we may approximate Equation 8.4-19 as:

\[ \beta_\nu \approx k_0 n_0 - \frac{(2\nu + 1)}{2x_0} \]  
8.4-20

Since \( k_0 = \frac{\omega}{c} \), the dispersion characteristics of parabolic-index guide consist of a set of parallel straight lines, as shown in Figure 8.4-5. This has extremely important consequences for signal transmission in fibres, which we shall discuss later on. Notice that we have deliberately obscured details of the cutoff conditions, when ray paths must stray out to the edge of the core, since these are not modelled realistically using the approximate index profile we have assumed.

![Figure 8.4-5 Dispersion characteristics of a one-dimensional parabolic-index guide.](image)

In the mean time, we simply note the following result, which is relevant to image transmission through a GRIN-rod lens. As we saw in Chapter 6, a general one-dimensional, forward-travelling field \( E_{in}(x) \) specified at an input plane \( z = 0 \) may be expanded in terms of guided modes as:

\[ E_{in}(x) = \sum_\nu a_\nu E_\nu(x) \]  
8.4-21

where \( a_\nu \) is the amplitude of the \( \nu \)th mode, and \( E_\nu(x) \) is its transverse field distribution. After propagating a distance \( z \), this field will be modified to:

\[ E_{out}(x) = \sum_\nu a_\nu E_\nu(x) \exp(-j\beta_\nu z) \]  
8.4-22

Inserting the propagation constants specified by Equation 8.4-20 into the above, we get:

\[ E_{out}(x) = \exp(-j\psi) \sum_\nu a_\nu E_\nu(x) \exp(j\nu z/x_o) \]  
8.4-23

where \( \psi = (k_0 n_0 - 1/2x_o)z \). We now note that if \( z = 2\pi x_o \) (i.e., if \( z = P \), Equation 8.4-23 reduces to:

\[ E_{out}(x) = \exp(-j\psi) \sum_\nu a_\nu E_\nu(x) \]  
8.4-24
At this point, \( E_{\text{out}} \) is equal to \( E_{\text{in}} \) multiplied by an unimportant phase factor. Propagation through one pitch length will therefore leave an image effectively unchanged, as we would expect from the ray model of a GRIN-rod lens described earlier. We can interpret the formation of an inverted real image by a half-pitch lens in a similar way.

### MODAL ANALYSIS OF PARABOLIC-INDEX FIBRE

We shall now show how scalar theory may be used to determine the actual guided modes in a two-dimensional parabolic-index fibre. We base the analysis on the TEM model introduced in Chapter 6, which assumes that the index changes forming the guide are weak. For the refractive index distribution of Equation 8.4-1, the scalar waveguide equation we must solve is:

\[
\nabla^2 \mathbf{E}(x, y) + \{n_0^2 k_0^2 [1 - (r/r_0)^2] - \beta^2\} \mathbf{E}(x, y) = 0
\]

Where \( \mathbf{E}(x, y) \) is the transverse field and \( \beta \) is the propagation constant. Since the fibre is cylindrical, we shall begin by looking for a solution that varies only in the radial direction. In this case, we may put \( \mathbf{E}(x, y) = \mathbf{E}(r) \), and evaluate its derivative with respect to \( x \) as:

\[
\frac{\partial \mathbf{E}}{\partial x} = \frac{\partial \mathbf{E}}{\partial r} \frac{\partial r}{\partial x} = \left(\frac{x}{r}\right) \frac{\partial \mathbf{E}}{\partial r}
\]

Similarly, its second derivative with respect to \( x \) is given by:

\[
\frac{\partial^2 \mathbf{E}}{\partial x^2} = \left[1/r - x^2/r^2\right] \frac{\partial \mathbf{E}}{\partial r} + \left(x^2/r^2\right) \frac{\partial^2 \mathbf{E}}{\partial r^2}
\]

\( \nabla_{\text{xy}}^2 \mathbf{E} \) may be evaluated in the same way, so we obtain:

\[
\nabla^2 \mathbf{E} = \left(\frac{1}{r}\right) \frac{\partial \mathbf{E}}{\partial r} + \frac{\partial^2 \mathbf{E}}{\partial r^2}
\]

Equation 8.4-28 may now be substituted into Equation 8.4-25, to give:

\[
\frac{\partial^2 \mathbf{E}}{\partial r^2} + \left(\frac{1}{r}\right) \frac{\partial \mathbf{E}}{\partial r} + \{n_0^2 k_0^2 [1 - (r/r_0)^2] - \beta^2\} \mathbf{E} = 0
\]

At this point, we shall make a guess at the solution. Somewhat arbitrarily, we assume that the transverse field is given by a Gaussian distribution, of the form:

\[
\mathbf{E}(r) = E_0 \exp\left(-r^2/a^2\right)
\]

where \( E_0 \) and \( a \) are constants. Substituting into Equation 8.4-29, and removing the common terms, we get:

\[
r^2 \left\{4/a^4 - n_0^2 k_0^2/r_0^2\right\} + \{n_0^2 k_0^2 - 4/a^2 - \beta^2\} E = 0
\]

where we have grouped the terms into two blocks, one dependent on \( r^2 \), the other a constant. Now, in order to satisfy Equation 8.4-31 for all \( r \), the two blocks must be zero, independently. The former condition leads to:

\[
a = \sqrt{2r_0/n_0 k_0}
\]

While the latter requires:

\[
\beta = n_0 k_0 \sqrt{\left\{1 - 2/(n_0 k_0 r_0)\right\}}
\]
This shows that the Gaussian transverse field we have assumed is indeed a possible mode of the parabolic-index fibre. Since its phase-front is plane, this mode will radiate into free space from an end-face of the fibre as a Gaussian beam; the beam waist (which has radius a) will be located at the fibre end, and behaviour in the far-field will be as described earlier in Chapter 4.

Using similar methods, but this time adopting rectangular coordinates, it is possible to show that the higher-order guided modal solutions are all Hermite-Gaussian modes, with transverse field variations of the form:

$$E_{\mu,\nu}(x, y) = H_{\mu}(\sqrt{2}x/a) H_{\nu}(\sqrt{2}y/a) \exp\{-x^2 + y^2\}/a^2\}$$

where $\mu$ and $\nu$ are two mode numbers, and $H_{\mu}(\zeta)$ is a Hermite polynomial of order $\mu$, defined as satisfying the differential equation:

$$d^2H_{\mu}(\zeta)/d\zeta^2 - 2\zeta dH_{\mu}(\zeta)/d\zeta + 2\mu H_{\mu}(\zeta) = 0$$

while the propagation constants $\beta_{\mu,\nu}$ are given by:

$$\beta_{\mu,\nu} = n_0 k_0 \sqrt{1 - 2(\mu + \nu + 1)/(n_0 k_0 r_0)}$$

The first few Hermite polynomials are as given in Table 8.4-1. From this, it should be clear that the Gaussian solution above (for which $H_{\mu}(\sqrt{2}x/a) = H_{\nu}(\sqrt{2}y/a) = 1$) is the lowest-order mode, with $\mu = \nu = 0$. This is confirmed by the agreement of Equations 8.4-33 and 8.4-36. Although the latter differs slightly from Equation 8.4-19, found using our earlier ray model, it still has the same general form. Consequently, we may deduce that the general characteristics of a parabolic-index fibre will indeed be qualitatively similar to those for a one-dimensional parabolic-index guide.

### Table 8.4-1. Some low-order Hermite polynomials.

<table>
<thead>
<tr>
<th>$H_0(\zeta)$</th>
<th>$H_1(\zeta)$</th>
<th>$H_2(\zeta)$</th>
<th>$H_3(\zeta)$</th>
<th>$H_4(\zeta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$2\zeta$</td>
<td>$-2 + 4\zeta^2$</td>
<td>$-12\zeta + 8\zeta^3$</td>
<td>$12 - 48\zeta^2 + 16\zeta^4$</td>
</tr>
<tr>
<td>$H_3(\zeta)$</td>
<td>$120\zeta - 160\zeta^3 + 32\zeta^5$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8.5 SIGNAL DISPERSION

We shall now consider a number of aspects of signal propagation in optical fibre, beginning with dispersion. In Chapter 6, we showed that there are generally two contributions to dispersion in waveguides: intermodal and intramodal dispersion. The latter can be further subdivided into contributions from material and waveguide dispersion. We shall consider them separately, and examine the dominant effects in both step- and graded-index fibre. First, we reiterate that all forms of dispersion arise from variations in the group velocity over the signal bandwidth. Intermodal dispersion results from differences in $v_g$ between different modes, and is therefore important only in multimode fibre. However, if it is present, it normally dominates all other contributions.

Clearly, to evaluate the group velocities of the relevant modes, we must have already made an accurate calculation of the dispersion diagram. Since we have not done this for step-index fibre, we must use an approximation. Our argument runs like this. Modes that are far from cutoff will have most of their energy contained inside the core, so their group velocity...
will be close to that of the core material. Ignoring all other forms of dispersion, this will be roughly equal to the phase velocity in a medium of refractive index \( n_1 \), namely \( c/n_1 \).

Similarly, modes that are close to cutoff will have their energy mostly propagating in the cladding, so that their group velocity will be roughly \( c/n_2 \). In a distance \( L \), the time-spread of a signal due to intermodal dispersion is then:

\[
\Delta t = L\{1/v_{g_{\text{min}}} - 1/v_{g_{\text{max}}}\} \approx (L/c)(n_1 - n_2)
\]

Assuming the typical value of \( n_1 - n_2 = 0.01 \), we obtain \( \Delta t/L = 33 \text{ ns/km} \). Now, in a digital communications system, pulses of duration \( T \) are normally sent separated by gaps of \( T \). Intersymbol interference will occur when consecutive pulses start to overlap, i.e. when \( \Delta t \approx T \). Consequently, intermodal dispersion limits the maximum bit rate \( B \) to:

\[
B \approx 1/(2\Delta t)
\]

This is normally expressed in terms of a bit-rate : length product \( BL \), given by:

\[
BL = c/[2(n_1 - n_2)]
\]

For the fibre above, we obtain \( BL \approx 15 \text{ Mbit/s km} \), implying that data may be sent at 15 MBit/s over a distance of 1 km, or at 1.5 MBit/s over 10 km. The small value of \( BL \) renders this fibre impractical for long-distance, high bit-rate communications.

A considerable improvement is offered by parabolic index fibre, for which we have already calculated the dispersion characteristics. Assuming weak guidance, Equation 8.4-36 may be differentiated to give \( d\omega/d\beta_{\mu,\nu} \approx c/n_0 \). This shows that the group velocities of all modes are approximately equal, a highly advantageous feature. The explanation is found in the ray paths of Figure 8.4-1a. Trajectories with small amplitudes of oscillation obviously travel through shorter physical distances than those with large amplitude. However, due to the parabolic index variation, paraxial rays spend more time in regions of high index, and therefore travel more slowly. The two effects - decreased physical distance, and increased refractive index - balance out, so that the transit times of all rays are roughly equalised. Consequently, a drastic reduction in intermodal dispersion is obtained, allowing a considerable increase in the bit-rate : length product, typically to 1 Gbit/s km. For this reason, multimode parabolic- index fibre has been very successful, allowing transmission at respectable data rates over short distances. However, for higher rates and longer distances, it is essential to eliminate intermodal effects entirely by using single-mode fibre. The major limitations on bandwidth then have different origins.

In Chapter 2, we showed that the broadening of a signal due to variations in the propagation constant with frequency could be written as

\[
\Delta t = L\Delta\omega\frac{d^2 k}{d\omega^2}
\]

where \( \Delta\omega \) is the signal bandwidth. Although this was originally derived for plane waves, we would expect that the following analogous expression would describe similar effects caused by variations in the \( \beta \)-value of a single guided mode:

\[
\Delta t = L\Delta\omega\frac{d^2\beta}{d\omega^2}
\]

This phenomenon is known as intramodal dispersion. In general, it consists of contributions from material and waveguide dispersion, as we now show. Firstly, we recall from Chapter 2 that dispersion in plane waves could be related to variations in the refractive index with wavelength, according to

\[
\Delta t = -(L\lambda_0\Delta\lambda_0/c)\frac{d^2n}{d\lambda_0^2}
\]

where \( \Delta\lambda_0 \) is the wavelength range of the signal. For a guided mode, we simply replace \( n \) by \( n_{\text{eff}} \), to obtain:
 Consequently, intramodal dispersion follows from variations in the effective index with wavelength. These will arise if the constituents of the core or the cladding display any wavelength-dependence, i.e. through material dispersion. Now, we have already plotted the dispersion characteristic of pure silica in Chapter 3, and noted that dispersion goes to zero at \( \lambda_0 \approx 1.27 \mu m \). This would be the ideal operating point if signal distortion were the only criterion, but (as we showed earlier) the optimum wavelength for transmission is around \( \lambda_0 = 1.55 \mu m \). However, several methods may be used to move the point of zero dispersion to longer wavelengths. Firstly, the addition of some dopants (e.g. GeO\(_2\)) to silica shifts the point in the required direction. Secondly, there is a further contribution to dispersion that we have not yet included: variations in effective index arise simply from the changes in angle between any guided ray and the fibre axis that must occur with changes in wavelength. This is known as waveguide dispersion, and is independent of material effects.

In step-index fibre, particular parameter choices can lead to waveguide dispersion of opposite sign to the material dispersion, resulting in cancellation of the total dispersion at the required wavelength. Unfortunately, this type of dispersion-shifted fibre normally has a rather small core, leading to problems with jointing. However, by using a more complicated refractive index profile (e.g. a W-shaped profile) it is possible to achieve similar cancellation using cores of larger cross-section.

### 8.6 MODE CONVERSION IN FIBRES

In our analysis, we have concentrated on infinite, perfect guides. The exception has been a discussion of tapers in Chapter 6, where we showed that variations in guide shape can induce mode conversion. Now, although fibre is nominally made with a uniform cross-section, there will inevitably be small irregularities (e.g., at the interface between core and cladding, as in Figure 8.6-1a). Any change in ray direction resulting from scattering by such an irregularity then corresponds to mode conversion. Slight changes in direction (Figure 8.6-1b) lead to similar effects. Over the large distances involved in optical fibre transmission, even a small amount of intermodal scattering can lead to a significant transfer of power between modes. This leads to a range of consequences, which assume differing importance in the various fibre types. We shall begin by discussing the effects in multimode fibre.

![Mode conversion](image)

Figure 8.6-1  Mode conversion, at (a) a small irregularity, and (b) a change in direction

Firstly, it might be thought that intermodal dispersion can be avoided by launching only a single mode. However, since the processes above lead to a gradual transfer of energy to higher-order modes, multimode operation must occur after a finite distance. Intermodal dispersion is therefore inevitable. Secondly, it is likely that scattering will lead to some rays striking the core/cladding interface at angles lower than the critical angle. These will not be reflected, leading to energy loss through the excitation of radiation modes. This is known as...
**microbending loss** if it is caused by small changes in guide direction (or as **bend loss**, if the changes are large). High-order modes, which strike the interface at angles closer to \( \theta_c \), are more susceptible to bend loss. Now, it is likely that intermodal scattering will sooner or later lead to the excitation of modes near to cutoff, which are very sensitive to radiation loss. Consequently, multimode guides show an increase in bend loss with distance. One way to alleviate this is to decrease the critical angle, which requires a large index difference between core and cladding. However, in step-index fibre, this increases intermodal dispersion.

As we have mentioned, intermodal dispersion is not a problem in single-mode guides. Nonetheless, intermodal **scattering** still leads to undesirable effects. Bend loss remains, although this can be be held to acceptable levels by maximising the index change forming the guide, and ensuring that the bend radius always exceeds a specified minimum value. We shall discuss this point more fully in Chapter 9. Here, we will concentrate on an alternative phenomenon: **polarization fluctuation**. This occurs in circularly symmetric fibre, due to random perturbations acting along the length of the fibre.

Now, it can be shown that energy may be coupled most effectively between any two modes by an external perturbation if it is a periodic one, with a spatial frequency that phase-matches the modes. We have already met one example of this in Chapter 7 (the phase matching of beams travelling in different directions by an optically thick grating) and we will encounter some others in Chapter 10. The general idea is that two codirectional modes with propagation constants \( \beta_{\nu} \) and \( \beta_{\mu} \) can be phase-matched by a perturbation with a spatial frequency \( K \) such that:

\[
K = \frac{2\pi}{\Lambda} = |\beta_{\nu} - \beta_{\mu}|
\]

Any random perturbation along the length of a fibre may be resolved into a set of Fourier components, each corresponding to a given spatial frequency of perturbation. Although the exact frequency spectrum cannot be specified, we can make reach some broad conclusions as to its general form.

i) It will be difficult to avoid a component at zero spatial frequency, since this represents the 'average' level of perturbation.

ii) The spectrum will fall off rapidly at higher frequencies, because anything else would require rapid variations in the fibre with distance; (iii) the spectrum will not be constant, because the most common disturbances (temperature changes and vibration) are inherently time varying.

As we have mentioned before, even a single-moded fibre actually supports two modes. These have orthogonal polarizations, and are degenerate in circularly symmetric fibre. Consequently, the zero- spatial-frequency component above can phase-match them. Random coupling between the two then leads to a rotation of the polarization. This is a problem in interferometric systems, where the polarization of the interfering beams must be held parallel. Clearly, these require a different type of fibre, which holds the plane of polarization. Naturally enough, this is known as **polarization- preserving fibre**. Its improved performance is obtained by adopting a non-circularly-symmetric core geometry, which eliminates the degeneracy between the two polarization modes. Since these now have different propagation constants (say, \( \beta_{x} \) and \( \beta_{y} \)) they are no longer phase-matched by the zero- spatial- frequency component of any external perturbation. In fact, the larger the difference between \( \beta_{x} \) and \( \beta_{y} \) the better, since this will result in a larger phase-matching frequency (which will have a smaller component in the spectrum of a typical perturbation).
The effectiveness of polarization-preserving fibre is normally measured in terms of the **beat length** \( \Lambda \) between the two modes, defined as:

\[
\Lambda = \frac{2\pi}{|\beta_x - \beta_y|}
\]

Values of \( \Lambda \) in available fibre are generally of the order of millimetres.

It will be appreciated that a difference in propagation constant between two polarization modes is just another manifestation of the familiar phenomenon of birefringence. Polarization-preserving fibre is therefore also known as **high-birefringence** or hi-bi fibre. There are two basic methods of inducing the required effect; the first operates by modifying the physical symmetry of the fibre cross-section, while the latter works by removing the isotropy of the fibre material. In each case, preferred axes are imposed on the overall geometry, which become the polarization axes of the modified fibre.

To obtain **form-induced birefringence**, the core is made non-circular, usually by adopting an elliptical cross-section. The two preferred polarizations are then aligned parallel to the semi-major axes of the ellipse. So that these may readily be identified, a flat is often ground on the cladding, parallel to one axis (Figure 8.6-2a). This also improves the accessibility of the core, which is often advantageous (for example, in the directional coupler structures that will be discussed in Chapter 10). For obvious reasons, this type of fibre is known as **D-fibre**.

Alternatively, in **stress-induced birefringence**, the circular shape of the core is retained, but an anisotropic stress is applied to it. The most effective method uses the segmented construction shown in Figure 8.6-2b. Here an annular region of the cladding is divided into quadrants, with alternating segments made of glasses with different thermal expansion coefficients. Silica (with expansion coefficient \( \alpha_1 \)) is used for most of the cladding, while the shaded areas consist of B\(_2\)O\(_3\) doped silica (which has a higher coefficient \( \alpha_2 \)). During the cooling stages of the fibre fabrication process, differential contraction causes a large uniaxial stress to be applied to the core, which in turn induces birefringence in the core material (GeO\(_2\)-doped silica) through the **photoelastic effect**. This type of fibre is known as **bow-tie fibre** (from the shape of the stress-producing sectors) and offers beat lengths as short as 0.5 mm at \( \lambda_0 = 0.633 \) \( \mu \)m.

Figure 8.6-2 Polarization-preserving fibres: (a) D-fibre (photo courtesy R.Dyott, Andrew Corporation), and (b) bow-tie fibre (photo courtesy D.Payne, Southampton University).

### 8.7 COUPLING TO FIBRES

We will now examine some of the technology required to connect optical fibre and other guided wave components in complete systems. This is important, because it is only the
availability of low-loss, low-cost joints and splices that has allowed the full potential of optical fibre to be realised outside the laboratory. Two aspects are involved: the coupling of free-space beams into guided-wave components, and the direct connection of waveguides. Both may be tackled using a general method known as transverse coupling, and analysed using similar mathematical apparatus. We shall concentrate on the former in this section, leaving the latter to the next.

The technique of coupling free-space beams into guided modes by transverse excitation is known as end-fire coupling. For multimode fibres, the process is easy to visualise. In a ray model, each guided mode is defined by a characteristic angle of propagation relative to the fibre axis. Ignoring reflections, a field comprising an arbitrary collection of modes will then exit through an end-face of the fibre as a cone of light, whose apex angle is set by the numerical aperture defined earlier. Reversing this argument, we would expect that a cone of external radiation impinging on a fibre end-face will excite a random set of guided modes, which will then propagate down the fibre. The required external field can be generated very simply by a lens, leading to the typical arrangement shown in Figure 8.7-1.

![Figure 8.7-1 End-fire coupling into a multimode fibre.](image)

Notice that this argument contains no details of the proportions in which the modes are excited. However, we can deduce that it will be unproductive to use a beam whose numerical aperture is much greater than that of the fibre, because a considerable fraction of the input power must then be wasted on the excitation of cladding modes. Equally, we can guess that a low-NA input field will launch a greater proportion of low-order modes, which have shallower ray angles. A simple design approach would therefore be to match the numerical aperture of the input to that of the fibre. For example, fibre #2 in Table 8.3-2 has an NA of 0.17. Assuming that the parallel beam has a diameter of 1 mm, the required focal length of the lens is approximately $0.5/0.17 \approx 3$ mm.

Another feature ignored by the model above is the relative positions of the components. For example, if the fibre is placed a considerable distance from the focus of the lens, we would expect low coupling efficiency. To include this aspect, we must consider the excitation of individual modes rather more carefully. This can be done by examining the fields at the boundary between the free-space and waveguide systems in more detail, as in Figure 8.7-2. On the left, we have the input wave. Near the focus, this can be specified by a transverse field $E_{in}$ which travels almost entirely in the axial direction. Exactly at the focal plane, $E_{in}$ will be a Gaussian function if the input beam is itself Gaussian. The particular mode of interest (say, the $\nu$th mode, whose transverse field is $E_{\nu}$) is on the right. According to the analysis of Chapter 6, the excitation efficiency of the $\nu$th mode may then be found in terms of overlap integrals, as:

$$\eta = \frac{|\langle E_{in}, E_{\nu} \rangle|^2}{\langle E_{in}, E_{in} \rangle \langle E_{\nu}, E_{\nu} \rangle}$$

8.7-1
Consequently, the main feature determining $\eta$ is the overlap between $E_m$ and $E_\nu$. For high efficiency, the two fields should coincide spatially as far as possible. Given that the field diameter will be of the order of $\mu m$ in single-mode fibre, this implies stringent alignment tolerance. Furthermore, the fields should match each other closely in shape and size. We shall investigate this using an example.

![Figure 8.7-2 Geometry for modal analysis of end-fire coupling.](image)

**DESIGN EXAMPLE**

A typical end-fire coupling operation might involve the launching of light from a He-Ne laser into a single-moded parabolic-index fibre using a microscope objective. We can estimate the dependence of coupling efficiency on the focal spot size, as follows. Firstly, from the results of Chapter 4, we can assume that since this type of laser emits a Gaussian beam, the phase-front at the focus of the lens will be planar, with the diffraction-limited amplitude distribution $E_L(r) = A_L \exp(-r^2/w_0^2)$. Similarly, we may take the guided mode to have a Gaussian transverse field, given by $E_1(r) = A_1 \exp(-r^2/a^2)$. The fibre is assumed to be aligned exactly on-axis, with its end face at the focal plane. In this case, the overlap integrals are best performed in radial coordinates. If this is done, we obtain:

$$< E_L , E_L > = \int_0^{\infty} 2\pi r E_L^2(r) \, dr = \left(\frac{w_0^2 \pi}{2}\right) A_L^2$$

$$< E_1 , E_1 > = \int_0^{\infty} 2\pi r E_1^2(r) \, dr = \left(\frac{a^2 \pi}{2}\right) A_1^2$$

$$< E_L , E_1 > = \int_0^{\infty} 2\pi r E_L(r) E_1(r) \, dr = \left\{\frac{w_0^2 a^2 \pi}{w_0^2 + a^2}\right\} A_L A_1$$

Substitution into Equation 8.7-1 then yields:

$$\eta = \frac{4w_0^2 a^2}{(w_0^2 + a^2)^2}$$

8.7-2

By differentiation, it can be shown that maximum coupling efficiency is obtained when the fields are exactly matched in size, so that $w_0 = a$; simple substitution then shows that $\eta_{\text{max}} = 100\%$. However, $\eta$ falls rapidly away from this condition; for example, when $w_0 = 2a$, we obtain $\eta = 64\%$. Although this figure might still seem reasonable, it actually corresponds to an insertion loss of $\approx 2$ dB, equivalent to the attenuation incurred in propagating through about 10 km of low-loss fibre!

**COUPLING BETWEEN LASER DIODES AND OPTICAL FIBRE**

End-fire coupling via a lens may be used with other types of laser. For example, laser diodes (to be covered in Chapter 12) emit radiation from a narrow stripe, so their output is
highly divergent, with a numerical aperture that is normally much larger than that of single-mode fibre. However, a lens may be used to collect the output, and redirect it towards the fibre in a converging beam of reduced NA. This process is equivalent to the formation of a magnified image of the lasing stripe on the fibre end. In practise, highly compact optical arrangements are often used. Figure 8.7-3 shows three examples. In Figure 8.7-3a, a separate GRIN-rod lens is used between the laser diode (LD) and the fibre. In Figure 8.7-3b, a hemispherical glass microlens is attached to the end of the fibre, which has been tapered down to a reduced diameter. Finally, in Figure 8.7-3c, the end of a tapered fibre has been melted to form a lens directly in the fibre material.

![Figure 8.7-3a](image1)  
(a) Laser diode-to-fibre connection schemes: a) using a GRIN-rod lens, b) using a microlens attached to the fibre, and c) using a tapered, lensed fibre.

LENSED FIBRE CONNECTORS

The end-fire coupling principle can be used to construct a demountable connector known as a **lensed fibre connector**. Figure 8.7-4a shows the basic idea.

![Figure 8.7-4a](image2)  
(a) principle of lensed fibre connector; (b) and (c): connectors based on spherical microlenses and GRIN-rod lenses (after A.Nicia, Elect. Lett. 14, 511 (1978)).

Each half of the connector consists of one prepared fibre end and one lens, which has a numerical aperture matched to that of the fibre. The left-hand lens is used to collimate the output from fibre #1, while the right-hand lens focusses this beam down onto the input of...
fibre #2. Because the expansion-contraction process results in a near-plane beam of relatively large diameter passing between the two connector halves, errors on longitudinal alignment have virtually no effect, and the significance of errors in lateral position is drastically reduced. Naturally, extremely small and cheap lenses are required. Figures 8.7-4b and 8.7-4c show two possibilities; in the former, small spherical glass balls are used, while the latter is based in GRIN-rod lenses.

8.8 FIBRE INTERCONNECTS

We now consider the direct interconnection of fibre components. In this case, the transverse coupling procedure is known as butt coupling. The basic idea is very simple; the two components to be joined have their end-faces prepared accurately orthogonal to their axes of propagation. The axes are then aligned, and the components are simply butted together. Permanent attachment may then be achieved by gluing or by fusion splicing (to be described later). However, this process is naturally subject to error.

Figure 8.8-1 shows examples of the difficulties that may arise in connecting two fibres. The fibres could be displaced from their ideal relative positions, by (a) a short axial distance $\Delta z$, (b) a small transverse distance $\Delta x$, or (c) a rotational misalignment $\Delta \theta$. All such errors result in reduced coupling efficiency. The effects may be calculated in terms of the overlap between the field on the left-hand side of the joint and that on the right, using an expression similar to Equation 8.7-1.

![Figure 8.8-1](image)

Figure 8.8-1  Misalignment of a fibre joint in a) axial, b) lateral and c) rotational orientation.

Alternatively, they may be measured directly. For example, Figure 8.8-2 shows experimental measurements for the variation in coupling efficiency with off-axis misalignment $\Delta x$, for joints between identical single-mode fibres. Two sets of data are shown:

Fibre #1 has a diameter of 8.6 $\mu$m, a V-value 2.41 and a numerical aperture of 0.056, while Fibre #2 has $d = 4.3$ $\mu$m, $V = 2.18$ and $NA = 0.102$. In each case, efficiency falls drastically for offsets measured in $\mu$m, although the larger- diameter fibre is more tolerant. The only answer is stringent alignment accuracy.
Figure 8.8-2 Experimental variation of coupling efficiency with transverse displacement in butt-coupled joints (after W.A.Gambling et al., Elect. Lett. 14, 54-55 (1978)).

FIBRE END PREPARATION

The problems involved in achieving the necessary alignment accuracy are enormous. To begin with, the ends of the components to be joined must be smooth, flat and accurately orthogonal to their axis of propagation. Two basic methods are used to achieve this: polishing and cleaving. In the former, the fibre is mounted in a chuck, allowing one end to be held perpendicular to a rotating lap. The end-face is first ground flat, using an abrasive such as carborundum or aluminium oxide. Several grinding stages are needed, using abrasive of successively finer quality. Finally, the end is polished using diamond grit, to remove as far as possible the scratches induced by grinding. Alternatively, a chemical polish such as syton (colloidal silica) may be used. Although the procedure appears laborious, automatic machines exist that may polish a number of fibres simultaneously.

In the latter, a free end is first stripped of its protective jacket for a few inches, using a chemical stripper. It is then placed in a cleaving tool, as shown in Figure 8.8-3. This usually consists of four main components: two clamps, which hold the fibre at points a few centimetres apart, a curved anvil, which is used to tension the fibre, and a diamond scoring blade.

The procedure is to score the surface of the fibre with the diamond, inducing a crack. This is developed right through the cross-section, by moving the anvil so that the tension in the fibre is increased past a critical point. The tip of the crack travels at the speed of sound, and,
provided the tool is correctly adjusted, can leave a smooth, hackle-free surface orthogonal to the axis to within 0.5°. Greater control of this angle may be obtained by using a focussed ultrasonic wave to supply the energy needed to propagate the crack.

FUSION SPLICING

Permanent splices between identical fibres may be made by a fused joint, using either an electric arc or an oxyhydrogen flame. The process is known as **fusion splicing**, and the basic steps involved are illustrated in Figure 8.8-4. The two ends to be spliced are first cleaved, and clamped into the splicing machine (8.8-4a). This normally contains a set of precision piezo-electric micromanipulators, which allow the two ends to be aligned under a microscope (8.8-4b). To improve the alignment, the transmission through this interim butt-joint may be optimised. For example, light may be launched into fibre #1 by injection through a bend, and detected by radiation from a bend in fibre #2.

![Figure 8.8-4](image)

The position of the stages may then be adjusted under the control of a microprocessor, which is driven by a hill-climbing algorithm programmed to maximise the transmitted power. When the alignment is satisfactory, the two fibres are separated axially by a short distance (8.8-4c), and the electric arc is used to pre-fuse their tips. Surface tension causes each one to melt into an accurate hemisphere (8.8-4d), removing any irregularities caused by cleaving. The two fibres are then fed together, and the electric arc is used to perform the final fusion (Figure 8.8-4e). In this stage, surface tension forces once again perform a useful function, pulling the fibres into improved transverse alignment (8.8-4f).

Modern fusion splicing machines are semi-automatic, and routinely achieve splices with extremely low loss (= 0.1 dB); Figure 8.8-4g shows a typical commercial splicing machine. Splicers have also been developed for jointing fibre ribbon cable. This is a highly demanding task, requiring the fabrication of joints between closely-spaced fibres. Besides the constraints of space, each new fusion operation must not disturb the previous one.
After completion of the splice, the unjacketed fibre is rather fragile, and some kind of reinforcement is normally used. A typical reinforcement kit consists of short sections of meltable ethylene vinyl acetate (EVA) tube and steel rod, held in a length of polyethylene (PE) heat-shrink tubing. The components are placed over the splice and heated, so that the EVA melts onto the fibre. Simultaneously, the PE tubing is collapsed, to hold the steel rod firmly against the EVA-coated splice (Figure 8.8-5).

**MECHANICAL SPLICES**

For less demanding applications, mechanical splices are often used. In this case, the two fibres are placed in a precision assembly, which is secured with epoxy. Figure 8.8-6 shows one example, based on a preferentially-etched Si substrate. This has been lithographically patterned and processed in an anisotropic etch (such as ethylene diamine pyrocatechol, or EDP). Crystal planes with a higher density of atoms are more resistant to attack by such a mixture, so the etching follows these planes. For a suitable crystal orientation, the result is a pattern of deep etched grooves with sloping side walls (V-grooves), which act as a precision kinematic mount for cylindrical fibre, locating it accurately in the transverse plane. Two fibres are therefore self-aligned simply by placing them both in a groove and butting them together. The whole assembly is then glued together under pressure, or held mechanically if the splice is to be reusable. For mass-production, plastic replicas may be made of the etched substrates. In the first stage, a negative copy is made of the grooved
surface by nickel electroforming. This 'master' is then used to produce large numbers of
positive copies by precision moulding.

![Diagram of optical guided waves](image)

Figure 8.8-6 A self-aligned mechanical splice based on preferentially-etched silicon.

**DEMOUNTABLE FIBRE CONNECTORS**

Reusable splices have only limited lifetime, and proper demountable connectors are often
required. We have already described lensed connectors; **butt-jointed connectors** form the
other main type. In single-fibre connectors, each fibre end is normally encased in a
precision ferrule, which increases the effective diameter and allows accurate coaxial
alignment of the two ends by a mating sleeve. The ferrules may be cylindrical (in which
case a cylindrical sleeve is used), but some misalignment inevitably follows from the
clearance between the ferrules and the sleeve. Errors also follow from any lack of
concentricity between each ferrule and its fibre, but these may be reduced by using a core-
centering lathe for assembly.

Connectors of this type are suitable for multimode fibres. For higher precision, conical
ferrules are used. Insertion of the ferrule into a tapered sleeve then allows the virtual
elimination of machining tolerance, so that dust particles set the final limit on the transverse
alignment. This is the principle underlying the operation of the **biconical taper connector**,
which is often used for single-mode fibres. Figure 8.8-7 shows a number of different fibre
connector plugs.

Clearly, any approach involving discrete ferrules is impractical for the multiple joints
required in ribbon fibre connectors. Most often, these are based on etched silicon
substrates, using an extension of the principle previously described for mechanical splicing.
For, example, the two ribbon fibre ends may be mounted on precision-etched baseplates, so
that each fibre in the array is accurately fixed in a V-groove. On the male half of the
connector, two larger grooves are used to locate a pair of precision steel pins, which mate
with corresponding grooves on the female half. A joint may then be made simply by
aligning the pins and sliding the two connector halves together. Figure 8.8-8 shows a
typical multifibre connector based on similar principles.
8.9 Fibre-based components

We now consider a number of simple passive components constructed from optical fibre, for use in multimode and singlemode systems. In the former, devices are often based around GRIN-rod lenses and bulk-optic beamsplitters. Figure 8.9-1 shows two examples. The first is a **3 dB power splitter**, which divides an input beam equally between two outputs (Figure 8.9-1a). The beam enters from the left-hand fibre, and is collimated by a 1/4-pitch GRIN-rod lens. The resulting parallel beam is then passed into a frustrated-total-internal-reflection beamsplitter, of the type described in Chapter 7. This divides the beam equally into two components, which are end-fire coupled into output fibres using further 1/4-pitch GRIN-rod lenses. With the addition of a fourth unused port, the device is symmetric, and may act as a bidirectional 2 x 2 splitter.

The second component (Figure 8.9-1b) is a **demultiplexer**, used to separate or combine light of two different wavelengths in a wavelength-division-multiplexed communications
system. Just as in the previous example, a bulk-optic component (in this case, a multilayer filter, similar to the Bragg reflection gratings discussed in Chapter 7) is sandwiched between 1/4-pitch GRIN-rod lenses. However, only two lenses are used, and the input and output fibres are not located on-axis. Light of wavelength $\lambda_1$ is therefore collimated by the left-hand lens as an off-axis parallel beam. This particular wavelength is reflected by the filter, so the parallel beam is returned off-axis by an equal and opposite angle, and is then end-fire coupled into the left-hand output fibre. Light of a wavelength outside the reflective band of the filter (say, at $\lambda_2$) passes straight through, and is end-fire coupled into the right-hand output. Of course, the operation of the device could be reversed, so that two inputs at $\lambda_1$ and $\lambda_2$ are combined into a single output.

![Diagram](image-url)

**Figure 8.9-1** Multimode fibre components based on GRIN-rod lenses: (a) beamsplitter, and (b) wavelength demultiplexer (after S.Sugimoto et al., Elect. Lett. 14, 15-17 (1978)).

Single-mode fibre components are rather different. Because of the stringent alignment tolerances, end-fire coupling operations are inherently lossy; in fact, the best results are normally obtained by sticking to single-mode operation throughout. A component that is functionally equivalent to the 3 dB splitter above is known as a **fused tapered coupler**. To fabricate the device, two sections of single-mode fibre are first stripped of their protective jackets, placed parallel, twisted together and clamped by their free ends. Flame heating is then applied to the twisted section, so that the fibres gradually melt into one another. At the same time, the clamped ends are gradually drawn apart, so that the central section is tapered. As each fibre is thinned, it approaches cutoff, and light propagating in either core is spread further into the cladding.

![Flame-heating](image-url)

**Figure 8.9-2** Cross-section of a fused tapered coupler (photograph courtesy D.Mortimore, British Telecom Research Laboratories).
Since the fibres are melting together, the cores approach each other, so the modal fields in the two fibres start to overlap. There is then a possibility for power to be transferred from one fibre to the other. In this case, the mechanism is complicated, and we will simply assume that power exchange takes place. However, in Chapter 10 we will examine a similar device - the **directional coupler** - where a reasonable analytic explanation of the phenomenon can be given. The process is sufficiently controllable that it can be terminated when any desired splitting ratio is achieved. After encapsulation of the fragile central section, the device is complete. Figure 8.9-2 shows the cross-section of a typical coupler.

More complicated components may be constructed using similar techniques. **Polarization-preserving couplers** can be made by fusing together two high-birefringence fibres - for example, two D-fibres, arranged with their flat surfaces in contact. Similarly, larger numbers of fibres may be fused together, to make 1 x N or N x N power splitters (where N might lie in the range 3 - 7). When N becomes much larger, the monolithic approach is impractical, and it is simpler to synthesise the required structure by splicing together a number of smaller devices in an array. For example, Figure 8.9-3 shows an 8 x 8 **star coupler**, built up from twelve 2 x 2 fused tapered couplers. Note that there is exactly one pathway from each input to each output. Furthermore, all paths pass through the same number of 2 x 2 couplers (three), and have the same transmission through each (50%). Consequently, light incident on any of the eight input ports is equally divided among all the outputs.

![8 x 8 star coupler](image)

Figure 8.9-3 An 8 x 8 star coupler, built using fused tapered components.

We now introduce an entirely different single-mode fibre device: a **polarization controller**. We have already noted the phenomenon of polarization fluctuation in circularly-symmetric fibre. In many situations, it is necessary to adjust the output polarization from a fibre, before coupling into another component (which may be polarization-sensitive). Now, in bulk optics, it can be shown that a plane wave of arbitrary polarization may be transformed into any desired output polarization state using two quarter-wave plates. We recall from Chapter 3 that these are slabs of birefringent material, with a thickness chosen to give a $\pi/2$ phase difference between the two orthogonal polarization components. An in-line single-mode polarization controller can perform a similar trick, using the fibre-optic equivalent of a quarter-wave plate.

Figure 8.9-4a shows the principle. The fibre is formed into a coil, of radius $R$ and length $L$. If $R$ is sufficiently small, bending the fibre in this way results in a uniaxial stress in the core, acting towards the centre of the loop. As we saw in bow-tie fibre, this causes stress-induced birefringence, and $L$ is simply chosen so that a $\pi/2$ phase shift is obtained between the two polarization modes in travelling round the loop. It is also necessary to be able to
achieve an effect analogous to the rotation of a wave plate. This is done by fixing the input and output of the coil (points A' and B'). The entire coil is then rotated about the line A' - B', twisting the sections A' - A and B - B'. This can induce the necessary rotation of the plane of polarization with respect to the principal axes of the coil. Two coils are needed in a complete polarization controller, which is then as shown in Figure 8.9-4b.

Figure 8.9-4 Single-mode fibre fractional wave devices: a) basic principle, and b) polarization controller.
PROBLEMS

8.1. Attenuation in silica glass is approximately 4 dB/km at \( \lambda_0 = 0.65 \) \( \mu \)m. What is the fraction of power transmitted through 10 km of silica-based fibre? Assuming that the main contribution to loss arises from Rayleigh scattering, estimate the attenuation at \( \lambda_0 = 1.3 \) \( \mu \)m. What is the fraction of power transmitted through the same length of fibre at this wavelength?

\[0.01\%; 0.25 \text{ dB/km}; 56.2\%\]

8.2. A multimode fibre, with a numerical aperture of 0.2, supports approximately 1000 modes at a wavelength of 0.85 \( \mu \)m. Estimate its diameter.

60.5 \( \mu \)m

8.3. Using (i) a one-dimensional ray-optic argument, and (ii) a modal argument, explain the formation of an inverted real image by a half-pitch GRIN-rod lens.

8.4. Show that the reflection coefficients for TE- and TM-polarized light at a dielectric interface both tend to exp(j\( \pi /2 \)), when the refractive index step tends to zero and the angle of incidence approaches 90°.

8.5. Calculate the half-width of the lowest-order mode of a parabolic-index fibre, assuming that the peak refractive index of the fibre is \( n_p = 1.5 \), the characteristic radius is \( r_0 = 25 \) \( \mu \)m and the wavelength is \( \lambda_0 = 0.85 \) \( \mu \)m. What will be the half-angle of divergence of the far-field radiation pattern of this mode? (Hint: refer to Equation 4.6-18.)

2.1 \( \mu \)m; 7.4°

8.6. The waveguide equation for an idealised model of a parabolic-index fibre may be taken as:

\[ \nabla^2_{xy} E_{\mu,\nu}(x, y) + \{ n_0^2 k_0^2 [1 - (r/r_0)^2] - \beta_{\mu,\nu}^2 \} E_{\mu,\nu}(x, y) = 0. \]

Show by direct substitution that permissible solutions for the transverse fields of higher-order guided modes are given by:

\[ E_{\mu,\nu}(x, y) = H_\mu(\sqrt{2}x/a) H_\nu(\sqrt{2}y/a) \exp\{-(x^2 + y^2)/a^2\} \]

where \( H_\mu(\zeta) \) is a Hermite polynomial of order \( \mu \) and \( a = \sqrt{2r_0/n_0 k_0} \), and that the corresponding propagation constants \( \beta_{\mu,\nu} \) are given by

\[ \beta_{\mu,\nu} = n_0 k_0 \sqrt{1 - 2(\mu + \nu + 1)/ (n_0 k_0 r_0)}. \]

8.7. Show that the first few low-order Hermite polynomials given in Table 8.4-1 do indeed satisfy the governing differential equation:

\[ d^2 H_\mu(\zeta)/d\zeta^2 - 2\zeta dH_\mu(\zeta)/d\zeta + 2\mu H_\mu(\zeta) = 0. \]

Sketch the variation of the transverse fields

\[ E_{\mu,\nu}(x, y) = H_\mu(\sqrt{2}x/a) H_\nu(\sqrt{2}y/a) \exp\{-(x^2 + y^2)/a^2\} \]

along the line \( y = 0 \). Under what conditions will the solution prove inaccurate?
8.8. What properties are the most desirable in optical fibre that is to be used in a long-distance telecommunications system? You should discuss loss, dispersion, ease of jointing, and the availability of other systems components such as sources and detectors.

8.9. The theoretical expression for the end-fire coupling efficiency between a Gaussian beam of characteristic radius \( w_0 \) and the lowest-order mode of a parabolic-index fibre (which also has a Gaussian transverse field, of radius \( a \)) is given by:

\[
\eta = \frac{4w_0^2a^2}{\left(w_0^2 + a^2\right)^2}.
\]

Sketch the variation of \( \eta \) with the parameter \( \alpha = \log(w_0/a) \). Prove that the variation is symmetric about the point \( \alpha = 0 \), when maximum efficiency of 100% is obtained.

8.10. Sketch the layout of a 16 x 16 star coupler, based on an array of 2 x 2 fused tapered couplers.
SUGGESTIONS FOR FURTHER READING

Koehler B.G., Bowers J.E. "In-line single-mode fiber polarization controllers at 1.55, 1.3 and 0.63 \( \mu \)m" Appl. Opt. 24, 349-353 (1985)
Miller C.M. "Fibre-optic array splicing with etched silicon chips" Bell Syst. Tech. J. 57, 75-90 (1978)
Miller T.J., Hart A.C., Vroom W.J., Bowden M.J. "Silicone- and ethylene-vinyl-acetate-coated laser-drawn silica fibres with tensile strengths > 3.5 GN/m2 (500 kp.s.i.) in > 3 km lengths" Elect. Lett. 14, 603-605 (1978)
Miya T., Terunuma Y., Hosaka T., Miyashita T. "Ultimate low-loss single-mode fibre at 1.55 \( \mu \)m" Elect. Lett. 15, 106-108 (1979)
Mortimore D.B. "Low-loss 8 x 8 single-mode star coupler" Elect. Lett. 21, 502-504 (1985)
Payne D.B., McCartney D.J., Healey P. "Fusion splicing of a 31.6 km monomode optical fibre system" Elect. Lett. 18, 82-84 (1982)
CHAPTER NINE

CHANNEL WAVEGUIDE INTEGRATED OPTICS

9.1 CHANNEL WAVEGUIDE TYPES

The second important system of integrated optics arises when the waveguides take the form of channel guides. These still lie on a planar substrate, but are now arranged to confine the optical field in two directions, x and y. Figure 9.1-1 shows cross sections through the most common geometries. We shall now describe these in some detail, together with a summary of typical fabrication methods; this information was omitted from our previous discussion of planar guides in Chapter 6, on the grounds that many of the structures used are common to each waveguide system. Further processing information can be found in Chapter 13.

![Channel waveguide types: a) buried, b) ridge and c) strip-loaded.](image)

1) BURIED CHANNEL GUIDES

Figure 9.1-1a shows a buried channel guide, which is made by modifying the properties of the substrate material so that a higher refractive index is obtained locally. Most fabrication processes result in a weak, graded-index guide buried just below the surface, and, although channel guides normally have one axis of symmetry, they usually lack a well-defined cross-sectional shape.

Diffusion is often used to fabricate guides of this type. For example, titanium metal can be diffused into lithium niobate or lithium tantalate substrates, by first depositing the metal in patterned strips of \( \approx 1000 \) Å thickness and then carrying out an in-diffusion at a high temperature (\( \approx 1000 \)°C) for several hours (3 to 9). This is known as the Ti: LiNbO\(_3\) process. The additional impurities cause a change in refractive index that is approximately proportional to their concentration, with a typical maximum value of \( \Delta n \approx 0.01 \). Figure 9.1-2 shows contours of constant refractive index that are obtained by diffusion of Ti metal into LiNbO\(_3\).

Alternatively, material can be exchanged with the substrate. For example, protons (H\(^+\) ions) can be exchanged with Li\(^+\) ions in LiNbO\(_3\). This is done by first covering the substrate with a masking layer, which is patterned so narrow stripe openings are made. The crystal is then placed in a suitable hot melt (in this case, benzoic acid, heated to \( \approx 200 \)°C), and exchange of material takes place through the crystal surface in regions exposed by the mask openings. Proton exchange yields an index change of \( \Delta n = +0.12 \) for one polarization mode (the TM mode), but the index change is actually negative for the other one. Consequently, a waveguide is only formed in the former case. Silver ions can be exchanged with Na\(^+\) ions in soda-lime glass in a similar way, by immersing the substrate in molten AgNO\(_3\), at temperatures of 200 - 350°C. In this case, the time required to form a guide is large (several
hours), but if a DC electric field is used to assist the ion migration, the process can be speeded up quite considerably. The index change obtained is \( \Delta n \approx 0.1 \).

![Figure 9.1-2 Contours of constant refractive index in a Ti-diffused channel waveguide.](image)

Finally, material can be implanted using an ion implanter. This is a large, expensive item of vacuum equipment, which can be used to select a particular species of ions and then accelerate the desired ions electrostatically, so that they strike the substrate surface at high speed and penetrate some distance. Because this process is ballistic, it causes damage to the lattice as well as a change in refractive index. Ion implantation is therefore normally followed by an annealing step, which shakes out the lattice damage. This allows properties that rely on a regular, undamaged crystal lattice (e.g., electro-optic behaviour) to be preserved, and greatly reduces waveguide propagation losses. Following annealing, He\(^+\) implantation has been found to be reasonably successful in LiNbO\(_3\).

2) RIDGE WAVEGUIDES

Figure 9.1-1b shows a ridge waveguide. This is a step-index structure, usually fabricated from a planar guide that has been patterned and etched to leave a ridge. Total internal reflection then takes place at the sides of the guide, as well as the top and bottom faces. Often, the etching is designed to follow specific crystal planes, so the edges of the ridge may be sloped. Similarly, material is often (especially in laser structures) regrown around the sides of the guide. Much higher field confinement is possible using a ridge than a buried guide. As mentioned above, the index change that can be obtained by material modification is very small, so the evanescent field of any guided mode must extend for some distance outside the core. In a ridge guide, on the other hand, there can be a large index difference between the ridge and its surround. This is an important advantage; for example, bends can be much tighter in ridge-guide systems, as we shall see later.

Waveguides are often made as ridges in semiconductor materials, because of the ease with which a mesa structure may be made by etching. Although the actual characteristics of semiconductors will not be covered until Chapter 11, we shall now describe the basic semiconductor waveguide types. There are two: heterostructure guides, which operate by the refractive index differences obtained between different materials, and homostructure guides, which use the index reduction following from an increase in the majority carrier density.

The two most common semiconductor materials used in optoelectronics are based on the ternary alloy gallium aluminium arsenide and the quaternary alloy indium gallium arsenide phosphide, respectively. To fabricate heterostructure waveguides in the GaAs/GaAlAs system, successive layers of material are first grown on a GaAs substrate, in differing compositions of the general form Ga\(_{1-x}\)Al\(_x\)As (where x is the mole fraction of the constituent). Because GaAs and GaAlAs have similar lattice parameters, these additional layers are lattice-matched to the substrate and the composite crystal can grow without strain.
Now, in Ga$_{1-x}$Al$_x$As, the refractive index ranges roughly linearly from $n \approx 3.57$ for $x = 0$ to $n \approx 3.36$ for $x = 0.35$ (at near infrared wavelengths). The variation of the alloy content between the layers then yields the refractive index difference necessary for a waveguide. For example, Figure 9.1-3 shows how a planar guide might be fabricated using a layer of high-index GaAs, grown epitaxially on top of a thick layer of Ga$_{1-x}$Al$_x$As (which acts as a low-index isolation layer, separating the guide from the high-index GaAs substrate). The guiding layer may then be etched down to a narrow rib, to form a ridge guide.

Waveguide fabrication is similar in the InP/InGaAsP system. This time, layers of the quaternary alloy In$_{1-x}$Ga$_x$As$_{1-y}$P$_y$ are grown on an InP substrate. However, lattice-matching to the substrate is only obtained for particular combinations of $x$ and $y$, which are detailed in Chapter 13. More complicated layered structures are also often used. For example, Figure 9.1-4 shows a double heterostructure waveguide, which has a low-index GaAlAs confining layer on both sides of the guide. This arrangement is used extensively in semiconductor lasers, since the difference in bandgap between the GaAs and GaAlAs layers can provide carrier confinement in addition to confinement of the optical field. It will be discussed in detail in Chapter 12.

Instead of using compositional variations to form a layered guiding structure, waveguides can also be formed by making use of the reduction in refractive index that follows from an increase in the free carrier concentration. As we showed in Chapter 3, this follows the general law:

$$\Delta n = -N\lambda_0^2e^2 / (8\pi^2\varepsilon_0\hbar m^*c^2)$$

9.1-1
where $N$ is the carrier density, and $m^*$ is their effective mass. For example, undoped GaAs may be grown on top of a doped (n-type) substrate to form a planar homostructure guide, as in Figure 9.1-5.

![Figure 9.1-5 A planar waveguide in GaAs formed by changes in free carrier concentration.](image)

Alternatively, a similar structure may be created by proton bombardment of an n-type substrate. This creates damage sites near the surface, which trap free carriers. This is known as free carrier compensation, and produces an increase in refractive index. Now, we showed in Chapter 3 that the index change obtained by doping may be found for n-type GaAs as $\Delta n = -0.01$, when $N = 5 \times 10^{18} \text{ cm}^{-3}$ and $\lambda_0 = 1 \mu\text{m}$. This implies that the index changes possible in homostructures are much lower than in heterostructures, giving reduced confinement in any guide structure. In fact, since $m_{e}^* \gg m_{h}^*$ in InP, the hole-induced index difference is much smaller than the equivalent electron-induced value, so only n-type material is usable for waveguiding. A further problem arising in homostructures is that the increased carrier concentration in the substrate causes a rise in absorption, which can result in unacceptable propagation loss. Heterostructures therefore offer several advantages.

Ridge guides can also be made in amorphous material on semiconducting substrates. One common system involves silicon oxynitride guides, constructed on a silicon dioxide buffer layer, which in turn lies on a silicon substrate. Silicon oxynitride has a relatively high refractive index, but one which is lower than that of the silicon, for which $n \approx 3.5$. The two are therefore separated by a thick layer of SiO$_2$, which acts as a low-index ($n \approx 1.47$) spacer. Other systems use doped silica (e.g. an SiO$_2$/TiO$_2$ mixture) on top of the silica buffer layer. These are important, since they allow the fabrication of waveguides in a form compatible with VLSI electronics.

3) STRIP-LOADED WAVEGUIDES

Two-dimensional confinement may also be less obvious. Figure 9.1-1c shows a strip-loaded guide. Here, a three-layer structure has been used, which has a substrate, a planar layer, and then a ridge. The planar layer is arranged as a guide, but one that is just cut off. However, the addition of a further high-index overlay is sufficient to induce guiding in a localised region near the ridge. Confinement of the optical field in the y-direction is obtained because the refractive index, when averaged in the x-direction, is higher in the region of the ridge. The additional ridge overlay may have a different refractive index to that of the planar guide, or (as in Figure 9.1-1c) the same index.

ANALYSIS OF CHANNEL WAVEGUIDES

The analysis of a channel guide is generally a complicated problem, and different approaches are used for ridge and channel guides. Since the index changes concerned in the former are so large, a TEM model cannot be used, and a rigorous solution is normally required. Figure 9.1-6 shows one possible model of a ridge guide, which consists of a
region of dielectric of rectangular cross-section surrounded by regions of differing refractive indices.

![Model of a rectangular dielectric waveguide.](image)

Figure 9.1-6  Model of a rectangular dielectric waveguide.

It has been shown that modes of such a guide can be found (at least approximately) by assuming solutions for the transverse variation of the electric field in the form of a product, i.e. as $E_t(x, y) = E_1(x) E_2(y)$. Since the layers are individually uniform, we may guess that $E_1(x)$ and $E_2(y)$ must have the form either of cosinusoidal or of exponential functions, by analogy with the solution for a slab guide. These solutions are then built up in a piecewise fashion, as follows:

- In Region 1, the field is assumed to vary cosinusoidally with both $x$ and $y$
- In Region 2, the field is assumed to decay exponentially with $x$, and vary cosinusoidally with $y$
- In Region 3, the field is assumed to vary cosinusoidally with $x$, and decay exponentially with $y$

And so on. These fields are matched rigorously at the interfaces between Regions 1 and 2, 1 and 3, 1 and 4, and 1 and 5, but only approximately in the shaded corner regions. The result is a set of two-dimensional mode patterns, which are neither truly TE nor truly TM. Often, the confinement provided by the layering is much stronger than that in the lateral direction, so this type of mode may be non-circular, with a high aspect ratio (e.g. 3 : 1).

Unfortunately, the index distribution obtained by other fabrication methods can be less well defined. For example, we have already seen in Figure 9.1-2 the refractive index distribution arising when a strip of Ti metal is diffused into a LiNbO$_3$ crystal. Since the index change is weaker in this case, the waveguide modes are TEM-like and the approximate model (Equation 6.4-3) developed in Chapter 6 may be used to find them. We shall not attempt to solve this equation here, but simply present the results. Figure 9.1-7 shows contours of constant intensity for the transverse field $E(x, y)$ in a single-mode guide, with the index distribution of Figure 9.1-2.

![Isointensity contours for the modal field in a Ti-diffused channel guide.](image)

Figure 9.1-7  Isointensity contours for the modal field in a Ti-diffused channel guide.
Notice that the mode is roughly circular, but there is increased confinement at the surface, due to the large index change at the substrate/air interface.

9.2 INPUT AND OUTPUT COUPLING

As with optical fibres, power is generally coupled into a channel guide by transverse excitation, using end-fire or butt coupling. For example, Figure 9.2-1 shows the experimental set-up for end-fire coupling into and out of a Ti:LiNbO₃ channel waveguide device. The Ti:LiNbO₃ chip is contained in the package at the centre of the photograph, and consists of an array of six directional coupler switches. A x10 microscope objective to the left of the device is used to focus the output from a 1.523 µm wavelength He-Ne laser onto a polished end-face of the chip, and micromanipulators are used to align one of the device inputs to the focused spot of light. A second microscope objective is used to couple light out of the chip, forming real images of the two device outputs at a convenient nearby plane. Detectors may be placed at this plane to measure the device response.

![Figure 9.2-1 Apparatus for end-fire coupling into a Ti:LiNbO₃ channel waveguide device.](image)

More usefully, a channel guide device can be joined to a different type of guide (e.g. a fibre) simply by polishing an end of each component, aligning the field-carrying cores together, and forming a permanent cemented joint. There are two requirements for high efficiency in butt coupling, as we saw in Chapters 6 and 8.

1. The modal field in the channel guide must be matched (in overall shape and dimensions) as far as possible with that in the fibre. Perfect mode matching is never possible, but considerable success has been achieved in many cases by careful choice of the fabrication conditions. For Ti:LiNbO₃ devices, for example, this may be done by adjusting the width and thickness of the Ti metal strip, and the diffusion temperature and time.

2. The field-carrying cores must be accurately aligned with respect to one another, in relative spatial position and angular orientation. Typical tolerances for the former are < 1 µm. For single joints, this can be achieved by micromanipulation, using piezoelectric positioners.

If conditions (1) and (2) are satisfied, the fibre-device-fibre coupling loss can be as low as 1.5 dB in Ti:LiNbO₃ devices, at near infrared wavelengths.

Unfortunately, micromanipulation is not a feasible method of adjusting the relative positions of closely-spaced arrays of fibres and channel guides. An alternative self-
alignment technique is therefore preferred. Figure 9.2-2 shows one method, used with Ti:LiNbO$_3$ devices.

![Diagram of precision alignment structure](image)

Figure 9.2-2 Precision alignment structure for multiple fibre end-fire coupling to an integrated optic channel waveguide chip (after E.J.Murphy et al., IEEE J. Quant. Elect. QE-22, 928-932 (1986)).

This is based on the V-grooved silicon substrates used in the fibre connectors and splices discussed in Chapter 8. The fibre array is simply sandwiched between two grooved substrates, and the end of the entire assembly is polished. It is then only necessary to position each array correctly, instead of each individual fibre. In the example above, one half of each fibre-carrying sandwich is extended as a ledge, which mates with the integrated optic chip. Only one adjustment (parallel to the substrate surface) is then needed to align each fibre array with the channel guides on the chip.

### 9.3 SOURCES OF PROPAGATION LOSS

#### FUNDAMENTAL LOSS MECHANISMS

The main causes of propagation loss in straight waveguides are absorption, and scattering from material inhomogeneities and boundary imperfections. Note that the latter is greatly reduced in fibre manufacture by the process of pulling a preform down to the final fibre. This 'smoothing' does not occur in the fabrication of an integrated device, which is produced as a 1 : 1 copy of a master mask set. Consequently, scattering losses are much higher in integrated optics. However, all forms of loss vary considerably, depending on the precise nature of the guide. We shall illustrate this with examples.

For Ti:LiNbO$_3$ buried channel guides, losses are typically 1.0 dB/cm @ $\lambda = 0.633$ $\mu$m, 0.5 dB/cm @ $\lambda = 1.15$ $\mu$m and 0.2 dB/cm @ $\lambda = 1.52$ $\mu$m. In this case, the major contribution to loss is the photorefractive effect. This is a form of optical damage, which results in the optically induced formation of refractive index changes. These scatter light out of the guide, causing increased loss. Photorefractive effects are very dependent on the level of optical power in the guide. They are extremely significant at visible wavelengths, but reduce enormously in the near infrared.
For GaAs ridge guides, losses were originally extremely high (≈ 4.0 dB/cm), but these have been reduced considerably in recent years, to ≈ 0.2 dB/cm. Much of the loss in homostructures arises from the presence of free carriers in the low-index confining layers outside the guide, which gives rise to loss through collision damping. For low loss, a mode must therefore be well confined inside the more resistive guiding layer. However, further loss arises from scattering from the guide walls. If these are etched down crystal planes, they may well adopt a ‘staircase’ profile, causing significant scattering.

**BEND LOSS**

A common requirement in a guided wave system is to be able to route light round a bend, for example to establish links between different devices on an integrated optic chip. However, as mentioned in Chapter 8, any curvature of a waveguide results in a loss additional to the normal propagation loss. Figure 9.3-1 illustrates the effect, which is caused by continuous conversion of the guided light into radiation modes. We shall now give an alternative physical explanation of the process.

One way to understand the underlying cause of the radiation is to consider the profile of the transverse field as the guided mode travels round the bend. To keep in step, this pattern must rotate around the centre of curvature of the guide, like a spoke in a rotating wheel. Consequently, the further from the centre, the faster the field must move. However, the outermost evanescent part of the field extends to infinite radius, so at some distance from the core the field would have to exceed the speed of light in the substrate material to keep up. The portion of the field outside this point must be radiated, reducing the power in the guided mode. Bend loss is therefore very dependent on the confinement offered by the guide. For indiffused guides, the index change forming the guide is small, so the confinement is poor. Consequently, if low loss is required, bends must be shallow, with typical radii of curvature of greater than 1 cm for Ti:LiNbO₃ devices. This limits the device packing density considerably. In ridge guide systems, the index change is larger, so bends can be considerably tighter.

A detailed analysis of the conversion of guided light into radiation modes in a bend can be performed using the method of local normal modes, which we previously used to investigate slow tapers. This time, however, the axis of propagation of the guide must change with distance around the bend, so the analysis is rather more complicated. An approximation to the coupled mode solution is therefore often used. Typically, it is found that the guided power falls off as:
where \( P_z \) is the power after travelling a distance \( z \) and \( P_0 \) is the input power. The attenuation coefficient \( \alpha \) has the semi-empirical form:

\[
\alpha = c_1 \exp(-c_2 r)
\]

where \( r \) is the radius of curvature and \( c_1 \) and \( c_2 \) are constants. Equation 9.3-1 can be interpreted as saying that the rate of power loss is proportional to the local guided power. This is reasonable, provided that no radiated power is subsequently coupled back into the guide. Since this process can and does occur, Equation 9.3-1 should be used with some caution.

Note that, for a bend of a given angle \( \theta \), the transmitted power is \( P_\theta = P_0 \exp\{-\alpha r \theta\} \). For bends of large radius, \( P_\theta \) tends to unity, since the loss coefficient \( \alpha \) tends to zero. However, for extremely small radius bends, \( P_\theta \) tends to unity once again. This is because the distance travelled round the bend now becomes very small, which counters the effect of an increasingly large value of \( \alpha \). In between, there is a minimum in \( P_\theta \). By differentiation, this can be shown to occur at a critical radius \( r = 1/c_2 \).

Bends are often used to provide offset transitions at the input and output of devices, or to separate guides at the edge of a chip for ease of coupling to fibres. In either case, an S-bend is used. This can take the form of an abrupt transition, as shown in Figure 9.3-2a, or a smooth curve (Figure 9.3-2b). Generally the former works better if the offset is small, and the latter otherwise.

\[ P_z = P_0 \exp(-\alpha z) \]  

9.4 POLARIZERS

We now consider a range of passive channel waveguide devices, which can be used as building blocks in a more complicated guided wave optical circuit. We begin with
polarizers. In bulk optics, these are devices that will pass one polarization component of a plane wave, while attenuating the other (preferably blocking it entirely). An integrated optic polarizer is a similar component, which will pass only one polarization mode. There are a number of ways to achieve this. The major strategies are:

(1) To radiate one polarization mode preferentially. This can be done by adding a birefringent overlayer to the guide, so that one polarization mode is cut off while the other still propagates. For example, Figure 9.4-1 shows an early polarizer for a waveguide formed in a glass substrate. This consisted of an overlayer of calcite, for which the ordinary and extraordinary indices are 1.658 and 1.486, respectively. Since the guide index is approximately 1.5, only the TE mode is guided with the crystal arranged in the orientation shown, and the TM mode is radiated. Although this approach produces a functionally adequate polarizer, it is not suitable for integration. More recently, fully integrated processes have been developed. One example combines short sections of proton-exchanged waveguide with guides formed by the Ti:LiNbO₃ process. This produces a TM-pass polarizer, because proton exchange only forms a waveguide for the TM mode.

(2a) To absorb one polarization mode preferentially. This can be done using a metal overlay, in conjunction with a dielectric. For Ti:LiNbO₃ guides, an Si₃N₄/Al combination is suitable. Strong absorption of one mode can then occur if it is coupled into a surface plasmon (previously encountered in Chapter 5) propagating in the metal layer. (2b) To couple one polarization mode preferentially into another guide. This can be performed using a polarization-selective directional coupler. Both of these involve the phenomenon of codirectional coupling, which will be covered in Chapter 10.

Figure 9.4-1  Waveguide polarizer using a calcite overlay (after S.Uehara et al., Appl. Opt. 13, 1753-1754 (1974)).

9.5 MIRRORS

We now consider the integrated optic equivalent of a mirror, which offers a simple method of changing the direction of a guided beam. It can occupy considerably less space than a bend, and can theoretically be more efficient. Such a device can operate by the reflection of guided modes from a flat, highly reflective interface, arranged exactly orthogonal to the guide. A suitable interface can be made by cutting a trench across the guide, as shown in Figure 9.5-1. This can be done using a highly anisotropic etch process - reactive ion etching (RIE) and focussed ion beam micro-machining are both suitable. The interface must then be metallised to increase its reflectivity.
While the integrated mirror is conceptually a simple device, it is hard to make in practise, because any reduction in surface flatness, orthogonality or reflectivity will lower the overall reflection efficiency, as we now show. This can be calculated by modelling the local mirror reflectivity as the product of a real function $R(x, y)$, which accounts for the amplitude reflectivity, and a phase function $\phi(x, y)$, which accounts for mirror tilt and flatness. An incident mode with transverse field distribution $E_\nu(x, y)$ will therefore be reflected as a new field, with transverse field distribution $E_\nu'(x, y)$, given by:

$$E_\nu'(x, y) = E_\nu(x, y) R(x, y) \exp{-j\phi(x, y)}$$  \hspace{1cm} 9.5-1

The problem can then be treated as one of end-fire coupling, where we simply wish to know the efficiency $\eta_r$ with which the new field $E_\nu'$ will excite the old field $E_\nu$. From Chapter 6, we know that this will be given by:

$$\eta_r = \frac{|\langle E_\nu', E_\nu \rangle|^2}{\langle E_\nu', E_\nu \rangle \langle E_\nu, E_\nu \rangle}$$  \hspace{1cm} 9.5-2

The important feature in Equation 9.5-2 is (as usual) the overlap integral between $E_\nu'$ and $E_\nu$. The reflection efficiency is clearly 100% if $E_\nu' = E_\nu$. However, it is reduced if there are significant differences between $E_\nu'$ and $E_\nu$.

**DESIGN EXAMPLE**

We may use this approach to estimate the tolerance on mirror tilt angle, assuming that the mirror is tilted about the y-axis through an angle $\theta$ from the vertical plane as shown in Figure 9.5-2. We shall evaluate the reflection efficiency $\eta_r$ using fields defined on the plane $z = 0$. We start by assuming that the phase-front of the incident mode is flat, so that $E_\nu(x, y)$ is real at $z = 0$. To calculate the reflected field $E_\nu'$, we simply assume that each part of the mode must travel a different optical distance (which depends on the transverse co-ordinate $x$) in moving from $z = 0$ to the mirror, and back again.

$E_\nu'$ is therefore given approximately by:

$$E_\nu'(x, y) = E_\nu(x, y) \exp{-j2\beta x \tan(\theta)}$$  \hspace{1cm} 9.5-3

where the exponential term accounts for the additional phase-shift accumulated in travelling the extra path.
The phase function $\phi$ in Equation 9.5-1 therefore has the form:

$$\phi(x, y) = 2\beta x \tan(\theta)$$

so the functional description of the tilted mirror is particularly simple: it imparts a linearly-varying phase-shift across the guided beam. We may now estimate that $\eta_r$ will fall to zero when there is a phase shift of $\pi$ radians across the beam, as the overlap integral in the numerator of Equation 9.5-2 will then average to zero. For a waveguide of depth $h$, this implies that:

$$2\beta h \tan(\theta) \approx 2\beta h \theta \approx \pi$$

Or:

$$\theta \approx \frac{\lambda_0}{4n_{\text{eff}} h}$$

Using the typical parameters of $\lambda_0 = 1 \mu m$, $h = 5 \mu m$ and $n_{\text{eff}} = 2.2$, we get $\theta \approx 1.3^\circ$. Since $\theta$ must be considerably lower than this for good reflectivity, the allowable tilt angle is extremely small.

REFLECTIVE BENDS AND BRANCHES

Mirrors can also be used to turn a beam through ninety degrees, for example by making an interface at $45^\circ$ at an intersection between two orthogonal ridge guides, as shown in Figure 9.5-3a. The device is similar to the direct reflector, but there is no need for coating to increase reflectivity - the interface itself is normally totally-reflecting, due to the large index difference between the ridge guide material and its surround (air). This particular technique is important, because it allows the construction of bendless integrated optical circuits.

Figures 9.5-3b and 9.5-3c show examples of reflective waveguide bends, formed using GaAs and high-silica waveguides, respectively.
CHAPTER NINE

Figure 9.5-3  a) Total-internal reflection mirror; b) and c) Reflective bends in GaAs ridge guides and high silica ridge guides, respectively (photographs courtesy P. Buchmann, IBM Research Division, Zurich, and A. Himeno, NTT); d) frustrated total-internal-reflection beamsplitter.

Figure 9.5-3d shows a related device, formed from two 90° reflectors separated by a small air-gap. This is made using a narrow etched trench, so that the light can cross the air-gap by optical tunnelling. The device then acts as a frustrated-total-internal-reflection beamsplitter, similar to the bulk-optic FTIR device described in Chapter 7, and the splitting ratio is controlled by adjusting the gap width.

9.6 TAPERS AND Y-JUNCTIONS

We now consider devices that can change the cross-section of a guided beam, by a gradual tapering of the guide parameters. The first such device is the waveguide horn, shown in Figure 9.6-1. This was important in the early days of integrated optics, since it was used as a conversion element between single-moded guides and much wider, multimode guides. Since the development of all single-mode integrated optics, it has been rather less in demand.

A horn operates by a slow expansion of the guide cross-section. As we saw in Chapter 6, the theory of local normal modes implies that, if this process is slow enough, the input mode will gradually evolve into the lowest-order, symmetric mode of the multimoded output guide, without mode conversion. Generally, the allowed rate of expansion at any point along the taper depends on the number of guided modes supported at that point. At the left-hand end, the number of modes is small, so the expansion rate can be relatively fast.
At the right hand end, the number is large, so the rate must be much slower. This leads to the characteristic parabolic variation in cross-section shown in Figure 9.6-1.

The second such device, the **Y-junction**, is the simplest kind of branching element. It consists of a single-mode input guide, which is gradually separated by a forked transition region into two similar single-mode output guides as shown in Figure 9.6-2.

![Figure 9.6-1 A channel waveguide horn.](image)

![Figure 9.6-2 Characteristic modes of a symmetric Y-junction.](image)

The operation of the device is slightly more complicated than it appears, and one useful viewpoint is provided by considering the characteristic modes of the entire structure at each end. This is particularly simple at the left-hand end. Here, the structure consists of one single-moded guide, which supports a characteristic mode with a transverse field distribution \( E_L(x, y) \), given by (say):

\[
E_L(x, y) = E(x, y)
\]

At the right-hand end, the device consists of two similar single-moded guides, separated by a distance \( S \). If we define the transverse field patterns of the upper and lower guides to be \( E_{RU} \) and \( E_{RL} \), respectively, we may write:

\[
E_{RU}(x, y) = E(x, y - S/2) \quad \text{and} \quad E_{RL}(x, y) = E(x, y + S/2)
\]

Now, the combined structure at the right-hand end can also be thought of as a composite guide, which supports its own set of characteristic modes, known as supermodes. We will now guess the form this set might take. Firstly, we shall assume that there are only two guided supermodes. This is reasonable, since the structure has twice the guiding cross-
section of a single guide. Secondly, we guess that each supermode can be written as a linear combination of the modes of each individual guide. This is also realistic; because the wave equation in the vicinity of each guide must look very like that for a single guide, the solutions must be similar. Finally, since the structure is inherently symmetric, we shall assume that the supermodes are symmetric and anti-symmetric field patterns. We shall label these patterns $E_{RS}$ and $E_{RA}$, respectively.

At the right-hand end of the device, therefore, we assume that the combined structure can support two guided supermodes, of the form:

$$E_{RS}(x, y) = E_{RU} + E_{RL}$$
$$E_{RA}(x, y) = E_{RU} - E_{RL}$$

We shall now consider the response to an input to the single guide the left-hand end, as shown in Figure 9.6-3.

![Figure 9.6-2 Excitation of a Y-junction at the left-hand end.](image)

At some reference point near the input, the transverse field can be taken as:

$$E_{in}(x, y) = a_{in} E_{l}(x, y)$$

where $a_{in}$ is the input mode amplitude. Propagation through the taper region can be analysed using the theory of local normal modes. According to this, the single input (which is a symmetric distribution) will be gradually converted into the symmetric supermode in the transition region. Because the taper itself is symmetric, no antisymmetric modes - especially the antisymmetric supermode - can be excited. Similarly, provided the taper rate is low enough, no symmetric radiation modes will be generated. In a carefully constructed Y-junction, conversion into the symmetric supermode will therefore be 100% efficient. At some reference point near the output, the transverse field can therefore be written as:

$$E_{out}(x, y) = a_{out} E_{RS}(x, y)$$

Assuming 100% efficiency, we can calculate the field amplitudes. The input power is given by:

$$P_{in} = (\beta/2\omega\mu_0) a_{in}^2 < E, E >$$

While the output power is:

$$P_{out} = (\beta/2\omega\mu_0) a_{out}^2 < E_{RS}, E_{RS} >$$

Now, if the two output guides are sufficiently far apart, Equation 9.6-7 can be approximated as:
\[ P_{\text{out}} = \left( \frac{\beta}{2\omega\mu_0} \right) a_{\text{opt}}^2 \{ \langle E_{RU}, E_{RU} \rangle + \langle E_{RL}, E_{RL} \rangle \} \]
\[ = 2 \left( \frac{\beta}{2\omega\mu_0} \right) a_{\text{out}}^2 \langle E, E \rangle \]

where we have assumed that \( \langle E_{RU}, E_{RL} \rangle = 0 \). Equating Equations 9.6-6 and 9.6-8, we see that:
\[ a_{\text{out}} = a_{\text{in}} / \sqrt{2} \]

We now change our viewpoint, and consider the device output in terms of the modes travelling in each of the two right-hand guides individually. Clearly, Equation 9.6-5 can be rewritten as:
\[ E_{\text{out}}(x, y) = a_{\text{in}} / \sqrt{2} E_{RU}(x, y) + a_{\text{in}} / \sqrt{2} E_{RL}(x, y) \]

Equation 9.6-10 implies that the Y-junction divides the input equally between the two output guides. Each has a mode amplitude of \( 1/\sqrt{2} \) times the input amplitude, and therefore carries half the power.

We now consider operation in the reverse direction, when light is incident from just one of the two channels on the right (as shown in Figure 9.6-4).

**Figure 9.6-4** Excitation of a Y-junction at the right-hand end.

This time the incident beam can be written as:
\[ E_{\text{in}}'(x, y) = a_{\text{in}}' E_{RU}(x, y) \]

Combining Equations 9.6-3, we see that this field can be expanded in terms of the supermodes, as:
\[ E_{\text{in}}'(x, y) = (a_{\text{in}}'/2) \left[ E_{RS}(x, y) + E_{RA}(x, y) \right] \]

This is physically reasonable; the supermodes add together to give a finite field amplitude in the upper guide, and cancel to give nothing in the lower one.

What happens to these two components at the junction? For the symmetric supermode, the situation is effectively that of Figure 9.6-3 in reverse, so it must be gradually converted into the single guide mode. This contribution can be written as:
\[ E_{\text{out}}'(x, y) = a_{\text{out}} E(x, y) \]

By analogy with Equation 9.6-9, we can say that:
Consequently, the amplitude of the emerging mode is $\sqrt{2}$ times that of the input mode, and hence the power carried by this contribution is half the input power. We now consider the fate of the component carried by the antisymmetric supermode. This cannot be converted into a symmetric distribution by a symmetrically tapered Y-junction, and hence cannot emerge as a guided mode at all. Instead, this component must be converted into antisymmetric radiation modes. Excitation of the Y-junction from the right hand side with a single input therefore results in an automatic loss of 50% of the power to radiation. This feature is not immediately apparent, and should therefore be considered carefully.

Y-junctions can be combined together in binary tree structures. These allow a single input to be split into $2^N$ outputs, all carrying equal power, using $N$ levels of two-way splitting. Figure 9.6-5 shows an example with three levels, which produces eight output beams from a single input. Y-junction trees are often used as input devices for parallel arrays of other components. They can also be used in reverse, to combine $2^N$ guided inputs in a single guided output. However, for the reasons just given, 100% efficiency is only achieved if all of the inputs have equal amplitude and phase.

Figure 9.6-5  A Y-junction tree structure.

9.7 PHASE MODULATORS

We now move on to discuss a range of electrically controllable devices, which require the fabrication of a waveguide in an electro-optic substrate. The Ti:LiNbO$_3$ process is an extremely simple and versatile method of constructing such devices in lithium niobate. However, many other materials (especially the semiconductor systems based on GaAs/GaAlAs and InP/InGaAsP) are also electro-optic. The first such device is a phase modulator. As we saw in Chapter 3, the most desirable configuration is a transverse modulator, where the electric field is applied in a direction orthogonal to the axis of propagation. The drive voltage required is proportional to $g/L$ where $g$ is the electrode gap and $L$ is the device length. In bulk optics, it is difficult to make $g/L$ small, which results in a high drive voltage. However, integrated optic devices require much lower voltages.

INTEGRATED OPTIC PHASE MODULATORS IN INSULATING CRYSTALS

Figure 9.7-1 shows typical geometries used with an insulating crystal such as LiNbO$_3$, which illustrate many of the differences between bulk and guided-wave electro-optic devices. Firstly, the electric field must now be applied through a pair of electrodes deposited on the same surface of the crystal, as shown in Figure 9.7-1a. Secondly, these must be arranged to lie somewhere above a waveguide, as in Figures 9.7-1b and c. (Note that, in LiNbO$_3$ devices, a thin dielectric buffer layer - usually $\approx 2000 \text{ Å}$ of SiO$_2$ - is inserted to space the electrodes away from the waveguides, to avoid any undesirable polarization effects.) The resulting static electric field may then be used to alter the refractive index in
the region of the guide, which in turn alters the phase of a guided optical wave. However, due to the strong confinement of the mode, the electrode gap can now be very small (of the order of a waveguide width), so that a high field can be created in the gap region using a greatly reduced voltage (≈ 10 V).

Calculation of the field distribution $E_s(x, y)$ produced by the electrodes is a difficult electrostatic problem, which requires the solution of Laplace's equation for a relatively complicated set of boundary conditions. Although this can be done, it requires knowledge either of an analytical technique known as conformal mapping, or of numerical methods for the solution of partial differential equations. Since both topics lie some way from our main theme, we shall simply assume the known solution directly.

The field distribution can be shown to be highly non-uniform, as indicated by the field lines in Figure 9.7-1a. Their direction defines the local field direction, and their spacing its strength. As can be seen, they run down from one electrode and up into the other; in between, they are roughly horizontal. The vertical component of the field therefore peaks just below the edge of each electrode, while the horizontal field peaks in the gap region. Since the optical field is significant only near the waveguide core, we can exploit either the vertical or horizontal parts of the electric field by placing the guide in different positions beneath the electrodes. Figure 9.7-1b shows how the guide should be placed if the vertical field component is to be used. For Z-cut LiNbO$_3$, with the optical wave propagating in the Y-direction, the TM mode is then affected by the strong $r_{33}$ electro-optic coefficient, while the TE mode is controlled by $r_{13}$. Figure 9.7-1c shows the placement for the horizontal field, which requires a different crystal cut. Using semiconductors, the construction of an electro-optic device is more complicated, and special precautions are required to obtain a high field in the guide core. These will be discussed in Chapter 12.

**CALCULATION OF THE INDUCED PHASE SHIFT**

The two effects described above - non-uniformity of electrical and optical fields - make it harder to work out the induced phase shift in integrated optic devices. However, this can be
done approximately using a perturbation method, as we shall now show. First, we suppose that the guide is single-moded, and oriented in the z-direction. If it is formed by a two-dimensional index distribution \( n(x, y) \), the modal field must satisfy the scalar waveguide equation:

\[
\nabla^2_{xy} E(x, y) + \left[ n^2(x,y)k_0^2 - \beta^2 \right] E(x, y) = 0
\]

9.7-1

We now suppose that the effect of the static field \( E_s(x, y) \) is to induce a two-dimensional refractive index perturbation \( \Delta n(x, y) \), in the region from \( z = 0 \) to \( z = L \), so that the new refractive index in this region is defined by \( n' = n + \Delta n \). Next, we assume that the effect on the guided mode will be to change its propagation constant to \( \beta' = \beta + \Delta \beta \), without altering its transverse field distribution. This is reasonable, if the mode is initially well confined. The waveguide equation for the modified guide is therefore as Equation 9.7-1, reading \( n' \) for \( n \) and \( \beta' \) for \( \beta \). If \( \Delta n \) and \( \Delta \beta \) are small, we may put:

\[
n'^2 = (n + \Delta n)^2 \approx n^2 + 2n\Delta n \quad \text{and} \quad \beta'^2 \approx \beta^2 + 2\beta\Delta \beta
\]

9.7-2

so that the waveguide equation for the perturbed mode is given by:

\[
\nabla^2_{xy} E(x, y) + \left[ \left( n(x, y)^2 + 2n(x, y)\Delta n(x, y) \right) k_0^2 - \left[ \beta^2 + 2\beta\Delta \beta \right] \right] E(x, y) = 0
\]

9.7-3

This looks rather formidable. However, we note that many of the terms can be cancelled out using the unperturbed waveguide equation, Equation 9.7-1. Removing these terms leaves:

\[
\{2n(x, y)\Delta n(x, y) k_0^2 - 2\beta\Delta \beta\} E(x, y) = 0
\]

9.7-4

Unfortunately, Equation 9.7-4 is a two-dimensional equation, since it contains a number of spatially varying quantities; really, we wish only to extract the scalar term \( \Delta \beta \). This can be done by multiplying both sides of the equation by \( E^\ast(x, y) \) and integrating over all space. If this is done, we get:

\[
k_0^2 < n\Delta n E , E > - \beta\Delta \beta < E , E > = 0
\]

9.7-5

so the change in propagation constant is given by:

\[
\Delta \beta = (k_0^2/\beta) < n\Delta n E , E > / < E , E >
\]

9.7-6

In Equation 9.7-6, the numerator is the most important term, since the denominator just represents a normalisation factor (which can be ignored if we have arranged that \( < E , E > = 1 \)). The change in propagation constant is then found as an overlap between the electrically-induced index change \( \Delta n(x, y) \) and the modal field \( E(x, y) \). This reflects the fact that the electric and optical fields are both localised; we must therefore arrange them to overlap spatially, or the index perturbation will have no effect on the guided mode. Assuming that \( < E , E > \) is indeed unity, and that the guide is weak, we may find the change in effective index as:

\[
\Delta \text{neff} = (k_0/\beta) < n\Delta n E , E > = < \Delta n , |E|^2 >
\]

9.7-7

Now, from the results of Chapter 3, we know that the index change \( \Delta n \) will be linearly proportional to one particular component of the electric field (which we define as \( E_s' \)), following:
\[ \Delta n(x, y) = -n^3 r \frac{E'_s(x, y)}{2} \]

Here \( n \) and \( r \) are the relevant refractive index and electro-optic coefficient, respectively. \( \Delta n_{\text{eff}} \) is then:

\[ \Delta n_{\text{eff}} = -n^3 r/2 \langle E_s', |E|^2 \rangle \]

The field \( E' \) will in turn be proportional to the drive voltage \( V \), and inversely proportional to the electrode gap \( g \). It is therefore often useful to write Equation 9.7-9 in the alternative form:

\[ \Delta n_{\text{eff}} = -n^3 rV \Gamma /2g \]

where we have introduced an overlap factor \( \Gamma \), defined as:

\[ \Gamma = g/V \langle E_s', |E|^2 \rangle \]

The overlap factor is normally calculated by numerical integration, and lies in the range \( 0 \leq \Gamma \leq 1 \). Incomplete overlap between the electrical and optical fields therefore causes a reduction in the effective index change over the value expected for uniform fields. For well-designed devices, \( \Gamma \) is 0.6 - 0.7. The phase shift \( \Phi = \Delta \beta L \) corresponding to the change in effective index above is then:

\[ \Phi = 2\pi \Delta n_{\text{ef}} L/\lambda_0 = -\pi n^3 rV L \Gamma /g \lambda_0 \]

In a device with a given electrode gap, the phase shift is linearly proportional to both drive voltage and device length. It is then convenient to characterise transverse modulators in terms of the voltage-length product \( (VL)_n \) required for a phase-shift of \( |\Phi| = \pi \) radians, given by:

\[ (VL)_n = g \lambda /n^3 r \Gamma \]

Note that this implies that larger drive voltages are required at longer optical wavelengths.

DESIGN EXAMPLE

We may estimate \( (VL)_n \) for Ti:LiNbO\(_3\) devices as follows. For Z-cut lithium niobate, for example, with propagation in the Y-direction and TM polarization, the refractive index is altered by the Z-component of the electric field via the \( r_{33} \) coefficient (note the use of uppercase letters to describe the crystal axes, which are slightly different from the \( x, y, z \) coordinates above). In this case, we may write \( (VL)_n = g \lambda /n^3 r_{33} \Gamma \). Assuming that \( \lambda_0 = 1.5 \mu m \), that the electrode gap is 7 \( \mu m \), and that the overlap factor \( \Gamma \) is 0.6, we obtain \( (VL)_n = 7 \times 10^{-6} \times 1.5 \times 10^{-6} / (2.23 \times 30.8 \times 10^{-12} \times 0.6) = 0.053 \) V m, or roughly 50 V mm. A 5 mm-long device would then require a 10 V drive.

CUTOFF MODULATORS

If the field is poorly confined, our assumption that the transverse field distribution of the mode will remain unaltered when the index change is applied becomes invalid. For example, the guide index may be depressed sufficiently far that the mode becomes cut off, and radiates into the substrate. This type of device is known as a cutoff modulator, and can be used to provide intensity modulation directly, rather than a simple phase change. Figure
9.7-3a shows a schematic of the device, which consists of a pair of phase-modulator electrodes operating in a region of weak optical confinement.

Figure 9.7-3 A cutoff modulator: a) schematic, b) operation with no applied voltage, and c) operation with a voltage sufficient to induce cutoff.

With no applied voltage, the guided beam is transmitted through the active section with little attenuation (Figure 9.7-3b). If a voltage of the correct sign is applied, this section approaches cutoff, and the guided beam suffers heavy attenuation through conversion to radiation modes (Figure 9.7-3c).

9.8 FREQUENCY SHIFTING AND HIGH SPEED OPERATION

A phase modulator can be used to obtain frequency modulation instead of phase shifting, by using a time-varying electrical drive. The simplest method, known as the Serrodyne, uses a sawtooth drive voltage as shown in Figure 9.8-1.

In the regions of linear ramp, the voltage varies linearly with time, so that \( V(t) \) is given by:

\[
V(t) = V_{\text{max}} t/T
\]

where \( T \) is the period of the sawtooth and \( V_{\text{max}} \) is the maximum voltage. This results in a sawtooth variation in phase-shift, which (again in the regions of linear ramp) varies as:

\[
\Phi(t) = \alpha V(t) = \Phi_{\text{max}} t/T
\]

where \( \alpha \) is a proportionality constant and \( \Phi_{\text{max}} \) is the maximum phase shift. Consequently, the electric field at the output of the phase-shifter is given by:

\[
E_{\text{out}}(x, y, t) = E_{\text{out}}(x, y) \exp(-j\Phi_{\text{max}} t/T)
\]
Now, this can be written as:

\[ E_{\text{out}}(x,y,t) = E_{\text{out}}(x, y) \exp(j\Delta \omega t) \]  

where:

\[ \Delta \omega = -\Phi_{\text{max}} / T \]

The effect of the time-varying phase shift is therefore equivalent to a constant downward frequency shift in the regions of linear ramp. If the drive is arranged so that the voltage steps between the linear regions correspond to discontinuities of \(2\pi\) in \(\Phi\), these jumps have no noticeable effect and the output is always a frequency-shifted wave. The Serrodyne method is very simple and effective, but the frequency shift is limited by the difficulty of generating a sawtooth wave at high frequencies.

![Figure 9.8-1 Variation of drive voltage and phase shift in a Serrodyne frequency shifter.](image)

**DESIGN EXAMPLE**

We may use the analysis above to calculate the frequency shift produced by a Serrodyne phase modulator of 5 mm length, driven by a 20 V peak-to-peak sawtooth drive voltage with a period of 1 \(\mu\)s, given that the static voltage-length product required for \(\pi\) radians phase shift is \((VL)_s = 50\) V mm. In this case, \(\Phi_{\text{max}} = \pi \times 20 \times 5 / 50 = 2\pi\) and \(T = 1\) \(\mu\)s, so \(\Delta f = 1\) MHz.

**MODULATION WITH SINUSOIDAL VOLTAGES**

Another common form of drive voltage (which is considerably easier to generate at high frequencies) is a sinusoid. If the modulation frequency is \(\omega_m\), the phase shift has the form:

\[ \Phi(t) = \Phi_{\text{max}} \sin(\omega_m t) \]

The wave emerging from the phase shifter is then given by:
We can then use a well-known mathematical identity to expand the exponential in Equation 9.8-7 as a sum of frequency-shifted waves, as:

\[ E_{\text{out}}(x, y, t) = E_{\text{out}}(x, y) \sum_n (-1)^n J_n(\Phi_{\text{max}}) \exp(jn\omega_m t) \]  

where \( n \) is an integer, and \( J_n \) is the \( n \)th order Bessel function. As can be seen, the use of a sinusoidal drive does not yield a single frequency-shifted component. Instead, a whole series of upper and lower sidebands are generated. The individual components have angular frequencies of \( \omega + n\omega_m \), so they lie at regular spacings of \( \omega_m \) in frequency space, and have Bessel function amplitude coefficients. This frequency spectrum is very similar to that of a FM wave in communications theory. Consequently, the effect of raising \( \Phi_{\text{max}} \) by increasing the drive voltage is not an increase in amplitude of the desired frequency-shifted component - rather, it is that more and more sidebands are generated. This is often undesirable, since it increases the overall signal bandwidth.

Figure 9.8-2 shows the variation of the three lowest-order Bessel functions \( J_0(\Phi) \), \( J_1(\Phi) \) and \( J_2(\Phi) \) with \( \Phi \). These are slowly varying functions, whose values are found tabulated in many standard mathematical texts. The fraction of the optical power \( P_n \) contained in the nth sideband is given by:

\[ P_n = J_n^2(\Phi_{\text{max}}) \]  

Thus, for an example maximum phase shift of \( \pi \) radians, the power contained in the first (upper) sideband is \( P_1 = J_1^2(\pi) = 0.0813 \), slightly more than 8% of the input power.

**HIGH-SPEED MODULATION**

The speed at which a phase modulator can be driven is of course limited, and the high-frequency cutoff depends strongly on the electrode design. Figure 9.8-3 shows a typical layout for a Ti:LiNbO₃ device, where the electrodes can be considered as **lumped elements**. These act as capacitors, which must be charged and discharged by the voltage source. Typically, this has a fixed 50 \( \Omega \) output impedance, so the high-frequency cut-off is...
determined by an RC time constant, where \( R = 50 \, \Omega \) and \( C \) is the electrode capacitance. Now, \( C \) can be reduced either by decreasing the electrode length, or by increasing the electrode gap. Unfortunately, both these strategies also increase the drive voltage, the first because the voltage-length product \( (VL) \) of a given design is fixed, and the second because it results in a weaker electric field. One possible solution is to match the electrodes to the generator using a matching network, but broad band matching is virtually impossible. Consequently, broad band lumped-element devices have a high-frequency cut-off at around 1 GHz.

![Figure 9.8-3](image1)

Figure 9.8-3  A channel waveguide phase modulator with lumped-element electrodes.

One way of improving high-frequency performance is to replace the lumped-element electrodes with electrodes that act as a waveguide at microwave frequencies, known as **travelling-wave electrodes**. Typically, the required structure is an asymmetric coplanar stripline, constructed from a pair of electrodes with greatly increased metal thickness \((\approx 15 \, \mu m)\), as shown in Figure 9.8-4.

![Figure 9.8-4](image2)

Figure 9.8-4  A channel waveguide phase-shifter with travelling wave electrodes.

This can be arranged to present a constant, real impedance at microwave frequencies, which can be matched to the output impedance of the generator quite simply. The electrical signal is then launched as a travelling microwave at one end of the electrodes, travels the length of the device, and is absorbed in a matched termination at the far end. This time, the high-frequency cut-off is determined by a mismatch in the velocity of the optical and electrical waves, due to the difference in dielectric constant at the two frequencies concerned. This is highly significant for LiNbO\(_3\) devices. The result is reduced modulation at high frequencies, but operation at 10 - 20 GHz is quite possible. Figure 9.8-5 shows a fully packaged Ti:LiNbO\(_3\) microwave phase modulator, fitted with fibre pigtails.
9.9 INTERFEROMETERS

Phase modulation can be converted into amplitude modulation, by inserting a phase shifter in one arm of an interferometer. The simplest is the Mach-Zehnder interferometer, shown in Figure 9.9-1.

![Mach-Zehnder interferometer](image)

Figure 9.9-1  An electro-optic channel waveguide Mach-Zehnder interferometer, constructed using Y-junctions.

The structure consists of two back-to-back Y-junctions, separated by a region containing a pair of identical electro-optic phase modulators. It operates as follows. The input wave is first split into two components by the left-hand Y-junction. Using our previous results, the transverse field across the device just before the phase shifter region can be written as:

\[ E(x, y) = \left( a_n / \sqrt{2} \right) [E_{RU}(x, y) + E_{RL}(x, y)] \]

As before, \( a_n \) is the modal amplitude of the single-guide input, and \( E_{RU} \) and \( E_{RL} \) are the transverse fields of the modes in the upper and lower guides, respectively. Each beam then travels through a phase-shifter section, and opposite shifts of \( \Phi \) are applied to each, to get a
push-pull effect. The transverse field across the whole device just beyond the phase-shifters can then be written as:

$$E(x, y) = (a_{in}/\sqrt{2}) \left[ E_{RU}(x, y) \exp(-j\Phi) + E_{RL}(x, y) \exp(+j\Phi) \right]$$  \hspace{1cm} 9.9-2

Now, this field can be expressed alternatively, in terms of the characteristic modes of the two-guide system forming the central region of the device. If this is done, we obtain:

$$E(x, y) = (a_{in}/\sqrt{2}) \left[ E_{RS}(x, y) \cos(\Phi) - j E_{RA}(x, y) \sin(\Phi) \right]$$  \hspace{1cm} 9.9-3

where $E_{RS}$ and $E_{RA}$ are the transverse fields of the symmetric and antisymmetric supermodes in the central region. This distribution now passes into the right-hand Y-junction. For the reasons given earlier, the component carried by the symmetric supermode emerges from the single guide output, while the antisymmetric component is radiated. The amplitude of the guided output is thus given by:

$$a_{out} = a_{in} \cos(\Phi)$$  \hspace{1cm} 9.9-4

so that the normalised output power $P = P_{out}/P_{in}$ can be found as:

$$P = \cos^2(\Phi) = [1 + \cos(2\Phi)] / 2$$  \hspace{1cm} 9.9-5

The normalised output therefore varies cosinusoidally with $\Phi$, as shown in Figure 9.9-2.

An interferometer of a rather different kind can be made by allowing two waveguides to cross at a shallow angle, as shown in Figure 9.9-3. This is known as an X-switch. In general, such a device has two possible modes of operation (one interferometric, the other operating via switchable total-internal-reflection). Exactly which occurs depends the design parameters, but here we will assume the interferometric mode is operative. If the device is fabricated in an electrooptic material, it can perform intensity modulation in a similar manner to the Mach-Zehnder interferometer. However, because it has two guided output ports, it can also function as a two-way switch.
In the central region of the switch, the guide is arranged to support two modes, which are as usual symmetric and antisymmetric in character. This can be done by making the index change forming the guide in this region double that elsewhere. Now, an input to one of the switch arms will excite both the characteristic modes equally. These travel at different speeds to the two output arms, where they are recombined. Depending on their relative phases, a high output is obtained from either of the arms.

Electrodes are now added to the device, with a narrow central gap. When a voltage is applied, the static electric field is high down the device centre line, and weak elsewhere. The consequence is an induced index change in the gap region. The overlap between the symmetric mode and this perturbation is strong, so its phase is altered according to the phase modulation principle described earlier. However, the overlap with the antisymmetric mode is much smaller, since this must have zero amplitude at its centre. A relative phase difference is then induced between the two modes, so that interferometric addition or cancellation occurs, and the power output from each arm is sinusoidally dependent on the applied voltage. Although the device is difficult to make, it has the advantage of requiring no bends.
PROBLEMS

9.1. Describe briefly the difference in structure of semiconductor waveguides fabricated using homo-structures and heterostructures. Calculate the maximum possible thickness of the guiding layer of a symmetric GaAs/Ga$_{0.85}$Al$_{0.15}$As double heterostructure guide, if single-mode operation is desired at a wavelength of 0.9 µm.

\[0.55 \mu m]\n
9.2. For a particular channel guide system, experimental measurements of the radiation loss incurred in 90° waveguide bends showed a 10 dB loss for a 1 mm bend radius, and 3 dB loss for 1 cm radius. Estimate the loss incurred in a 45° bend of 5 mm radius.

\[5.27 \text{ dB}\]

9.3. A ninety-degree bend is formed in a single-mode semiconductor ridge waveguide system using a total-internal reflection mirror. Unfortunately, due to an error, the reflecting interface is offset by $\delta$ from its optimum position, as shown below. Explain how you would estimate the transmission through the bend. Can the transmission reach 100%, if $\delta = 0$?

9.4. The figure below shows a Y-junction tree, used as a beam combiner for eight single-mode inputs. If the input mode amplitudes in guides 1 ... 8 are 1, -2, 3, 0, -1, 2, 4 and -1, respectively, calculate the efficiency of the beam-combining process. How does the efficiency improve, if corrections are made to the input phases so that the contributions sum in-phase?

9.5. Sketch the electric field lines associated with the following surface electrode geometries. In each case, indicate where you would place a pair of channel guides, if the electrodes are to be used to provide push-pull phase-shifting in the central region of a Mach-Zehnder interferometer.
9.6. A cutoff modulator is based on a single-moded Ti:LiNbO₃ channel waveguide with an effective index of \( n_{\text{eff}} = 2.201 \). Assuming that the substrate index is \( n_e = 2.2 \), that the device will operate via the \( r_{33} \) electro-optic coefficient \( (30.8 \times 10^{-12} \text{ m/V}) \) using surface electrodes with a 7 \( \mu \text{m} \) gap width, and that overlap between the electrical and optical fields is \( \Gamma = 0.6 \), estimate the voltage required to induce cutoff. Why might your estimate prove inaccurate?

[71.1 V]

9.7. A 10 mm-long electro-optic phase modulator is driven with a 10 V peak-to-peak sinusoidal voltage, at a frequency of 10 MHz.

Assuming that the voltage-length product required for \( \pi \) radians phase shift is \( (V\ell)_{\pi} = 50 \text{ V mm} \), find the percentage of the output power contained in a bandwidth ranging from 20 MHz below the optical carrier frequency to 20 MHz above it.

[72.5%]

9.8. The figure below shows a travelling-wave electro-optic phase modulator. Assuming that the angular frequencies of the electrical drive and the optical wave are \( \omega_m \) and \( \omega_0 \), respectively, and that the effective indices of the stripline waveguide and the optical guide are \( n_m \) and \( n_o \), calculate and sketch the dependence of the induced optical phase shift \( \Phi \) on \( \omega_m \). At what frequency does \( \Phi \) first fall to zero?

9.9. A Mach-Zehnder interferometer is constructed using two identical phase-shifters, which provide push-pull modulation. Each is 1 cm in length, and has a voltage-length product for \( \pi \) radians phase-shift of \( (V\ell)_{\pi} = 50 \text{ V mm} \). Sketch accurately the variation of guided output power with time, for a constant optical input and the D.C.-biassed triangle-wave drive shown below.
9.10. The figure below shows single-mode channel waveguide integrated optic circuit. Explain the operation of each component, and of the circuit as a whole. How does the output power vary, as a function of the path imbalance \( L \)? What might the device be used for?
SUGGESTIONS FOR FURTHER READING


Buchmann P., Kaufmann H., Melchior H., Guebos H. "Totally reflecting mirrors: fabrication and application in GaAs rib devices" Proc. 3rd European Conf. on Integrated Optics, Berlin, Germany, Pp 135-139, 6th-8th May (1985)


Ctyroky J., Henning H.-J. "Thin-film polarizer for Ti:LiNbO 3 waveguides at \( \lambda = 1.32 \mu m \)" Elect. Lett. 22, 756-758 (1986)

Findakly T. "Glass waveguides by ion exchange: a review" Opt. Eng. 24, 244-250 (1985)


Minford W.J., Korotky S.K., Alferness R.C. "Low-loss Ti:LiNbO 3 waveguide bends at \( \lambda = 1.3 \mu m \)" IEEE Trans. Micr. Theory & Tech. MTT-30, 1790-1794 (1982)


CHAPTER TEN

COUPLED MODE DEVICES

10.1 INTRODUCTION

Until now, we have been concerned mainly with the properties of perfect, uniform guides. We have shown them all to support a set of characteristic modes, which are (by definition) field structures that propagate down the guide with only a change in phase. We have encountered some exceptions to this rule - most notably, waveguide tapers, bends, gratings and beam couplers, where the mode amplitudes also change due to power redistribution between guided modes, or conversion from guided to radiation modes - but we have managed to avoid any quantitative calculations of the amplitude changes concerned. It is now the time to rectify this omission.

It turns out that a common physical principle underpins the behaviour of nearly all perturbed waveguide structures: namely, that the introduction of a perturbation to an otherwise perfect guide can cause an interchange of energy amongst the modes of that guide. This interchange can be highly significant, if a number of conditions are satisfied. In fact, one individual mode can be converted into another with close to 100% efficiency in many cases. This principle, known as mode coupling, is ubiquitous, and occurs in many branches of physics other than guided wave optics (most notably, in quantum mechanics). As far as we are concerned, it is important to note that besides explaining the performance of waveguides suffering from some imperfection or distortion, the principle can also be exploited for use in new guided wave devices. We shall illustrate this point by considering some of the wide range of devices that operate through coupling between co- and contra-directional modes.

10.2 THE DIRECTIONAL COUPLER - BASIC PRINCIPLES

We will start with one of the most successful and versatile devices in both integrated and fibre optics, the directional coupler, which works by coupling together two modes travelling in the same direction. In its simplest form, this acts as a beam splitter, but more complicated devices can be used as two-way switches or modulators; further variants can be used as filters or polarizers. In broad outline, the device works as follows.

From previous experience, we know that there is an evanescent field extending outside any dielectric waveguide. If two parallel guides are placed sufficiently close together, these parts of the field must overlap spatially. Usually, the interwaveguide gap required for this overlap to be significant is of the order of the guide width. For example, Figure 10.2-1 shows the situation for the lowest-order modes in two identical, adjacent symmetric slab waveguides. The refractive index distribution of the twin-guide system is also shown. The indices of the two guide cores are $n_{g1}$ and $n_{g2}$, respectively, and the index in the substrate material outside the guide has the uniform value $n_s$.

Because of the intermodal overlap, it is hard to decide which guide any light in the overlap region actually belongs to. This implies that there must be a mechanism for light in one guide to be transferred to the other. It turns out (as we shall see later) that, if the two guides satisfy a particular set of conditions, and run parallel for sufficient distance, this interchange of power can be highly significant, reaching almost 100% in many cases. The light then starts coupling back, so the power transfer process must be periodic with distance. We can represent this as shown in Figure 10.2-2.
Here the input is to Guide 1, at \( z = 0 \). We shall assume that complete coupling occurs at \( z = L \), when all the light has been transferred to Guide 2. At \( z = 2L \), therefore, the light will have been completely coupled back to Guide 1, and so on. In fact, the process is a continuous, bidirectional one, of a type known as multiple scattering. The condition the guides must satisfy for 100% power transfer is therefore that they are synchronous (i.e. have the same propagation constants) so that all the scattered contributions add up in-phase. This usually requires them to be identical.

We can use this effect to make a device that can switch light between two waveguides, as follows. We start with two identical single-moded guides, and choose the length of the coupling region so that the device runs from \( z = 0 \) to \( z = L \). Under normal circumstances, an input to Guide 1 will then be completely coupled across, and emerge as an output from Guide 2. If we can now arrange to desynchronise the guides under electrical control, we can prevent the power transfer process occurring. In this case, the light emerges from Guide 1, without any cross coupling. Since the choice of output guide is now electrically alterable, the device functions as a two-way switch.

INTEGRATED OPTIC DIRECTIONAL COUPLERS IN INSULATING CRYSTALS

Figure 10.2-3 shows the cross-section of a structure suited to realisation in integrated optics. This is made from two channel waveguides buried in an insulating, electro-optic...
substrate, with a pair of surface electrodes. The Ti:LiNbO$_3$ process is now well developed for fabrication of such devices. We have already discussed the effect of this type of electrode pair in Chapter 9. When a D.C. voltage is applied between the two electrodes, the result is a static electric field, represented by the field lines in Figure 10.2-3. Note that the field is directed mainly downwards through the left-hand guide, and upwards through the right-hand one. This will produce a change in the refractive index through the linear electro-optic effect. The exact result will depend on the precise field variation, but the crystal orientation will also affect matters. For simplicity, let us assume that the crystal is chosen so that the vertical field component is the important one for the particular polarization of light involved.

![Figure 10.2-3 Cross-section of an electro-optic directional coupler.](image1)

Now, the linear electro-optic effect is (by definition) linear. In this case, therefore, we can deduce that the index change must be of opposite sign in the two guides. Consequently, any change in guide propagation constant that results must also be of opposite sign. The application of a D.C. voltage to the electrodes can therefore be used to desynchronise two otherwise identical guides. The difference between the propagation constants is often written as $\Delta \beta = \beta_2 - \beta_1$, and is proportional to voltage. The electrode structure shown is consequently often known as a pair of $\Delta \beta$ electrodes. An electro-optic device using such electrodes is called a **directional coupler switch**, shown in plan view in Figure 10.2-4. Here, additional bend sections have been included to separate the waveguides at the input and output, for example for coupling to pairs of fibres (which have a large core-to-core separation due to their cladding). Integrated optic directional couplers in semiconducting crystals require a rather different electrode construction, which will again be covered in Chapter 12.

![Figure 10.2-4 Plan view of a directional coupler switch.](image2)
CHAPTER TEN

10.3 THE DIRECTIONAL COUPLER - THEORETICAL ANALYSIS

We will now form a theoretical model of the directional coupler. The derivation of the relevant equations is, unfortunately, rather lengthy, so for simplicity we base the analysis on scalar theory, and concentrate on the central region of the device where the guides run parallel. We start by assuming that the two guides are oriented in the z-direction, and are described in isolation by the refractive index distributions \( n_1(x, y) \) and \( n_2(x, y) \), respectively. Figure 10.3-1 shows how these distributions might look for a pair of symmetric slab waveguides. Here the refractive index outside the guides is everywhere \( n_s \), while the indices in the two guiding layers are \( n_{g1} \) and \( n_{g2} \) respectively. Assuming that the electric field is polarized in the y-direction, the scalar wave equation governing the variation of the electric field \( E_{yi}(x, y, z) \) for each guide in isolation is then given by:

\[
\nabla^2 E_{yi}(x, y, z) + n_i^2(x, y)k_0^2 E_{yi}(x, y, z) = 0
\]

Both guides are now assumed to be single-moded, with guided eigenmode solutions defined by:

\[
E_{yi}(x, y, z) = E_i(x, y) \exp(-j\beta_i z)
\]

Where \( i = 1 \) or \( 2 \), and where \( E_i(x, y) \) is the transverse field of the mode in the \( i^{th} \) guide, and \( \beta_i \) is its propagation constant. These solutions must satisfy the individual waveguide equations:

\[
\nabla^2_{xy} E_i(x, y) + [n_i^2(x, y)k_0^2 - \beta_i^2] E_i(x, y) = 0
\]

Here the subscripts on the Laplacian imply that the differentiation is to be performed with respect to \( x \) and \( y \) only. We take it that these equations have previously been solved by some means, so that we know both of the transverse field distributions \( E_i(x, y) \) and both propagation constants \( \beta_i \).
Now, we assume that the complete coupler is described by a refractive index distribution \( n_T(x, y) \), which corresponds to both guides together. We have already shown in Figure 10.2-1 how this distribution might look for the slab waveguide model used above. The scalar wave equation for the complete coupler is then:

\[
\nabla^2 E_y(x, y, z) + n_T^2(x, y)k_0^2 E_y(x, y, z) = 0
\]

Now, because we expect power to be coupled slowly from one guide to the other, we assume that we can describe the solution for \( E_y(x, y, z) \) as a superposition of the modes in the two isolated guides, i.e. as a linear combination of the two modal solutions previously found. We therefore put:

\[
E(x, y, z) = A_1(z) E_1(x, y) \exp(-j\beta_1 z) + A_2(z) E_2(x, y) \exp(-j\beta_2 z)
\]

Carrying out the necessary differentiation, we can show that:

\[
\nabla^2 E_y(x, y, z) = \left[ A_1 \nabla_{xy}^2 E_1 + (d^2 A_1/dz^2 - 2j\beta_1 dA_1/dz - \beta_1^2 A_1) E_1 \right] \exp(-j\beta_1 z) + \left[ A_2 \nabla_{xy}^2 E_2 + (d^2 A_2/dz^2 - 2j\beta_2 dA_2/dz - \beta_2^2 A_2) E_2 \right] \exp(-j\beta_2 z)
\]

We now substitute this expression, and the assumed solution of Equation 10.3-5, into the scalar wave equation (Equation 10.3-4). The resulting equation can then be drastically simplified, by eliminating the \( \nabla_{xy}^2 \) terms using the waveguide equations for the individual guides (10.3-3). The relevant terms are merely collected together, and set equal to zero. The result is:

\[
\left[ d^2 A_1/dz^2 - 2j\beta_1 dA_1/dz + k_0^2(n_T^2 - n_1^2) A_1 \right] E_1 \exp(-j\beta_1 z) + \left[ d^2 A_2/dz^2 - 2j\beta_2 dA_2/dz + k_0^2(n_T^2 - n_2^2) A_2 \right] E_2 \exp(-j\beta_2 z) = 0
\]

Note the appearance in Equation 10.3-7 of two new functions, \( n_T^2 - n_1^2 \) and \( n_T^2 - n_2^2 \). Bearing in mind the square-law relationship between refractive index and relative dielectric constant, \( n^2 = \varepsilon_r \), we can interpret these as follows. The first represents a perturbation in dielectric constant \( \Delta \varepsilon_{1r} \) to Guide 1, caused by the introduction of a neighbouring waveguide, Guide 2. This will clearly be non-zero only in the neighbourhood of Guide 2. The second is a similar perturbation \( \Delta \varepsilon_{2r} \) caused to Guide 2 by Guide 1, and will in turn be significant only near Guide 1.

Of course, we can make further simplifications. For example, we can neglect second derivatives of the amplitudes (e.g. \( d^2 A_i/dz^2 \)), on the grounds that these represent modal envelopes, which will probably vary only slowly with distance. We can also divide through by \( \exp(-j\beta_i z) \). This yields:

\[
[-2j\beta_1 dA_1/dz + k_0^2(n_T^2 - n_1^2) A_1] E_1 + [-2j\beta_2 dA_2/dz + k_0^2(n_T^2 - n_2^2) A_2] E_2 \exp(-j\Delta \beta z) = 0
\]

where \( \Delta \beta = \beta_2 - \beta_1 \) represents the mismatch in propagation constant between the guides. We are now nearly there; Equation 10.3-8 is almost (but not quite) useable. The problem is that we really wish to find an equation governing the changes in the wave amplitudes \( A_i \) with distance. Since these are one-dimensional functions, it is not sensible to solve a three-dimensional equation. Somehow, we must remove the x- and y-variations contained in the
quantities $E_i$ and $n_i^2 - n_e^2$. We do this in two steps. First, we multiply the whole of Equation 10.3-8 by the complex conjugate of the transverse field in Guide 1 ($E_1^*$), and integrate over the device cross-section. This yields the following new equation:

$$[-2j\beta_1 < E_1 , E_i > dA_1 /dz + < k_0^2(n_t^2 - n_e^2) E_1 , E_i > A_1 ] +$$

$$[ -2j\beta_2 < E_2 , E_i > dA_2 /dz + < k_0^2(n_t^2 - n_e^2) E_2 , E_i > A_2 ] \exp(-j\Delta\beta z) = 0$$

Here we have used the inner product notation introduced in Chapter 6 to represent all the integrals that arise. These are of course independent of $x$ and $y$, so the operation has had the desired effect.

Secondly, we now make a number of assumptions concerning the relative sizes of the remaining terms. (1) We assume that the overlap term $< E_2 , E_i >$ is negligible, i.e. that the spatial overlap of the two modes is small. This is reasonable, if the guides are not too close together. (2) We also assume that the self-coupling term $< k_0^2(n_t^2 - n_e^2) E_1 , E_i >$ is negligible. This is physically reasonable, since the integral is only non-zero in the region of the perturbation $\Delta E$, which lies in the evanescent region where the field function $E_i$ is itself small. (3) We assume on the other hand that the self-overlap term $< E_1 , E_i >$ is non-negligible, since $E_i$ must of course overlap completely with itself. (4) We also assume that the coupling term $< k_0^2(n_t^2 - n_e^2) E_2 , E_i >$ is non-negligible, since $E_i$ is significant in the region where $\Delta E$ is non-zero, the core of Guide 1. Neglecting terms (1) and (2), and retaining terms (3) and (4), leaves Equation 10.3-9 in the form:

$$-2j\beta_1 < E_1 , E_1 > dA_1 /dz + < k_0^2(n_t^2 - n_e^2) E_2 , E_1 > A_2 \exp(-j\Delta\beta z) = 0$$

Now, we could of course repeat the process above, this time multiplying Equation 10.3-8 by $E_2^*$. We can guess that the result will be the slightly different equation:

$$-2j\beta_2 < E_2 , E_2 > dA_2 /dz + < k_0^2(n_t^2 - n_e^2) E_2 , E_2 > A_1 \exp(+j\Delta\beta z) = 0$$

We now make two more simplifications. Firstly, since the propagation constants will not change by very much during the switching process, we assume that $\beta_1 \approx \beta_2 \approx \beta_0$ for the purpose of division. Secondly, we assume that the symmetry of the device will ensure that:

$$< k_0^2(n_t^2 - n_e^2) E_2 , E_1 > / < E_1 , E_1 > \approx < k_0^2(n_t^2 - n_e^2) E_1 , E_2 > / < E_2 , E_2 >$$

With these assumptions, Equations 10.3-10 and 10.3-11 become:

$$dA_1 /dz + j\kappa A_1 \exp(-j\Delta\beta z) = 0$$

$$dA_2 /dz + j\kappa A_1 \exp(+j\Delta\beta z) = 0$$

Here we have introduced a new quantity $\kappa$, which is known as the coupling coefficient. This is arguably the most important parameter of the coupler, and is defined as:

$$\kappa = (k_0^2/2\beta_0) < (n_t^2 - n_e^2) E_2 , E_1 > / < E_1 , E_1 >$$

Equations 10.3-10 and 10.3-11 are known as coupled mode equations, and represent the final stage in the derivation. Essentially, they imply that variations in the amplitude of the mode in Guide 1 are linked to the amplitude of the mode in Guide 2 through the coupling coefficient $\kappa$, and vice-versa. In other words, the mode amplitudes are coupled together. We shall see how to solve the equations very shortly; first, we will briefly consider factors affecting the value of the coupling coefficient.
FACTORS AFFECTING THE COUPLING COEFFICIENT

The expression for $\kappa$ in Equation 10.3-14 is relatively complicated. However, we note that the first term, $k_0^2/2\beta_0$, is really a constant, since it depends mainly on the optical wavelength. Similarly, the last term, $\langle E_1, E_1 \rangle$, is a normalisation factor. The salient features are all contained in the central term, $\langle (n_0^2 - n_2^2) E_2, E_1 \rangle$. This is, in effect, an overlap integral between $\Delta \varepsilon_{\alpha}$ (the dielectric constant perturbation seen by Guide 2 due to Guide 1) and the two fields $E_1$ and $E_2$.

In our slab guide model, $\Delta \varepsilon_{\alpha}$ is non-zero only within the core of Guide 1, so this region will give the only contribution to the integral. The main requirement for $\kappa$ to be large is therefore that the evanescent tail of the field $E_2$ penetrates Guide 1 to a significant extent. We can then say that $\kappa$ is most strongly affected by the interwaveguide gap $g$, which should be small (but not too small, or our equations will be invalid). Because evanescent fields fall off roughly exponentially, it turns out that $\kappa$ also depends exponentially on $g$, to good approximation. Typically, $g$ should be of the order of the guide width (usually, a few $\mu$m).

$\kappa$ is also affected by the confinement of the modes. For example, a poorly confined mode will have an evanescent field that extends a long distance from the guide core, which will result in a strong coupling coefficient. However, weak confinement may be undesirable because it increases bend loss, so the optimum confinement is something of a compromise. Finally, $\kappa$ is affected by polarization. Even though we have assumed a scalar model here, the use of a birefringent electro-optic substrate generally causes different coupling rates for TE and TM modes. This is clearly unfortunate, since the device will no longer be polarization-independent. Fortunately, careful choice of the fabrication parameters can equalise $\kappa_{TE}$ and $\kappa_{TM}$ in some cases (for example, in the Ti:LiNbO$_3$ process).

10.4 SOLUTION OF THE EQUATIONS AT SYNCHRONISM

We shall see first what happens when the two guides are synchronous, i.e. when there is no propagation constant mismatch, so that $\Delta \beta = 0$. The two coupled mode equations then reduce to:

$$\frac{dA_1}{dz} + j\kappa A_2 = 0; \quad \frac{dA_2}{dz} + j\kappa A_1 = 0$$

We can solve these equations as follows. Differentiating the left-hand one, we get:

$$\frac{d^2A_1}{dz^2} + j\kappa \frac{dA_2}{dz} = 0$$

10.4-2

Substituting for $dA_2/dz$ using the right-hand one then gives:

$$\frac{d^2A_1}{dz^2} + \kappa_2 A_1 = 0$$

10.4-3

This is a standard second-order differential equation, with the general solution:

$$A_1 = C_1 \cos(\kappa z) + C_2 \sin(\kappa z)$$

10.4-4

Here $C_1$ and $C_2$ are constants that must be chosen to satisfy the boundary conditions. If we initially launch light only into Guide 1, with unit modal amplitude, these must be that:
A_1 = 1, \text{ and } A_2 = 0 \text{ on } z = 0 \tag{10.4-5}

Now, from Equation 10.4-1, we know that:

\[ A_2 = (-1/j\kappa) \frac{dA_1}{dz} \tag{10.4-6} \]

So the boundary conditions could be written in the alternative form:

\[ A_1 = 1, \text{ and } \frac{dA_1}{dz} = 0 \text{ on } z = 0 \tag{10.4-7} \]

It is easy to see the solution must then be:

\[ A_1 = \cos(\kappa z) \tag{10.4-8} \]

The solution for \( A_2 \) can be found by differentiation, using Equation 10.4-6. The result is:

\[ A_2 = -j \sin(\kappa z) \tag{10.4-9} \]

Note that \( A_2 \) is not real in Equation 10.4-9, showing that \( A_1 \) is in quadrature with \( A_2 \). It is probably more useful to consider the powers in the two guides, rather than mode amplitudes. These can be found in the usual way, via Poynting's Theorem. Normalising relative to the input power, we get:

\[ P_1 = A_1A_1^* = \cos^2(\kappa z) ; \quad P_2 = A_2A_2^* = \sin^2(\kappa z) \tag{10.4-10} \]

This demonstrates one of the nicer features of coupled mode theory. After a string of complicated manipulations, we still end up with a physically convincing result: total power, which is of course given by \( P_1 + P_2 \), is conserved by the device. This is what we would expect, since there is no loss. We can plot the variations of \( P_1 \) and \( P_2 \) with distance as shown in Figure 10.4-1.

![Figure 10.4-1](https://example.com/figure10.4-1.png)

As can be seen, the variations are in complete agreement with our earlier qualitative discussion, and the power distribution between the guides is an oscillatory function of \( z \). All the power is transferred from Guide 1 to Guide 2 at the point \( z = L \), when \( \kappa L = \pi/2 \). Complete transfer also occurs when \( \kappa L = 3\pi/2, 5\pi/2 \) and so on. Generally, however, the device parameters are chosen for the lowest of these values. We can arrange this either by altering the length \( L \) or the coupling coefficient \( \kappa \). For Ti:LiNbO_3 channel waveguide devices, \( L_{\pi/2} \) is typically 5 - 10 mm.
10.5 ASYNCHRONOUS SOLUTION

What happens if there is a mismatch in propagation constant, so that $\Delta \beta$ is non-zero? Well, we can still solve the coupled mode equations. First, it is best to make the substitutions:

$$A_1 = a_1 \exp(-j\Delta \beta z/2) \quad \text{and} \quad A_2 = a_2 \exp(+j\Delta \beta z/2)$$

10.5-1

Clearly, we can obtain derivatives of these new variables, in the form:

$$\frac{dA_1}{dz} = \left[\frac{da_1}{dz} - j\frac{\Delta \beta}{2} a_1 \right] \exp(-j\Delta \beta z/2)$$

10.5-2

and so on. Substituting into Equations 10.3-13, we find that all the exponentials cancel out, leaving only constant terms, so that:

$$\frac{da_1}{dz} - j\frac{\Delta \beta}{2} a_1 + j\kappa a_2 = 0$$

$$\frac{da_2}{dz} + j\frac{\Delta \beta}{2} a_2 + j\kappa a_1 = 0$$

10.5-3

These equations are exactly equivalent to Equations 10.3-13, but much easier to solve. We won't repeat the mathematics involved, but merely quote the general solution. For the same input as before, the power output from Guide 2 is given by:

$$P_2 = \sin^2\left[\sqrt{(\nu^2 + \xi^2)}\right] / (1 + \xi^2/\nu^2)$$

10.5-4

where we have introduced two new parameters, a normalised coupling length $\nu$ and a normalised dephasing parameter $\xi$, given by:

$$\nu = \kappa L \quad ; \quad \xi = \Delta \beta L/2$$

10.5-5

Equation 10.5-4 looks a bit like a sinc$^2$ function (remember that sinc(x) = sin(x) / x). We can plot it out as a function of $\xi$ (together with the power in Guide 1, given by $P_1 = 1 - P_2$), for our design value of $\nu = \pi/2$. This gives the typical switching characteristic shown in Figure 10.5-1. When $\xi = 0$ (i.e., at synchronism) all the power emerges from Guide 2 as before, but as $|\xi|$ increases the power transfer is destroyed and $P_2$ falls, with a sort of 'filter' characteristic. Similarly, $P_1$ rises as $|\xi|$ increases, so that total power is once again conserved. Now, the first zero in Equation 10.5-4 occurs when:

$$\sqrt{(\nu^2 + \xi^2)} = \pi$$

10.5-6

Given that we chose a normalised coupling length of $\nu = \pi/2$, this implies that:

$$\xi = \pi \sqrt{3}/2, \quad \text{or} \quad \Delta \beta L = \pi \sqrt{3}$$

10.5-7

To make a switch, we therefore need only to be able to define two states:

i) state A, when no voltage is applied to the electrodes (so $\Delta \beta L = 0$). The light then emerges from Guide 2.

ii) state A', when a voltage is applied such that $\Delta \beta L = \pi \sqrt{3}$. The light then emerges from Guide 1.
DESIGN EXAMPLE

We can compare the voltages required to switch a coupler and a Mach-Zehnder interferometer, as follows. For example, we know from Chapter 9 that the output of the interferometer will fall to zero when there is $\pi$ radians phase difference between the arms. The equivalent value in a coupler can be found from Equation 10.5-7 as $\pi \sqrt{3}$ rads. At first sight, therefore, the coupler seems less efficient, since it will need $\sqrt{3}$ times the voltage for the same electrode geometry. However, coupler electrodes provide push-pull modulation from a narrow electrode gap. This would not be practical in the Mach-Zehnder modulator, since the guides are further apart (a bipolar drive is actually required in Figure 9.9-1 for push-pull operation). A more realistic figure for the voltage ratio is therefore $\sqrt{3}/2$.

PRACTICAL SWITCHES

The astute reader will have noticed that one switch state in Figure 10.5-1 is electrically-adjustable, but the other (at $\xi = 0$) is set by the coupling length. Since $\kappa$ is dependent on modal confinement, an error in fabrication may easily lead a value of $\nu$ different from $\pi/2$.

It is natural to ask whether the two switch states may then still be obtained. The short answer is no: the cross-coupled state will in general suffer crosstalk. As a result, an alternative set of electrodes, known as **stepped $\Delta\beta$ electrodes**, are often used. Each electrode is divided into two at its midpoint, giving two pairs of electrodes, both of length L/2. Using two separate control voltages (one for each pair) it is then possible to tune the device so that both states are always achievable. However, this can only be done if the value of $\nu$ is greater than $\pi/2$.

Directional coupler switches are often integrated together as a **switch array**. If a number of couplers are suitably arranged on a substrate, they can connect 2N inputs to 2N outputs very simply. Figure 10.5-2 shows an example of a 4 x 4 switch, connecting four input fibres to four outputs. 8 x 8 Ti:LiNbO$_3$ switch arrays are now common, but there is a limit to array size set by the dimensions of available crystal substrates.
10.6 FIBRE DIRECTIONAL COUPLERS

Directional couplers can also be made in single-mode optical fibre form. The fibre is first mounted in a silica or glass block, which has a curved groove cut in it, and the block is then polished down until the fibre core is just exposed. Two such blocks are then placed together, so that the polished areas are in contact, with a little index-matching fluid at the join. The evanescent fields in the two fibres then overlap, so power is transferred between them much as in the integrated device above. This arrangement is known as a polished fibre coupler, and is shown in Figure 10.6-1.

The main differences from the integrated optic device are that the coupling is non-uniform, because the spacing between the fibre cores varies with distance, and that the fibre version is not switchable. However, the coupling efficiency may easily be tuned, simply by sliding the blocks across each other as shown in Figure 10.6-2a. This alters the core-to-core separation, and hence the coupling coefficient, allowing simple variation of the splitting ratio. Often, this is adjusted for a coupling length of $\nu = \pi/4$, so the device acts as a 50 : 50 splitter. Figure 10.6-2b shows a tunable single-mode fibre coupler.

If ordinary circular fibre is used to fabricate a polished coupler, the resulting component will effectively scramble the input polarization, because there will be coupling between orthogonal polarization modes. To make a polarization-preserving coupler, this process must be suppressed. The easiest way is to use polarization-preserving fibre, so that orthogonal polarization modes are not synchronous. D-fibre and bow-tie fibre (described in Chapter 8) are both suitable. The only complication is that the polarization axes of the
fibres in the two blocks must be parallel; the usual approach is to arrange them to be parallel and perpendicular to the polished surfaces.

Figure 10.6-2 a) Arrangement for tuning the coupling strength of a polished fibre coupler, and b) a commercial tunable fibre coupler (photograph courtesy A.Tobin, Sifam Ltd.).

THE FIBRE LOOP MIRROR

Polarization-preserving couplers allow the construction of fibre circuits that operate on interferometric principles. We have already encountered the integrated optic Mach-Zehnder interferometer, and we will describe the use of fibre MZIs in Chapter 14. Here, we will consider an alternative circuit, the Sagnac interferometer, which can be used as an all-fibre mirror. Figure 10.6-3 shows the circuit, which consists of a fibre coupler with two of its outputs connected together in a loop.

To show that this acts as a mirror, we will consider the generation of a reflected output. There are two possible paths between input and output that lead to a reflection. In the first, light traverses the coupler without cross-coupling, travels clockwise round the loop, and is then cross-coupled on meeting the coupler again. For unity input, the resulting amplitude is then:

$$A_{ri} = \cos(\kappa L) \exp(-j\phi) - j \sin(\kappa L)$$ 10.6-1
Here the first term is the coupler transmission without cross-coupling, the second is the phase-shift incurred in travelling round the loop, and the third is the coupler transmission with cross-coupling.

In the second pathway, light is cross-coupled on the first encounter with the coupler, travels anti-clockwise round the loop, and passes through the coupler without cross-coupling on the second encounter. Since the phase accumulated in the loop is independent of direction (except under the special circumstances discussed in Chapter 14), the resulting amplitude is:

\[ A_{r1} = -j \sin(\kappa L) \exp(-j\phi) \cos(\kappa L) \]  

Clearly, the total reflected amplitude \( A_r \) must be the sum of these contributions, so that:

\[ A_r = -j \sin(2\kappa L) \exp(-j\phi) \]  

The reflected power, \( P_r = |A_r|^2 \), is then given by:

\[ P_r = \sin^2(2\kappa L) \]  

Equation 10.6-4 implies that the reflected power can reach 100% if the normalised coupling length \( \kappa L \) is suitable (for example, when \( \kappa L = \pi/4 \)), independent of the length of the loop. At first sight, this seems curious. To check we have not made a mistake, we can evaluate the transmitted power in a similar way. The total transmitted amplitude \( A_t \) is again the sum of two components:

\[ A_t = \cos(2\kappa L) \exp(-j\phi) \]

Hence, the reflected power \( P_t = |A_t|^2 \) is given by:

\[ P_t = \cos^2(2\kappa L) \]  

Comparing Equations 10.6-4 and 10.6-6, we see that the result is self-consistent; the two outputs sum to unity (to be expected, since there is no loss) and the transmission is indeed zero when the reflectivity reaches 100%. The physical principle in use is one of interference - the transmission falls to zero when the two components in Equation 10.6-5 cancel out.

### 10.7 COUPLING BETWEEN DISSIMILAR WAVEGUIDES

Significant coupling can still occur between two entirely different waveguides, provided they are synchronous and their modal fields overlap suitably. This principle can be exploited in a range of different devices, of which the most important are directional coupler filters and polarizers.

#### THE DIRECTIONAL COUPLER FILTER

We shall begin with the directional coupler filter, which is typically formed from two quite dissimilar, parallel dielectric waveguides in close proximity. For example, these might be structurally similar, but have different core widths and refractive indices. Alternatively, they might be quite different in shape - Figure 10.7-1 shows a section through a coaxial fibre
**coupler**, which operates by coupling between a central cylindrical guide and an outer annular guide.

![Cross-section of a coaxial fibre coupler.](image)

**Figure 10.7-1** Cross-section of a coaxial fibre coupler.

We can understand the effect of using two different guides by considering their dispersion characteristics. For example, Figure 10.7-2 shows typical $\omega - \beta$ diagrams for two slab guides. Note that for most of the frequency range, the guides have different $\beta$-values. However, the curves have been arranged to cross (by careful choice of the waveguide parameters) so that at one frequency $\omega_p$ both propagation constants are equal to $\beta_p$. Now, within a narrow range of this frequency, both curves are approximately linear. Similarly, the variations of $\beta_1$ and $\beta_2$ with $\lambda_0$ are roughly linear within a small range of the phase-matching wavelength $\lambda_p$, as in Figure 10.7-3a.

![Typical dispersion curves for two different guides.](image)

**Figure 10.7-2** Typical dispersion curves for two different guides.

We may approximate these curves by:

$$\beta_1 \approx \beta_p + (\lambda_0 - \lambda_p) \frac{\partial \beta_1}{\partial \lambda_0} \quad \text{and} \quad \beta_2 \approx \beta_p + (\lambda_0 - \lambda_p) \frac{\partial \beta_2}{\partial \lambda_0}$$

where $\frac{\partial \beta_1}{\partial \lambda_0}$ and $\frac{\partial \beta_2}{\partial \lambda_0}$ represent the slopes of the dispersion characteristics near $\lambda_p$. This allows us to write the difference in propagation constant $\Delta \beta = \beta_2 - \beta_1$ between the two guides as:

$$\Delta \beta = (\lambda_0 - \lambda_p) \left\{ \frac{\partial \beta_2}{\partial \lambda_0} - \frac{\partial \beta_1}{\partial \lambda_0} \right\}$$

Consequently, $\Delta \beta$ depends linearly on the difference in wavelength from $\lambda_p$. Without doing any more mathematics, we can deduce that the variation of the cross-coupled output $P_2$ with wavelength will be as previously defined for a switched coupler in Equation 10.5-4, but assuming that $\Delta \beta$ is as given above. A typical response is shown plotted in Figure 10.7-3b. It is clear that the device functions as a band-pass filter, cross-coupling significant power only in the region of the synchronous wavelength $\lambda_p$. The filter bandwidth is set by the size of the proportionality constant $\left\{ \frac{\partial \beta_2}{\partial \lambda_0} - \frac{\partial \beta_1}{\partial \lambda_0} \right\}$. For maximum selectivity, the difference in slopes of the two dispersion characteristics should be large. Unfortunately,
with the relatively weak guides in common use, there are restrictions in the possible range of slopes. This in turn limits typical filter bandwidths to \( \approx 100 \text{ Å} \) using Ti:LiNbO\(_3\) devices.

![Figure 10.7-3 Directional coupler filter operation: a) dispersion characteristics, and b) filter response.](image)

**POLARIZERS AND POLARIZATION SPLITTERS**

The principle of codirectional coupling between two dissimilar guided modes can also be adapted to provide polarizing and polarization-splitting functions, if one of the modes concerned is a surface plasma wave (which we previously introduced in Chapter 5). Surface plasmon polarizers represent a useful alternative to the birefringent devices described in Chapter 9.

We recall that a surface plasmon is a particular field structure that may be supported at the interface between a metal and a dielectric. We have already shown that it can only exist for TM-polarized light, and that is an extremely lossy, short-range mode. Let us suppose that by appropriate choice of parameters its propagation constant may be matched to that of a mode in a nearby parallel dielectric waveguide. Synchronous coupling between these modes may then occur. However, energy transferred to the surface plasma wave will be quickly dissipated as heat due to ohmic loss, so the coupling process will be largely one-way. In this way, the TM mode in the dielectric guide will be heavily attenuated, while the TE mode (which has no plasmon mode to couple to) is virtually unaffected. Such a device therefore acts as a TE-pass polarizer.

![Figure 10.7-4 shows the construction of a surface plasmon fibre polarizer.](image)

It is based on a polished fibre block, of the type previously used to fabricate a fibre directional coupler. However, there are two additional overlayers: a thin metal layer, and a dielectric loading layer. In fact, the surface plasmon mode is supported by a three-layer structure, comprising the two layers just mentioned, and an additional layer provided by the polished block.
Consequently, the form of the plasmon mode is rather more complicated than our previous analysis suggests. Furthermore, detailed analysis shows that more than one such mode can exist. However, the principle described above is nonetheless valid: by careful choice of the layer parameters, it is possible to phase-match one of the possible surface plasma waves to the TM fibre mode, which can then be heavily attenuated. Typical results show 50 - 60dB extinction ratios, with excess losses for the transmitted polarization of only $\approx 0.5 \, \text{dB}$.

![Diagram of a fibre polarizer based on a polished fibre block.](image)

Figure 10.7-4 A fibre polarizer based on a polished fibre block.

By placing another uncoated, polished fibre block on top of such a polarizer, a polarization-splitter can be made. This operates by using the surface plasmon as an intermediary between the two fibres. Now, we have already seen how TM light from the original fibre can be coupled into the plasmon mode. If the additional fibre is also synchronous, the plasmon can in turn couple light into this guide, effectively transferring light between the two different fibres. Since the plasmon now acts only as a 'stepping stone' for a short period, little power is dissipated in the process. Naturally, only TM light is transferred in this way; TE light passes through without coupling.

### 10.8 SIDELOBE SUPPRESSION USING TAPERED COUPLING

We now consider a slight modification to the basic directional coupler design, which can lead to significantly improved performance. Returning to Figures 10.5-1 and 10.7-3b, we notice a characteristic of both responses: the existence of significant sidelobes. This can be unsatisfactory, since switched states must be accurately located in the sidelobe 'nulls' for good crosstalk performance. Figure 10.8-1 shows an alternative coupler design, which avoids this problem. The modification introduced is simply that the interguide separation is made to be slowly varying with distance, reaching a minimum at the device centre and increasing at either end.

![Diagram of a directional coupler with tapered coupling due to varying interguide separation.](image)

Figure 10.8-1 Directional coupler with tapered coupling due to varying interguide separation.

What is the effect of this slow tapering? Well, we already know that the coupling coefficient $\kappa$ depends on $g$, so it is reasonable to assume that a spatially-varying $g(z)$ will lead to a varying $\kappa(z)$. Since $\kappa$ depends roughly exponentially on the gap, a parabolic change in $g$ (due to circularly curved guides) would lead to a Gaussian variation in $\kappa$. In this case, we might expect that the coupled mode equations for the device will be as Equations 10.3-13,
but with the constant $\kappa$ replaced by $\kappa(z)$. For simplicity, we shall just consider the equation for $A_2$, which becomes:

$$\frac{dA_2}{dz} + j \kappa(z) A_1 \exp(+j\Delta\beta z) = 0$$

10.8-1

Now, if we knew the form of $A_1(z)$, this equation could be integrated over the device length, to get:

$$A_2(L) - A_2(0) = -j \int_0^L \kappa(z) A_1(z) \exp(+j\Delta\beta z) \, dz$$

10.8-2

Without solving the equations, we do not know $A_1(z)$. However, we can make an approximation, valid for weakly-coupled devices: if little power is transferred, we may take $A_1(z)$ to be virtually constant. For unity input to Guide 1, we then get:

$$A_2(L) \approx -j \int_0^L \kappa(z) \exp(+j\Delta\beta z) \, dz$$

10.8-3

Equation 10.8-3 now shows that there is an integral relationship between the output $A_2(L)$ and $\kappa(z)$. For constant coupling, we may evaluate this very simply, to get:

$$A_2(L) = -jkL \exp(+j\Delta\beta L/2) \text{sinc}(\Delta\beta L/2)$$

10.8-4

This implies that the cross-coupled output power $P_2$ is given by:

$$P_2 \approx \nu^2 \text{sinc}^2(\xi)$$

10.8-5

where $\nu = \kappa L$ and $\xi = \Delta\beta L/2$, as before.

Equation 10.8-5 could of course have been found from the exact solution (10.5-4) merely by assuming a small value of $\nu$. However, this derivation shows an interesting feature: the amplitude response (Equation 10.8-3) looks very like a Fourier transform of the coupling distribution. The sinc function obtained in Equation 10.8-4 therefore arises from having uniform coupling over a finite distance, i.e. because $\kappa(z)$ is a rectangle function. To suppress the sidelobes, we should choose alternative variations for $\kappa(z)$ whose Fourier transforms have lower lobes. One example is the Gaussian variation described earlier, since this transforms to another Gaussian. If the coupling is tapered in this way, the levels of the sidelobes may be greatly reduced. Most usefully, it turns out that this principle is still valid for strongly coupled devices, even though the relation between the switch response and the coupling variation is no longer a simple Fourier transform.

### 10.9 THE REFLECTION GRATING FILTER - BASIC PRINCIPLES

We now move on to consider another extremely useful device, which operates this time by contradirectional coupling - the reflection grating filter. This is very similar to the planar waveguide gratings described in qualitative terms in Chapter 7. It can be made by etching a periodic corrugation on the surface of a waveguide, as shown in Figure 10.9-1.

The device can be used to convert a mode travelling in the forward direction into a backward-travelling mode, through the Bragg effect. Each groove acts like a weak mirror, which would individually give only a very small reflection. However, if the components all
sum constructively, the result can be a very strong combined reflection. For this to occur, we require the components reflected from adjacent grooves to add in-phase.

![Schematic of a corrugated reflection grating](image)

Figure 10.9-1 a) Schematic of a corrugated reflection grating, and b) electron microscope view of a reflection grating on a ridge guide formed in InP (photograph courtesy G.Heise, Siemens AG).

Now, the optical path between grooves is \( n_{\text{eff}} \Lambda \), where \( n_{\text{eff}} \) is the effective index and \( \Lambda \) is the grating wavelength. If a component reflected from one groove is to sum in-phase with the contribution of its neighbour, we require that twice this path is a whole number of optical wavelengths (the factor of two arises from the reversal in direction). We can write this as:

\[
2n_{\text{eff}} \Lambda = m \lambda_0
\]

10.9-1

Where \( m \) is an integer. This is a special case of Bragg’s Law, and if we take \( m \) to be unity, the guided wave is incident at the first Bragg condition. We shall assume this is the case from now on.

Now, with this assumption we can clearly re-arrange Equation 10.9-1 as:

\[
2 \left( \Lambda/2\pi \right) = \left( \lambda_0/2\pi n_{\text{eff}} \right)
\]

10.9-2

As in Chapter 7, we now define a grating vector \( \mathbf{K} \), whose direction is normal to the corrugations, and whose magnitude is given by:

\[
\mathbf{K} = 2\pi/\Lambda
\]

10.9-3

Bearing in mind that the propagation constant \( \beta \) is equal to \( 2\pi n_{\text{eff}} f/\lambda_0 \), Equation 10.9-2 implies that:

\[
\mathbf{K} = 2\beta
\]

10.9-4

Thus, when the Bragg condition is satisfied, the magnitude of the grating vector is exactly twice \( \beta \). This corresponds to the K-vector closure condition discussed in Chapter 7, and could be represented by a K-vector diagram. However, an alternative viewpoint is provided by Figure 10.9-2.

This shows the dispersion characteristics of both a forward- and a backward-travelling mode; clearly, the latter merely has a negative value of \( \beta \). Here Equation 10.9-4 is represented by the dashed-line construction, which shows that the forward and backward modes are phase-matched by the grating vector for incidence exactly at the phase-matching frequency \( \omega_p \).
Figure 10.9-2 Phase matching by a grating, illustrated on the $\omega$–$\beta$ plane.

DESIGN EXAMPLE

We can work out the period needed to satisfy the Bragg condition for a typical design, as follows. We take the optical wavelength to be $\lambda_0 = 1.5 \, \mu m$, a common near-infrared wavelength. We assume that the guides are made in titanium-indiffused LiNbO$_3$ (whose refractive index is $\approx 2.2$). Because guides made by indiffusion are very weak, we may take $n_{\text{eff}} \approx n_s$, so $\beta \approx 2\pi n_s / \lambda_0 = 9.2 \times 10^6$. Hence $K = 2\beta = 18.4 \times 10^6$, and $\Lambda = 2\pi/K = 0.34 \, \mu m$.

The small figure for $\Lambda$ found above implies that gratings are hard to make. Generally, a surface pattern is first prepared, either by electron beam lithography or by exposure to a holographic interference pattern. This is then etched into the guide using ion beam milling, reactive ion etching or wet etching. In semiconductor materials, the refractive index is generally higher, which results in an even smaller value of $\Lambda$. The required pattern is then so hard to generate that a different grating with period $2\Lambda$ or $3\Lambda$ is often used instead, with incidence at a higher Bragg condition.

OFF-BRAGG INCIDENCE

Now, following the argument above, we would expect a strong reflection from the grating when Equation 10.9-4 holds, because the scattered components will add constructively. What happens if this is not the case (if we change the optical wavelength, for example)? Well, there will still be scattered components, but these will now sum with different phases. Clearly, the further we are off-Bragg, the greater the differences involved, and there will soon come a point when they are so large that the effective sum of the components is zero. At this point, there is no net reflection. The grating is therefore wavelength-selective, and can be used as a filter.

10.10 THE REFLECTION GRATING FILTER - THEORETICAL ANALYSIS

We can also analyse gratings using scalar coupled mode theory. As with the coupler, we first briefly consider the properties of the unperturbed guide (i.e. the guide without the grating). This is again oriented in the $z$-direction, and is formed by a two-dimensional index distribution $n(x, y)$, so wave propagation will be governed by a scalar wave equation similar to Equation 10.3-1 (ignoring the subscript $i$). As before, we assume that the guide is single-moded, so a forward-travelling mode $E_y(x, y, z) = E(x, y) \exp(-j\beta z)$ will satisfy a waveguide equation similar to Equation 10.3-3, as does a backward-travelling mode $E_y(x, y, z) = E(x, y) \exp(+j\beta z)$. We now assume that the guide is perturbed by a periodic index change $\delta n(x,
y, z), which represents the grating. In the simplest model, $\delta n$ could be defined by a periodic function of the form:

$$
\delta n(x, y, z) = \Delta n(x, y) \cos(Kz)
$$

Here $\Delta n(x, y)$ describes the variation of the index perturbation in the transverse direction, while the cosine accounts for the periodicity of the grating in the $z$-direction. The scalar wave equation for the complete system (i.e. guide plus grating) is then:

$$
\nabla^2 E_y(x, y, z) + [n(x, y) + \delta n(x, y, z)]^2 k_0^2 E_y(x, y, z) = 0
$$

Given that the perturbation is likely to be small, we can approximate Equation 10.10-2 by:

$$
\nabla^2 E_y + [n^2 + 2n\delta n] k_0^2 E_y = 0
$$

Now, we know from the qualitative discussion above that a guided mode incident in the $+z$ direction will be reflected as a mode travelling in the $-z$ direction, provided the Bragg condition is nearly satisfied, so that $K \approx 2\beta$. We therefore assume a solution to Equation 10.10-3 as the sum of a forward- and a backward-travelling mode, in the following form:

$$
E_y(x, y, z) = A_F(z) E(x, y) \exp(-j\beta z) + A_B(z) E(x, y) \exp[-j(\beta - K)z]
$$

Here $A_F$ and $A_B$ are the amplitudes of the forward- and backward-modes, respectively, and their $z$-dependence accounts for the gradual conversion of one mode into the other. Note that the backward mode can be recognised as varying approximately as $\exp(+j\beta z)$, if $K \approx 2\beta$. Doing the necessary differentiation, we can obtain the Laplacian as:

$$
\nabla^2 E_y = [A_F \nabla^2 x y E + (d^2 A_F/dz^2 - 2j\beta dA_F/dz - \beta^2 A_F) E] \exp(-j\beta z) + [A_B \nabla^2 x y E + (d^2 A_B/dz^2 - 2j(\beta - K)dA_B/dz - (\beta - K)^2 A_B) E] \exp[-j(\beta - K)z]
$$

We now substitute this and the assumed solution into the wave equation (10.10-3). This produces a very lengthy equation, which can however be simplified using much the same techniques as we have used before. We first eliminate all the $\nabla^2 x y$ terms, using the unperturbed waveguide equation, and neglect second derivatives of $A_F$ and $A_B$ (on the grounds that these vary slowly). This gives:

$$
\{-2j\beta dA_F/dz \exp(-j\beta z) + [-2j(\beta - K)dA_B/dz - \{(\beta - K)^2 - \beta^2\}A_B] \exp[-j(\beta - K)z] + 2n\Delta n k_0^2 \cos(Kz) [A_F \exp(-j\beta z) + A_B \exp[-j(\beta - K)z]] \}E = 0
$$

We then note that the cosine term above can be written as:

$$
\cos(Kz) = (1/2) [\exp(jKz) + \exp(-jKz)]
$$

With this substitution, we can write:

$$
\cos(Kz) \{A_F \exp(-j\beta z) + A_B \exp[-j(\beta - K)z] = (1/2) \{A_F \exp[-j(\beta - K)z] + A_F \exp[-j(\beta + K)] + A_B \exp[-j(\beta - 2K)] + A_B \exp(-j\beta z)\}
$$

Notice that terms containing exponentials like $\exp[-j(\beta + K)]$ and $\exp[-j(\beta - 2K)]$ have not appeared before, and are not contained in our assumed solution (Equation 10.10-4). In fact, these represent higher diffraction orders of the grating, and we can safely ignore them here.
because they will not be phase-matched. With this approximation, Equation 10.10-6 reduces to:

\[
\{-2j\beta \frac{dA}{dz} \exp(-j\beta z) - [2j(\beta - K) \frac{dA}{dz} + \{(\beta - K)^2 - \beta^2\} A_B] \exp[-j(\beta - K)z] + n\Delta n k_0^2 [A_F \exp[-j(\beta - K)z] + A_B \exp(-j\beta z)] \} E = 0
\]

10.10-9

Now, if Equation 10.10-9 is to be true for all \( z \), the coefficients of \( \exp(-j\beta z) \) and \( \exp[-j(\beta - K)z] \) must vanish separately. We therefore equate them both with zero, to get the following equations:

\[
\{-2j\beta \frac{dA}{dz} + n\Delta n k_0^2 A_B \} E = 0
\]
\[
\{-2j(\beta - K)dA_B/dz - \{(\beta - K)^2 - \beta^2\} A_B + n\Delta n k_0^2 A_F \} E = 0
\]

10.10-10

Further approximations are also possible. Because the grating will only have a significant effect if we are close to the Bragg condition, we can put:

\[
\beta - K/2 = \Delta\beta
\]

10.10-11

Where \( \Delta\beta \) is a small quantity. Consequently, we can write:

\[
(\beta - K)^2 - \beta^2 \approx -4\beta\Delta\beta
\]

10.10-12

We will use this in the lower equation in 10.10-10. We also need an approximation for \( \beta - K \); a reasonable substitution is simply to put \( \beta - K \approx -\beta \). If this is done, Equations 10.10-10 become:

\[
\{dA_F/dz + j \left( k_0^2/2\beta \right) n\Delta n A_B \} E = 0
\]
\[
\{dA_B/dz - j2\Delta\beta A_B - j \left( k_0^2/2\beta \right) n\Delta n A_F \} E = 0
\]

10.10-13

These look very like coupled differential equations once again, apart from the \( x \)- and \( y \)-variations contained in the terms \( n(x, y), \Delta n(x, y) \) and \( E(x, y) \). We can get rid of these by multiplying by \( E^* \) and integrating over the waveguide cross-section. This gives:

\[
dA_F/dz + j\kappa A_B = 0 \quad \text{;} \quad dA_B/dz - j2\Delta\beta A_B - j\kappa A_F = 0
\]

10.10-14

These are now true coupled mode equations; once again, the rate of change of each mode amplitude depends on the amplitude of the other mode, via a coupling coefficient \( \kappa \), given by:

\[
\kappa = (k_0^2/2\beta) < n\Delta n E , E > / < E , E >
\]

10.10-15

Notice how \( \kappa \) depends on an overlap between the index perturbation \( \Delta n \) and the transverse field \( E \). If we put the grating in the wrong place, therefore, the field will simply not see it, and there will little reflection. In the Figure 10.9-1, the corrugation is at the upper surface of the guide. Consequently, the grating will only affect the evanescent part of the field, where the amplitude is weak. This will give a low value of \( \kappa \), so a long grating will be needed for high reflectivity.

### 10.11 SOLUTION OF THE EQUATIONS AT SYNCHRONISM

As with the directional coupler, we will start by investigating the solution of the equations at synchronism (i.e. when \( \Delta\beta = 0 \), so that \( K = 2\beta \)). The coupled mode equations are then:
\[
\frac{dA_F}{dz} + j\kappa A_B = 0 \quad ; \quad \frac{dA_B}{dz} - j\kappa A_F = 0
\]

10.11-1

Notice how similar these are to the equations for a synchronous coupler (10.4-1) - the only real difference is a sign change in the lower equation. We can therefore solve them in a very similar way. We start by differentiating the upper equation, and then eliminating \(A_B\) using the lower equation. This gives a second-order equation for \(A_F\), in the form:

\[
\frac{d^2A_F}{dz^2} - \kappa^2 A_F = 0
\]

10.11-2

Because of the sign change mentioned above, solutions of this equation are not trigonometric functions; instead, they are hyperbolic. We may therefore put:

\[
A_F = D_1 \cosh(\kappa z) + D_2 \sinh(\kappa z)
\]

10.11-3

Here \(D_1\) and \(D_2\) are constants, chosen to satisfy the boundary conditions. With a forward-going input, \(A_F\) is clearly specified at \(z = 0\). However, the backward wave must grow from zero at the far end of the device, which is of length \(L\). Suitable boundary conditions are therefore:

\[
A_F = 1 \quad \text{at} \quad z = 0, \quad A_B = 0 \quad \text{at} \quad z = L
\]

10.11-4

It is then quite simple to show that the solutions must be:

\[
A_F(z) = \cosh[\kappa(L - z)] / \cosh(\kappa L) \quad ; \quad A_B(z) = -j \sinh[\kappa(L - z)] / \cosh(\kappa L)
\]

10.11-5

Notice that the modal amplitudes at any point inside the grating now depend on the precise value of \(\kappa L\). Now, the power carried in the +z direction by the forward wave is clearly \(P_F = A_F A_F^*\). However, the power carried in the same direction by the backward wave is \(-P_B\), where \(P_B = A_B A_B^*\). Once again, it can be shown that power is conserved by the system of equations, since \(P_F - P_B = \text{constant}\).

Figure 10.11-1 shows the variation of \(P_F\) and \(P_B\) with distance through the grating, for a typical normalised thickness of \(\kappa L = 1.5\). As can be seen, the incident wave decays as it travels through the grating. However, as it loses energy, the reflected wave grows.

![Figure 10.11-1 Variation of reflected and transmitted powers with distance.](image)

The separation of the two curves is constant, consistent with the power conservation relation above. In general, we will actually be most interested in the two outputs from the grating.
These are the grating transmissivity $T$ and reflectivity $R$, defined as $T = A_f A_f^*$ at $z = L$ and $R = A_h A_h^*$ at $z = 0$. It is easy to show these are given by:

$$T = \frac{1}{\cosh^2(\kappa L)} \quad \text{and} \quad R = \tanh^2(\kappa L) \quad 10.11-6$$

Equation 10.11-6 implies that 100% peak reflectivity is possible, simply provided the normalised coupling length $\kappa L$ is large enough. For example, if $\kappa L = 2$, we obtain $R = 0.93$; similarly, for $\kappa L = 4$, we get $R = 0.999$. Figure 10.11-2 shows a plot of $T$ and $R$ versus $\kappa L$. Power conservation is again demonstrated, since the sum of $T$ and $R$ is unity throughout.

![Figure 10.11-2 Grating reflectivity and transmissivity versus normalised length.](image)

**10.12 ASYNCHRONOUS SOLUTION**

Now we consider the case when the optical wavelength is slightly incorrect, so that $\Delta \beta \neq 0$. Once again, the solutions can be found by quite straightforward methods at each of the outputs as:

$$T = \left| \frac{1}{\left[ \cosh(\Psi) - j \left( \frac{\xi}{\Psi} \right) \sinh(\Psi) \right]} \right|^2$$

$$R = \left| \frac{1}{\left[ \left( \frac{\xi}{\mu} \right) + j \left( \frac{\Psi/\mu}{\coth(\Psi)} \right) \right]} \right|^2 \quad 10.12-1$$

Where $\mu$, $\xi$ and $\Psi$ are defined by:

$$\mu = \kappa L, \quad \xi = -\Delta \beta L \quad \text{and} \quad \Psi = \sqrt{\mu^2 - \xi^2}$$

These solutions can be plotted against $\xi$ (i.e. as a function of the difference in $\beta$ from the design value), for a given normalised thickness. Figure 10.12-1 shows the response for $\kappa L = 3$. As expected, the results are in keeping with our earlier qualitative discussion. For $\xi = 0$, there is high reflectivity. As $|\xi|$ increases, the reflected power falls, with a typical filter response. This is accompanied by a corresponding rise in transmissivity, and for sufficiently large $|\xi|$, there is no significant reflection.

**DESIGN EXAMPLE**

We can estimate the bandwidth of a typical reflection grating filter as follows. For a reasonable range of normalised lengths, the half-power point in curves like Figure 10.12-1 is reached when:
\[ |\xi| \approx 3 \]

So that:
\[ |\Delta\beta| \approx 3/L \]

Since \( \beta = 2\pi n_{\text{eff}}/\lambda_0 \), we can write:
\[ |\Delta\beta| = 2\pi n_{\text{eff}} |\Delta\lambda_0| / \lambda_0^2 \]

And therefore that:
\[ |\Delta\lambda_0| \approx 3\lambda_0^2 / 2\pi n_{\text{eff}}L \]

For the typical parameters used before, namely \( n_{\text{eff}} \approx 2.2 \) and \( \lambda_0 = 1.5 \mu m \), and a grating length of \( L = 1 \text{ mm} \), we obtain \( |\Delta\lambda_0| \approx 0.5 \text{ nm} \). Grating filters are therefore highly wavelength-selective.

**Figure 10.12-1 Typical filter response of a reflection grating.**

### 10.13 FIBRE GRATINGS

We have already described how gratings can be made in integrated optics. They can also be made in fibre optic form, by etching a corrugation over a fibre that has been polished back to expose its core. Figure 10.13-1 shows how this is done, starting with one half of a polished fibre coupler. A fibre grating might be used as a frequency selective mirror in a fibre laser, or as a channel selector in a wavelength-division multiplexed (WDM) communications system. Notice that the distance from the fibre core to the grating corrugation varies through the device, due to the curvature of the fibre. Consequently, the coupling coefficient must also vary. It can be shown that tapered coupling of this type has much the same effect as was found in the directional coupler, so we might expect the sidelobes of the filter response to be suppressed.

**Figure 10.13-1 Optical fibre grating based on half a polished fibre coupler.**
More recently, an alternative method of grating fabrication has been investigated, based on the inherent photosensitivity of Ge-doped silica to short-wavelength radiation. This allows a periodic change in refractive index to be obtained inside the fibre core itself, rather than in the cladding. The fibre is simply exposed to an optical interference pattern, either derived from two coherent, counter-propagating guided waves or from two external beams. In the latter case, a grating period suitable for reflection of a guided infrared beam may be obtained, allowing the construction of filters for use in fibre lasers. Figure 10.13-2 shows a typical set-up for external recording. The exposure source is normally an Ar+ laser, operating either at $\lambda = 0.488 \mu m$ or at the frequency-doubled wavelength of $0.244 \mu m$. Although the work is at an early stage, and the index change obtained is rather weak ($\Delta n \approx 6 \times 10^{-5}$), high-efficiency gratings have already been produced by this method.

![Figure 10.13-2 Writing and detection of a reflection grating in germanosilicate fibre (after Kashyap R. et al., Elect. Lett. 26, 730-732 (1990))](image)

**10.14 OTHER COUPLED MODE INTERACTIONS**

It turns out that the directional coupler and the reflection grating are merely the simplest examples of a wide range of devices that operate by some combination of co- and contra-directional coupling. We therefore conclude with a few examples illustrating further coupled mode interactions.

**CONVERSION BETWEEN GUIDED AND RADIATION MODES (THE GRATING COUPLER)**

The first example occurs in the grating coupler, which we previously met in Chapter 7 as an input-output coupling component. There, its operation was described in terms of a K-vector diagram, but an alternative view is provided by the $\omega - \beta$ diagram of Figure 10.14-1.

Essentially, this is the same as for a simple reflection grating (Figure 10.9-2), but the frequency has been raised (or the wavelength lowered) so that the left hand-end of the grating vector no longer lies in the region where backward-travelling, guided modes exist; instead, it falls in the area of radiation modes. Consequently, we would expect the grating exactly to phase-match a forward-travelling guided mode to a backward-travelling radiation mode, and that this interaction could be described by a pair of coupled mode equations. This simple picture does not, unfortunately, represent the whole story. We have already seen that significant power can be coupled to modes that are not exactly phase-matched, provided they are reasonably close to synchronism. Since the radiation modes are a continuum, the guided mode will be coupled to an entire group of such modes, whose
propagation constants lie close to the phase-matched value. The longer the grating, however, the more selective it becomes, and the more restricted the group; eventually, the description above must become a reasonable one. When this occurs, the radiated output travels predominantly in a defined direction, which must of course be wavelength-dependent. This allows a grating coupler to be used as a form of spectrometer.

Figure 10.14-1 Phase matching by a grating between guided and radiation modes, illustrated on the $\omega-\beta$ plane.

GRATING-ASSISTED COUPLING

A grating can also be used to achieve phase-matching between two otherwise asynchronous, codirectional guided modes. These might be the two different polarization modes in a single-moded guide formed in a birefringent substrate (e.g. LiNbO$_3$). Phase-matching is obtained when one mode is excited as the first diffraction order of the grating, so that:

$$|K| = \beta_{TE} - \beta_{TM} = 2\pi \frac{\Delta n_{eff}}{\lambda_0}$$

where $\Delta n_{eff}$ is the difference in effective indices of the TE and TM modes. Since $\Delta n_{eff}$ is usually only moderate, $|K|$ is small, and the grating wavelength $\Lambda$ is large ($\approx 7 \mu m$ in Ti:LiNbO$_3$ at $\lambda_0 = 0.6 \mu m$). To obtain coupling between orthogonal modes, a periodic perturbation to an off-diagonal element of the dielectric tensor is needed. This cannot be achieved using a simple surface corrugation, but it can be induced electrically through an off-diagonal electrooptic coefficient, by using a periodic electrode as shown in Figure 10.14-2.

Figure 10.14-2 A TE-TM mode converter filter.

This type of device is known as a **TE-TM mode converter**, and can be used as a polarization rotator or a filter. Filter operation depends on the difference in dispersion of
the two modes, as for the dissimilar-guide directional coupler described earlier. However, greater wavelength selectivity ($\approx 10$ Å in Ti:LiNbO$_3$) can be obtained in this case, because larger differences in slope of the two dispersion characteristics are possible with orthogonal modes.

It is not, however, necessary for the two modes concerned to exist in the same guide. If a grating is combined with a directional coupler, it may be used to phase-match codirectional modes of the same polarization in the two guides (if they are dissimilar) or orthogonal polarizations (if they are the same). Equally, the two modes do not have to be codirectional. Figure 10.14-3 shows a device which acts as a track-changing reflector, by using a reflection grating to phase-match contradirectional modes in a fibre coupler formed between entirely different guides.

Without the grating, any codirectional coupling would be asynchronous, as shown in the $\omega$ - $\beta$ diagram of Figure 10.14-4. Similarly, the grating period is chosen so that it cannot retroreflect light in either Guide 1 or Guide 2. However, at the correct frequency, it can phase-match a forward mode in Guide 1 to a backward mode in Guide 2. This occurs when:

$$|K| = \beta_{g1} + \beta_{g2}$$

10.14-2

The result is that the reflected output now emerges from a different fibre to that used for the input - this is particularly convenient for channel-dropping operations.

Figure 10.14-4 Phase-matching by a grating between forward- and backward-travelling modes in different guides, as illustrated on the $\omega$–$\beta$ plane.
PROBLEMS

10.1. Discuss the approximations inherent in the derivation of the coupled mode equations for a directional coupler. Are they valid? If not, can you suggest the likely consequences for device performance?

10.2. An alternative view of a synchronous directional coupler can be based on the well-known optical principle that the output from a linear device is given by the sum of all possible routes between input and output.

Figure (i) below shows one path between an input to guide 1 and an output from guide 2. This involves a single scattering at an arbitrary position z, such that 0 ≤ z ≤ L. If the scattering amplitude per unit length is -jκ, and the input amplitude is A_{10}, the net contribution from all such paths is:

\[ A_{2L} = -jκ A_{10} \int_0^L dz = -jκL A_{10} \]

Figure (ii) below shows another possible route, involving three scatterings.

(a) If the sum of all such paths is denoted by A_{2L}^3, evaluate A_{2L}^3.
(b) Draw the paths contributing to A_{2L}^5, which has five scatterings, and evaluate the integral for this case.
(c) Sum the series A_{2L} = A_{2L}^1 + A_{2L}^3 + A_{2L}^5 ..., and show that it corresponds to the normal solution for a synchronous directional coupler.

10.3 A directional coupler is constructed from two identical, symmetric slab waveguides. The guide thickness is h, and the guide and substrate indices are n_g and n_s, respectively. Using the methods outlined in Chapter 6, derive a solution for the transverse field of the lowest-order, symmetric TE mode of a single such guide. Then, by direct evaluation of the overlap integral concerned, show that the coupling coefficient κ decreases exponentially with the interguide separation g in the twin-guide device.

10.4 The coupled differential equations for an asynchronous directional coupler are given by:

\[ \frac{dA_1}{dz} + jκ A_2 \exp(-jΔβz) = 0 \quad \frac{dA_2}{dz} + jκ A_1 \exp(+jΔβz) = 0. \]

Assuming that total power is defined as \( P = |A_1|^2 + |A_2|^2 \), show (without solving the equations) that power is conserved by the device.
10.5. From a series of experiments with Ti:LiNbO$_3$ directional couplers made with different interwaveguide gaps $g$, the following data were obtained for the length required for complete coupling: $g = 4 \, \mu m; \, L_{\pi/2} = 5 \, mm; \, g = 6 \, \mu m; \, L_{\pi/2} = 7 \, mm; \, g = 7.5 \, \mu m, \, L_{\pi/2} = 9 \, mm$. Estimate the length required to achieve 50% coupling efficiency when $g = 7 \, \mu m$.

[4.14 mm]

10.6 A fibre directional coupler is to be made by polishing two glass blocks, which have fibres embedded in curved grooves. (a) Suggest how the polishing process can be monitored, so it can be terminated when the core is reached. (b) When the coupler is assembled, it is found that the amount of cross-coupled light is very small, despite attempts to design it for 100% coupling. Suggest two possible explanations, and the remedy in each case.

10.7 The output of a synchronous reflection grating can also be found as the sum of all possible routes between the input and output. Figure (i) below shows one path between a forward-going input and a backward-travelling output. This involves a single scattering at an arbitrary position $z$, such that $0 \leq z \leq L$. If the scattering amplitude per unit length is $-j\kappa$, and the input amplitude is $A_{F0}$, the net contribution from all such paths is:

$$A_{B0} = -j\kappa A_{F0} \int_0^L dz = -j\kappa L A_{F0}$$

Figure (ii) below shows another possible route, involving three scatterings.

(a) If the sum of all such paths is denoted by $A_{B0}^3$, evaluate $A_{B0}^3$.

(b) Draw the paths contributing to $A_{B0}^5$, which has five scatterings, and evaluate the integral for this case.

(c) Sum the series $A_{B0} = A_{B0}^1 + A_{B0}^3 + A_{B0}^5 + \ldots$, and show that it corresponds to the normal solution for a reflection grating, when replayed at the Bragg wavelength.

10.8 The coupled differential equations for a reflection grating at Bragg incidence are given by:

$$\frac{dA_F}{dz} + j\kappa A_B = 0 \quad ; \quad \frac{dA_B}{dz} - j\kappa A_F = 0$$

Verify that the general solutions to these equations, subject to the boundary conditions $A_F = 1$ at $z = 0$, $A_B = 0$ at $z = L$, are indeed:

$$A_F(z) = \cosh[\kappa(L - z)] / \cosh(\kappa L) \quad ; \quad A_B(z) = -j \sinh[\kappa(L - z)] / \cosh(\kappa L).$$
10.9 A Bragg reflector is to be used with an integrated optic channel waveguide. Past measurements with this system suggest that the coupling coefficient is $\kappa = 200 \, /m$ at the design wavelength. How long should the grating be, for $\geq 99\%$ reflectivity? [15 mm]

10.10 Compare the wavelength selectivity of waveguide filters based on (a) codirectional coupling between dissimilar waveguides, and (b) contradirectional coupling via a grating. What are the factors limiting selectivity in each case?
SUGGESTIONS FOR FURTHER READING

Alferness R.C. Veselka J.J. "Simultaneous modulation and wavelength multiplexing with a tunable Ti:LiNbO₃, directional coupler filter" Elect. Lett. 21, 466-467 (1985)
11.1 WAVE PARTICLE DUALITY

In Chapter 3, we described the interaction of light and matter in terms of electric and magnetic fields imposing forces on discrete charges. These charges - electrons - were considered to be either basically free (as in a metal) or bound (as in a dielectric). Furthermore, the dynamics of the interaction were explained in classical terms, so that the electron was regarded as a particle with charge and mass, obeying the laws of mechanics. There is considerable evidence to support this view - for example, measurements of the charge and charge-to-mass ratio of an electron - and it has lead to a variety of hugely successful devices, many still of great importance (the oscilloscope, the TV tube, and high-power, high-frequency valves). Equally, the model leads to a perfectly adequate explanation of many optical phenomena - for example, the slowing of a wave in matter. This has allowed us to analyse many guided wave optical devices merely by assuming different distributions of refractive index.

Unfortunately, light displays some features beyond the scope of this model. So far, we have treated optical waves purely as solutions of Maxwell's equations. This provides a perfectly good explanation of the propagation of electromagnetic energy on a large scale, but it is seriously inadequate for the description of energy transfer between the electromagnetic field and matter. The need for a different analysis was first proposed by Max Planck in 1900, in connection with black body radiation. Planck found that the wavelength distribution of the energy radiated by a body in thermal equilibrium could only be explained if the energy exchanged between the field and the body (idealised as a collection of oscillators) took place in discrete amounts, comprising an integral number of elementary units known as quanta. He also showed that the magnitude of these quanta must be $hf$, where $h$ is a constant, subsequently called Planck's constant (with the value $6.62 \times 10^{-34}$ J s), and $f$ is the frequency. The quanta of visible radiation have become known as photons. Further support for this idea came later through observations by P. von Lenard of the photoelectric effect, where the illumination of a metal surface under vacuum was found to result in the ejection of electrons. Here, the critical feature was that the threshold of electron emission was determined by the frequency of the radiation and not by its power. This effect was explained completely by Albert Einstein in 1905.

Black body radiation and the photoelectric effect both relate to the interaction of radiation with matter. We might argue that a classical electromagnetic picture is still valid in the absence of absorption or emission effects. In fact, even this cannot really be justified. The first evidence came with an analysis of the scattering of X-rays by electrons, by A.H.Compton in 1923. Compton showed that there is a change in the X-ray wavelength with scattering angle; this is only consistent with a billiard-ball type of collision, in which the X-rays are represented by quanta of energy $\hbar \omega$, carrying momenta of $\hbar k$. At this point, we introduce the notation $\hbar = h/2\pi$, which allows us to write:

$$E = \hbar \omega \quad \text{and} \quad p = \hbar k$$

It is therefore clear that light actually possesses some of the attributes of both waves and particles, a property known as wave-particle duality. In practice, the relative importance of each facet of behaviour depends on the circumstances. As mentioned earlier, a wave analysis
is often adequate, but the quantum nature of radiation should be invoked if any energy interchanges take place.

It is also a reasonable guess that if a quantity we have previously considered to be a wave may have particle-like properties, something described earlier as a particle may display wave-like behaviour. This turns out to be true: electrons also share the properties of both particles and waves, and their wave nature is important in explaining their behaviour in crystalline solids, especially the semiconductor materials used for optoelectronic devices. If we regard the electron as being similar to a photon, so that in each case we are dealing with a localised, particle-like entity, we might say that the momentum of an electron \( p_{\text{electron}} \) corresponds to \( p_{\text{photon}} \), while its energy \( E_{\text{electron}} \) is analogous to \( E_{\text{photon}} \). In general, the electron energy will the sum of kinetic and potential contributions, so that:

\[
E = \frac{1}{2} mv^2 + \mathcal{V} \tag{11.1-2}
\]

where \( m \) is the electron mass, \( v \) is its velocity and \( \mathcal{V} \) is its potential energy. Naturally, since \( p = mv \), we could write this alternatively as:

\[
E = \frac{p^2}{2m} + \mathcal{V} \tag{11.1-3}
\]

We now take one more step and say that if a photon has momentum and energy while at the same time having angular frequency \( \omega \) and propagation constant (or wavenumber) \( k \), it is possible that an electron - which clearly has momentum and energy - should also have a frequency and wavelength, i.e. wave-like characteristics associated with a particle-like model. If this were true, we could introduce an electron wavenumber and angular frequency, following:

\[
\hbar k_{\text{electron}} = p_{\text{electron}} \quad \text{and} \quad \hbar \omega_{\text{electron}} = E_{\text{electron}} \tag{11.1-4}
\]

Consequently, we might also write:

\[
\hbar \omega = \frac{p^2}{2m} + \mathcal{V} = \hbar \omega k^2/2m + \mathcal{V} \tag{11.1-5}
\]

Equation 11.1-4, first suggested by L. de Broglie in 1924, is the key to the link between the particle and wave nature of electrons. So far, it is just conjecture; the most we can say is that it is plausible, and direct experimental evidence is required for confirmation. This has been provided in many ways. The earliest was the observation of electron diffraction - an interference phenomenon associated only with waves - by C.J. Davison and L.H. Germer in 1927. As we saw in Chapter 7, diffraction can be observed from a periodic structure with a spacing comparable to the wavelength in question. For electrons, we can easily calculate the wavelength for a given energy. When \( \mathcal{V} = 0 \), Equation 11.1-3 implies that:

\[
p = \sqrt{(2mE)} \tag{11.1-6}
\]

so that the equivalent wavelength may be found as:

\[
\lambda = h/p = h/\sqrt{(2mE)} \tag{11.1-7}
\]

Now, for electrons accelerated through \( V \) volts, we have \( E = eV \). For \( V = 1 \) kV, this corresponds to \( \lambda = 0.04 \) nm, and for 10 kV, \( \lambda = 0.012 \) nm. Thus for typical oscilloscope accelerating voltages, we obtain wavelengths roughly corresponding to X-rays. This should result in observable diffraction from the periodic array of atoms in a crystal. In the event,
Davisson and Germer showed that the scattering of electrons from a perfect crystal of nickel was not isotropic, but showed a clear dependence on angle (as expected from Bragg's law). This and other experiments established the validity of the idea of the wave nature of electrons, and the accuracy of the de Broglie relation.

11.2 PHOTONS

We shall begin our discussion of these new ideas by considering the description of radiation in terms of photons. To use a classical analogy, we might regard these localised fields as groups of plane waves, or wavepackets. For propagation in free space, we might write:

\[ E(z, t) = \int_{-\infty}^{\infty} E(\omega) \exp\{j(\omega t - kz)\} \, d\omega \]

for a one-dimensional wavepacket. This will travel at a speed given by the group velocity \( v_g = d\omega/dk \) (which is equal to the phase velocity \( c \) in free space).

The relation between the shape of the packet and its frequency spectrum \( E(\omega) \) is determined by the laws of Fourier transformation. For example, Figure 11.2-1 shows a snapshot of a burst of radiation of uniform amplitude and length \( \Delta z \). The corresponding angular frequency spectrum (which is directly analogous to one previously computed in Chapter 2) is also shown. In this case, the angular frequency \( \omega_c \) of the radiation determines the centre frequency of the spectrum and the length \( \Delta z \) of the packet the range of frequencies present. As \( \Delta z \) increases, the range decreases, until in the limit we are left only with a single plane wave of frequency \( \omega_c \).

A typical radiation field will comprise many photons of random lengths, separated by random intervals. The classical and the quantum pictures can then be reasonably brought together, by treating the classical power density as being a measure of the photon density at any place. Except for very low power levels, the photon density is likely to be sufficiently high that the inherent ‘graininess’ of the radiation field is effectively smoothed out (although it will still be responsible for a level of noise).
We must be more careful in our description of a radiation field, however, if it is confined inside a volume. For example, Figure 11.2-2 shows a cube of side \( h \), consisting of material of refractive index \( n \). This structure is effectively a three-dimensional resonant cavity, so we can guess from our earlier analysis that only particular standing-wave patterns may exist inside it. Any localised field we care to postulate - for example, a wavepacket - may then only be constructed from these allowed fields.

To determine the allowed fields, we might solve the time-independent scalar wave equation:

\[
\nabla^2 E(x, y, z) + n^2 k_0^2 E(x, y, z) = 0
\]

subject to the simplified boundary conditions of \( E = 0 \) on all faces of the cube. By analogy with our previous results for a one-dimensional cavity, we might assume the solution:

\[
E = E_0 \sin(k_x x) \sin(k_y y) \sin(k_z z)
\]

Substitution into Equation 11.2-2 shows that this will indeed be valid, provided:

\[
k_x^2 + k_y^2 + k_z^2 = n^2 k_0^2
\]

while the boundary conditions will be satisfied if:

\[
k_x = \nu_x \pi/h \quad ; \quad k_y = \nu_y \pi/h \quad \text{and} \quad k_z = \nu_z \pi/h
\]

where \( \nu_x, \nu_y \) and \( \nu_z \) are integers. Equations 11.2-4 and 11.2-5 imply that the possible values of \( k_0 \) are restricted. In particular, we find that the allowed frequencies are specified by:

\[
\omega = ck_0 = c\nu \pi/\hbar
\]

where \( \nu^2 = \nu_x^2 + \nu_y^2 + \nu_z^2 \). Any localised field may therefore only be constructed as a summation of components at these frequencies, rather than as a continuous integral. Because this is a complicated process, it is now sensible to consider the allowed solutions (or states) as being of significance in themselves. One question we might ask is, how many of them are there? Well, if \( h \) is large, the minimum difference in frequency between any two states is small, so that the allowed frequencies approximate to a continuum. Note that if \( \nu \) is
constant, then so is $\omega$. Consequently, if the allowed frequencies are plotted in a three-dimensional space, with axes $\nu_x$, $\nu_y$, and $\nu_z$, then a sphere of radius $\nu$ represents a surface of constant $\omega$. In this space, the distance between neighbouring states is one unit, when measured in the $x$-, $y$-, or $z$-directions, so the number of states contained in a cube of unit volume is one. The number of states with frequency less than $\omega$ might therefore be taken to be equal to the volume of the sphere, namely $4\pi\nu^3/3$. However, the indices $\nu_x$, $\nu_y$, and $\nu_z$ must all actually be positive for the results to be meaningful, so the true figure should really be $1/8$th of this value. Including the effect of polarization (which doubles the number of states), we finally obtain the result that the number of states with frequency less than $\omega$ is $\pi\nu^3/3$.

However, since:

$$\nu = C \omega$$

where $C = nh/\pi c$, it can easily be seen that:

$$\pi\nu^3/3 = (\pi/3) C^3 \omega^3$$

If we now define this quantity to be equal to the integral of a new function $s(\omega)$, i.e. as $\int_0^{\omega} s(\omega) d\omega$, where $s(\omega)d\omega$ is the number of states in an interval $d\omega$, then we may extract $s(\omega)$ by differentiation, as:

$$s(\omega) = \pi \frac{C^3}{\omega^2}$$

Often, we will be more interested in a function $S(\omega)$, the number of states per unit frequency interval per unit volume (generally known as the density of states). This may be obtained very simply by dividing Equation 11.2-9 by the volume of the cavity $(h^3)$ to get:

$$S(\omega) = \frac{n^3\omega^2}{\pi^2 c^3}$$

Note that this result is entirely independent of the cavity dimension $h$.

We may also care to enquire whether any of these possible states are indeed filled in a practical cavity, i.e. how much radiation exists at a given temperature. Using the methods of statistical thermodynamics can be shown that the probability of a state of frequency $\omega$ being filled at absolute temperature $T$ is:

$$F(\omega) = \frac{1}{\exp(hbar\omega/kT) - 1}$$

where $k$ is Boltzmann's constant. Equation 11.2-11 is known as the Bose-Einstein probability function, and it was originally derived under two assumptions: (1) that radiation exists only in quanta whose energy is a multiple of $hbar\omega$, and (2) that there is no restriction on the occupation of any state. We may use it to deduce the number of filled states lying in a frequency range between $\omega$ and $\omega + d\omega$. If we write this quantity as $n(\omega) d\omega$, then clearly $n(\omega) d\omega = S(\omega) F(\omega) d\omega$, or:

$$n(\omega) d\omega = \left[\frac{n^3\omega^2}{\pi^2 c^3}\right] \left[\frac{1}{\exp(hbar\omega/kT) - 1}\right] d\omega$$

Equally, we may use it to find the energy density of the radiation field in the same range. Since each state of frequency $\omega$ carries energy $hbar\omega$, this is given by $\rho(\omega) d\omega = hbar\omega S(\omega) F(\omega) d\omega$, or:
\[ \rho(\omega) \, d\omega = \left[ \frac{(\hbar n^3/\pi^2 c^3)}{\exp(\hbar \omega/kT) - 1} \right] \, d\omega \]

This is the famous **black-body radiation law** mentioned earlier, and its agreement with experiment went far to establish the quantum theory. As we shall see later, we can also interpret the existence of radiation in terms of the interaction of three processes for the absorption and emission of light. In equilibrium, these balance, so that the radiation density is stabilised at the black body level. However, by manipulating them away from their equilibrium rates, they may be exploited for device applications.

### 11.3 ELECTRONS

We now turn our attention to electrons. Having established their wave nature, it would be natural to try to develop a suitable electron wave equation. In Chapter 2, the wave equation for light arose from Maxwell’s equations. The problem here is that we do not have a corresponding set of ‘electron laws’. However, recall that we used the electromagnetic wave equation to deduce the dispersion characteristic of light. By contrast, we already know the dispersion characteristic of electrons; it must be Equation 11.1-5. An alternative route would therefore be to work backwards from this. We start by plotting the dispersion characteristic in Figure 11.3-1, for free electrons (i.e. with \( V \) = constant, say zero). Note firstly that it is parabolic, and hence quite different from the linear characteristic of light waves, and secondly that the curve might also represent the variation of electron energy with \( k \) (since \( E = \hbar \omega \)).

![Figure 11.3-1](image)

Figure 11.3-1 The dispersion characteristic of an electron.

Remember that the dispersion characteristic provides a relation between \( \omega \) and \( k \) for a travelling wave, which is itself descriptive of the electron in a uniform potential. Now, in our earlier electromagnetic analysis, a \( z \)-propagating plane wave solution for the electric field was written in the vector form:

\[ \mathbf{E}(z, t) = \mathbf{E}_0 \exp\{j(\omega t - kz)\} \]

For an electron wave, we choose a similar type of solution, writing:

\[ \Psi(z, t) = \Psi_0 \exp\{j(\omega t - kz)\} \]
where the quantity $\Psi$ (known as the **wavefunction**) describes some characteristic of electrons that is roughly analogous to the electric field, but is now a scalar. Differentiating with respect to $t$, we find:

$$\frac{\partial \Psi}{\partial t} = j\omega \Psi_0 \exp\{j(\omega t - kz)\} = j\omega \Psi$$  \hspace{1cm} 11.3-3

while differentiating with respect to $z$ gives:

$$\frac{\partial \Psi}{\partial z} = -jk \Psi \quad \text{and} \quad \frac{\partial^2 \Psi}{\partial z^2} = -k^2 \Psi.$$  \hspace{1cm} 11.3-4

We can now form a wave equation from the dispersion relation, by substituting for $\omega$ and $k^2$. If this is done, we get:

$$j\hbar \frac{\partial \Psi}{\partial t} = \hbar \frac{\partial^2 \Psi}{\partial z^2} - V \Psi$$  \hspace{1cm} 11.3-5

We could easily extend this argument to three dimensions, where we would find:

$$j\hbar \frac{\partial \Psi}{\partial t} = \hbar \nabla^2 \Psi - V \Psi.$$  \hspace{1cm} 11.3-6

This equation was first derived by Erwin Schrödinger in 1926, and is known as the **time-dependent Schrödinger wave equation**.

One obvious question is, what does the wave function $\Psi$ mean? Since the Schrödinger equation is complex, it is a reasonable guess that $\Psi$ will also be complex, and therefore of no direct physical significance. In fact, the exact interpretation of $\Psi$ was the cause of some controversy. After much debate, it was concluded that the quantity $\Psi^* \Psi \, dv$ (or $|\Psi|^2 \, dv$) could be taken to represent the probability that an electron will be found in the volume $dv$. This situation is not so different from that occurring in electromagnetic theory. We may feel more familiar and comfortable with the fields $E$ and $H$, and hence tend to regard them as real and observable, but when we want to observe or detect an electromagnetic wave, we have to extract energy from it, and the measure of energy content is proportional to $|E|^2$ (or $|H|^2$). We are therefore inevitably sampling the field intensities, much as we sample $|\Psi|^2$.

Naturally, in order to obtain the realistic result that the probability of finding a given electron somewhere in all space must be unity, the wavefunction must be normalised, so that:

$$\iint_v |\Psi|^2 \, dv = 1$$  \hspace{1cm} 11.3-7

Clearly, the value of $|\Psi|^2 \, dv$ is the same everywhere for the wave of Equation 11.3-2. We therefore conclude that a single electron wave is spread throughout space; in order to model a mobile, localised electron we must again group a number of plane wave components together into a wave-packet. This will travel at the group velocity, which may be found by differentiating the $\varepsilon$ - $k$ characteristic to get:

$$v_g = \frac{d\omega}{dk} = \hbar k/m = \sqrt{2\varepsilon/m}$$  \hspace{1cm} 11.3-8

A moment's thought shows that this is identical to the speed of our earlier particle-like electron, showing that our new model is consistent with the old. Note, however, that the speed of the electron wavepacket is variable, and depends on its energy. Consequently, when we increase the kinetic energy of an electron (e.g. by accelerating it in an electric field), its velocity increases monotonically.
The Schrödinger wave equation thus fills the role played by the electromagnetic wave equation that we have used so far. Although we have based the derivation on a simple travelling wave, we would expect that other solutions will describe electron behaviour under more complicated boundary conditions. For example, we would expect the ‘trapping’ of an electron wave to be possible, since we have already seen that an optical wave may be confined in a waveguide. Here, the details of the guiding structure were introduced through a variation of the refractive index; for an electron, we now have variations in the potential energy caused by some distribution of charge or electric potential.

Before examining this aspect, we adopt a simplification used earlier, namely to eliminate any time-dependence by assuming a harmonic solution. We therefore write:

$$\Psi(x, y, z, t) = \Psi(x, y, z) \exp(i\omega t)$$ \hspace{1cm} 11.3-9

If this is done, the Schrödinger wave equation reduces to its time-independent form:

$$\nabla^2 \Psi(x, y, z) + \left(\frac{2m}{\hbar^2}\right) \left( E - V \right) \Psi(x, y, z) = 0$$ \hspace{1cm} 11.3-10

**ELECTRON IN A POTENTIAL WELL**

We shall now attempt to find solutions to Equation 11.3-10 for a particular situation analogous to an electromagnetic problem considered in Chapter 6, the propagation of an optical wave in a slab guide. This time, however, we will examine the behaviour of an electron wave confined to a one-dimensional *potential well*, and propagating in the z-direction. To describe the well, we will assume a potential energy distribution of the form $V = 0$ for $0 \leq x \leq h$; $V = V_0$ otherwise. The geometry is shown in Figure 11.3-2; this distribution is analogous to the index variation of a symmetric slab guide, if decreased potential is taken to correspond to increased refractive index.

![Figure 11.3-2 Potential energy distribution forming a one-dimensional well.](image)

Clearly, in this geometry there will be no y-variation in the solution, so that $\partial \Psi / \partial y = 0$. The solution we assume in each of the three regions may then be taken in the separated form:

$$\Psi_i(x, z) = \psi_i(x) \exp(-j\beta z)$$ \hspace{1cm} 11.3-11

With this substitution, Equation 11.3-10 reduces to:

$$\frac{d^2 \psi_i}{dx^2} + \left(\frac{2m}{\hbar^2}\right) \left( E - V_i \right) \psi_i = 0$$ \hspace{1cm} 11.3-12
in each region. This is similar to the waveguide equation for the dielectric guide, which had the form:

\[
d^2E_i/dx^2 + (n_i^2k_0^2 - \beta^2) E_i = 0
\]

In fact, the analogy is perfect if we take the term \((2m/hbar)^2 \{E - \psi_i\}\) to correspond to \(n_i^2k_0^2\). We may therefore assume the features of the likely solutions directly from previous results. Firstly, if the term \((2m/hbar)^2 \{E - \psi_i\}\) is greater than \(\beta^2\), the solution for \(\psi_i(x)\) will be in the form of sines and cosines; otherwise, it will be exponential. We would expect the former situation to occur inside the well (where \(\psi = 0\)) when \((2m/hbar)^2 E > \beta^2\). In Region 1, therefore, we could put:

\[
\psi_1(x) = \psi_0 \cos(\kappa x - \phi)
\]

with:

\[
\kappa = \sqrt{(2m/hbar^2) E - \beta^2}
\]

However, when \(\psi_i\) is large (e.g. outside the well), we might find \((2m/hbar^2) \{E - \psi_i\} < \beta^2\), so the solution will be an exponential instead. In Regions 2 and 3, we would then expect the solutions:

\[
\psi_2(x) = \psi_0' \exp(\gamma x) \quad ; \quad \psi_3(x) = \psi_0'' \exp\{-\gamma(x - h)\}
\]

with:

\[
\gamma = \sqrt{\beta^2 - (2m/hbar^2) \{E - \psi_0\}}
\]

The solutions may be matched at \(x = 0\) and \(x = h\), by imposing the reasonable requirement that \(\psi\) and \(d\psi/dx\) must be continuous. Since the mathematics is identical to that used in Chapter 6, it should be obvious that bound solutions will exist, provided the eigenvalue equations \(\tan(\kappa h/2) = \gamma/\kappa\) or \(\tan(\kappa h/2) = -\kappa/\gamma\) are satisfied. Just as we found for the slab guide, only a discrete number of solutions can be found, each corresponding to a state in which the electron is 'trapped' inside the well. A typical solution - with the lowest energy - is shown in Figure 11.3-3. Clearly, the probability of finding the electron is highest inside the well, and diminishes rapidly outside it, although there is a finite expectation of locating it close to the well. States with a higher energy (analogous to higher-order guided modes) penetrate the region outside the well to a greater extent. In fact, if the energy is high enough, the electron is not confined at all, but is free to roam. This is analogous to the onset of cutoff in a waveguide, and the un-bound states are roughly equivalent to radiation modes. We will not examine these in detail; instead, we focus our attention on the energy levels of the bound states.

Now, in Figure 11.3-3, the exponential decay of the wavefunction outside the well is defined by \(\gamma\), which in turn depends on \(\beta\), \(E\) and \(\psi_0\). If we assume additionally that \(\psi_0\) is very large (tending to infinity), \(\gamma\) will be so great that we can take \(\psi\) to be zero outside the well. This geometry is known as the infinite potential well. Since the electron is now completely confined, the boundary conditions resemble those of the metal-walled guide, and the eigenvalue equations reduce to:

\[
\sqrt{(2m/hbar^2) E - \beta^2} h = n\pi.
\]
We then obtain for the possible electron energies:

$$\varepsilon = \nu^2 \left( \frac{\pi^2 \hbar^2}{2m} \right) + \frac{\hbar^2 \beta^2}{2m}$$ \hspace{1cm} \text{(11.3-19)}$$

These results are plotted in Figure 11.3-4; they suggest that the energies are quantized, and can take only specific, discrete values. Note firstly that the levels are not equally separated - the energy difference between the $\nu$th and $\nu+1$th levels increases rapidly with $\nu$. Furthermore, the separation of the levels depends inversely on the size of the well, and is only apparent for small values of $\hbar$; as $\hbar \rightarrow \infty$, the energies revert to those of a free electron. In the absence of motion in the z-direction (i.e., when $\beta \approx 0$), the energy is entirely potential. In this case, Equation 11.3-19 reduces to:

$$\varepsilon = \nu^2 \left( \frac{\pi^2 \hbar^2}{2m} \right)$$ \hspace{1cm} \text{(11.3-20)}$$

However, as $\beta$ increases, the total energy rises, since there are now kinetic and potential contributions.
DESIGN EXAMPLE

We can estimate the size of well for which quantization effects become significant, by comparing the value of the term $\frac{\pi^2 \hbar^2}{2m}\hbar^2$ with the thermal energy of a free electron. From the principle of equipartition of energy, the latter will be $E_{\text{th}} = \frac{3}{2} kT$, where $k$ is Boltzmann’s constant ($1.38 \times 10^{-23}$ J/K). At room temperature (when $T \approx 300$ K), $E_{\text{th}} = 6.21 \times 10^{-21}$ J. $\pi^2 \hbar^2 / 2m\hbar^2$ reaches this value when $\hbar = \sqrt{\left(\frac{(6.62 \times 10^{-34})^2}{(8 \times 9.1 \times 10^{-31} \times 6.2 \times 10^{-21})}\right)} = 3.1 \times 10^{-9}$ m. Quantization effects will thus be important for well sizes approaching interatomic distances.

TWO- AND THREE-DIMENSIONAL EFFECTS

In a one-dimensional well, we cannot easily restrict electrons to propagate in the $z$-direction alone, so that motion in the $y$-direction should really be allowed. The first question we might ask is, does this make any difference? Not really; the analysis above can be repeated, assuming the more general solution $\Psi(x, y, z) = \psi(x) \exp(-j(\beta_y y + \beta_z z))$; if this is done, we reach similar conclusions, with $\beta^2 = \beta_y^2 + \beta_z^2$. Equally, we might wish to confine the electron in more than one direction. For example, we might have an infinite well, in the form of a cube of sides $h$ in each of the $x$-, $y$-, and $z$- directions (as in our earlier Figure 11.2-2). This does make more of a difference, as we now show.

In this case, we must satisfy the equation:

$$\nabla^2 \Psi(x, y, z) + \left\{\frac{(2m/\hbar^2)}{E}\right\} \Psi(x, y, z) = 0$$  

11.3-21

inside the well, subject to the boundary conditions of $\Psi = 0$ on all faces of the cube. This situation is again analogous to the three-dimensional Fabry-Perot resonator, so we should look for standing wave solutions in each of the three directions. We might put:

$$\Psi(x, y, z) = \psi_0 \psi_x(x) \psi_y(y) \psi_z(z)$$  

11.3-22

where $\psi_0$ is a constant, and:

$$\psi_x = \sin(k_x x) ; \quad \psi_y = \sin(k_y y) ; \quad \psi_z = \sin(k_z z)$$  

11.3-23

Substituting into Equation 11.3-21, we see that this solution will be valid, provided:

$$k_x^2 + k_y^2 + k_z^2 = k^2 = \left(\frac{2m/\hbar^2}{E}\right)$$  

11.3-24

Similarly, the boundary conditions will be satisfied, if we choose:

$$k_x = v_x \pi/h, \quad k_y = v_y \pi/h \quad \text{and} \quad k_z = v_z \pi/h$$  

11.3-25

This yields the final result for the allowed energies:

$$E = (v_x^2 + v_y^2 + v_z^2) \left(\frac{\pi^2 \hbar^2}{2m\hbar^2}\right)$$  

11.3-26

This result should be compared with Equation 11.3-20; clearly, confinement in each direction has contributed a similar amount to the total energy. However, there is a new feature - the spacing of the levels is different. While the single index $v^2$ may only have the values 1, 4, 9 and so on, the term $v_x^2 + v_y^2 + v_z^2$ may assume the values 1, 2, 3, 4, 5, 6 ...,
depending on the exact choices of $\nu_x$, $\nu_y$ and $\nu_z$. Interestingly, some of the values of $\nu_x^2 + \nu_y^2 + \nu_z^2$ may be made up in more than one way. For example, a total of 3 may only be obtained if $\nu_x = 1$, $\nu_y = 1$, $\nu_z = 1$, but 6 may be obtained from $\nu_x = 1$, $\nu_y = 1$, $\nu_z = 2$, or from $\nu_x = 1$, $\nu_y = 2$, $\nu_z = 1$, or from $\nu_x = 2$, $\nu_y = 1$, $\nu_z = 1$. As a result, different states may now have the same energy; such states are described as degenerate. Furthermore, as the energy rises, the number of states with the same (or nearly the same) energy will also increase. We will return to this later, when we consider the density of states for electrons.

**ELECTRONIC TRANSITIONS**

It might be argued that the quantization of energy levels in a potential well is of purely academic interest. However, we do not need to add many refinements to describe situations of fundamental importance. For example, the function:

$$V(r) = -\frac{e^2}{4\pi \varepsilon_0 r}$$

11.3-27

describes the potential energy of an electron at a distance $r$ from a single positive charge, e.g. a proton. This is clearly a three-dimensional well, albeit of a more complicated form. If we now substitute Equation 11.3-27 into the Schrödinger equation and look for bound solutions, we will actually determine the electronic energy levels of a single hydrogen atom. As it turns out, this problem has a complete analytic solution. Most importantly, differences between the energy levels can be shown to correspond exactly to the energies of photons that are absorbed or emitted by hydrogen, demonstrating not only that the levels themselves are substantially correct, but also that the absorption and emission of radiation may be linked directly to electronic transitions. This is illustrated by Figure 11.3-5, which shows the interaction of a photon with a simple two-state system. If an incident photon has energy $hf = E_2 - E_1$, it may cause an electron to jump from the low energy state $E_1$ to the high energy state $E_2$, being itself absorbed in the process (Figure 11.3-5a). Similarly, a transition in the reverse direction may cause the emission of a photon, again of energy $hf = E_2 - E_1$ (Figure 11.3-5b).

![Figure 11.3-5](image)

This view of absorption and emission is clearly rather different from the mechanistic model previously given in Chapter 3. Most importantly, it was found necessary to postulate the existence of a third process in order to explain the black body radiation law. Here, an incident photon of the correct energy may induce a downward electronic transition, resulting in the emission of a photon (Figure 11.3-6). Since the original photon is not absorbed, the result is that two photons actually emerge.
At first sight, this is a very strange effect, since we might expect that when a photon hits an electron in the state $E_2$, it would force the electron to move to a still higher state (say, $E_3$). To appreciate why this does not occur, we must remember that the photon is essentially a wave train of electric field with some characteristic frequency, and that the electron states $E_1$ and $E_2$ are simply normal modes of the unperturbed atom. If these modes are perturbed by interacting with an oscillating field, then mode coupling will occur strongly if the frequency of the field matches the difference in frequency between the two modes. This is clearly analogous to the mode coupling situations encountered in Chapter 10, where a spatially periodic perturbation was used to phase-match two propagating modes.

Since the new photon has been emitted in response to the stimulation of the driving field, the phases of the incident and stimulated photons must be linked. In fact, they are identical, so that the two photons are effectively joined together in a single wavetrain. We may describe this as a 'doubly occupied photon mode'. This effect, known as **stimulated emission**, is the key to the amplification of light in optoelectronic devices. Spontaneous emission can also be interpreted in terms of coupling between higher and lower states, except that in this case the stimulating perturbation is not derived from an external agency (an incident photon) but from fluctuations in the local field within the atom itself.

For the other elements, a similar model may be constructed. For example, helium has two protons and two electrons, so each electron will derive potential energy from its companion, as well as from the protons in the nucleus. Unfortunately, the relevant mathematical description turns out to be of the type known as a many-body problem, which cannot be solved analytically. In fact, the higher the atomic number of the element, the worse the difficulty. However, through suitable approximations, quantum mechanics has allowed good estimates of the energy levels of all the elements to be made.

Two additional refinements are needed to account completely for all the spectroscopic observations. Firstly, some energy levels are found experimentally to be split into two closely-spaced levels. This feature may be explained by attributing to electrons an additional property known as **spin**, which may have one of two states, spin-up and spin-down. Originally, the physical description of spin was mechanistic - electrons were assumed to spin about their own axes - but no such analogy really exists; it is simplest to view spin as a vehicle for increasing the number of possible states to conform with experiment. Secondly, to account for details of the actual occupation of the levels, it is necessary to assume that there can only be one electron in any given state at any time. This idea was first postulated by W. Pauli in 1925, and is known as the **exclusion principle**.
11.4 BAND THEORY

The arguments above apply to isolated potential wells or to single atoms. However, most electronic and optoelectronic devices are based on crystalline solids. Here, the electron waves are propagating in a lattice, i.e. in a medium that contains a regular periodic array of fixed, positively-charged nuclei, each of which generates its own potential distribution. To model this geometry, we must somehow combine the results for single wells with the additional feature of periodicity.

To do so, we must in general express $V$ as a periodic function of all three co-ordinates. Solution of the resulting Schrödinger equation is extremely difficult, forcing some simplifications. Firstly, we shall assume a one-dimensional variation of potential energy, $V(x)$. Secondly, we might guess that any new features in the solution will emerge from the periodic nature of $V(x)$, and not from its detailed form.

In this, we follow Krönig and Penney, who gave a particularly simple explanation of the behaviour of electrons in crystals - known (naturally) as the Krönig-Penney model - in 1931. Since any periodic variation will do, we will use the simplest consistent with the physics. For a one-dimensional lattice, the potential energy of an electron will be a maximum exactly between the ions, falling to a minimum at each ion. We shall therefore adopt the potential distribution of Figure 11.4-1.

![Figure 11.4-1 Variation of $V(x)$ used in the Krönig-Penney model.](image)

Here, the potential energy near each ion is modelled as a well of height $V_0$ and width $h$. The edges of neighbouring wells are $g$ apart, so that the lattice periodicity is $L = g + h$. For simplicity, we shall ignore electron motion in the y- and z-directions, and concentrate on the effect of the lattice on the variation of the wavefunction in the x-direction. The equation that must be solved is therefore:

$$\frac{d^2 \psi(x)}{dx^2} + \left(\frac{2m}{\hbar^2} \right) \left(\mathcal{E} - V\right) \psi(x) = 0$$

11.4-1

with $V = 0$ inside the wells, and $V = V_0$ outside.

Now, we have tackled wave propagation problems in a layered structure many times before. In each case, a solution was first found for each layer in turn; the solutions were then matched at the interfaces, using suitable boundary conditions. The difficulty here is that we will have to solve an infinite number of simultaneous equations, since there are an infinite number of layers. Clearly, we must try to find some relation between the values of the
wavefunction at different points along the lattice, which will reduce the number of equations. This was originally done by Bloch in 1928, who based his analysis on a mathematical theorem governing the solutions of differential equations with periodically-varying coefficients (known as Floquet's theorem). The physical argument runs like this. Since the lattice is infinite, a displacement of the coordinate origin by exactly one period can make no difference to the boundary conditions. Consequently, if $\psi(x)$ is a solution, $\psi'(x) = \psi(x + L)$ must also be one. The two can only be related by a simple constant, so we may write:

$$\psi(x + L) = C \psi(x)$$  \hspace{1cm}  11.4-2

Similarly, for a displacement of more than one period (say, $N$ whole periods), we may write:

$$\psi(x + NL) = CN \psi(x)$$  \hspace{1cm}  11.4-3

This follows from repeated application of Equation 11.4-2. Now, in order for the solution to be finite everywhere, we must have $|CN| = 1$. Consequently, $C$ can only be a complex number of modulus unity, so that $C = \exp(j\phi)$. As a result, $CN$ must be of the form $CN = \exp(j\phi N)$. If we now define a new quantity $k$ as $k = \phi/L$, Equation 11.4-3 can be written as:

$$\psi(x + NL) = \exp(jkNL) \psi(x)$$  \hspace{1cm}  11.4-4

Since $NL$ is a distance, this can be seen to be a strange type of travelling electron wave if $k$ is interpreted as the wavenumber. Because $|\psi|$ is not constant, it is not a plane wave, but it has features that are clearly wave-like. For example, within each period, the wave-functions are all similar to $\psi(x)$, but each solution is shifted in phase by a progressive amount along the lattice, and the pattern repeats itself after a distance $\lambda = 2\pi/k$. Most importantly, Equation 11.4-4 is the desired connection between the value of $\psi$ at two different points along the lattice. We may now return to our original problem.

We begin by assuming that the energy $E$ is less than $V_0$, so that the electrons tend to become trapped inside the potential wells. In these regions, we would therefore expect the wavefunction to be a standing-wave pattern. Locating the $x$-origin conveniently at the right-hand edge of one of the wells, we might therefore guess a solution inside the well (i.e., for $-h < x < 0$) of:

$$\psi_1(x) = \psi_0 \exp(j\kappa x) + \psi_0' \exp(-j\kappa x)$$  \hspace{1cm}  11.4-5

where $\psi_0$ and $\psi_0'$ are constant coefficients, and:

$$\kappa = \sqrt{(2m/\hbar^2) E}$$  \hspace{1cm}  11.4-6

Using Equation 11.4-4, we can write the wavefunction in the next well along (i.e. for $g < x < L$) in terms of this solution, as $\psi_1'(x) = \exp(jkL) \psi_1(x - L)$, or:

$$\psi_1'(x) = \exp(jkL) [\psi_0 \exp(j\kappa(x - L)) + \psi_0' \exp(-j\kappa(x - L))]$$  \hspace{1cm}  11.4-7

The solution inside other wells may be found in the same way.

Outside each well, on the other hand, we can guess that the solution will be exponentially varying. However, since the wells are close together, we would expect that an electronic tunnelling analogous to the frustrated total-internal reflection occurring in similar optical geometries might take place between them. It is then prudent to allow a solution in the form
of exponentially growing and decaying waves. In the region to the right of the well of interest (i.e. for \(0 < x < g\)) we might put:

\[
\psi_2(x) = \psi_0'' \exp(\gamma x) + \psi_0''' \exp(-\gamma x)
\]

11.4-8

where \(\psi_0''\) and \(\psi_0'''\) are further constants, and:

\[
\gamma = \sqrt{(2m/h^2) \{V_0 - \Xi\}}
\]

11.4-9

We can now match the solutions at the boundaries. At \(x = 0\), for example, we must equate \(\psi_1\) with \(\psi_2\), and \(d\psi_1/dx\) with \(d\psi_2/dx\). This yields:

\[
\psi_0 + \psi_0' = \psi_0'' + \psi_0'''
\]

\[
j\kappa \{\psi_0 - \psi_0'\} = \gamma \{\psi_0'' - \psi_0'''\}
\]

11.4-10

Similarly, at \(x = g\), we must equate \(\psi_1\) with \(\psi_2\) and \(d\psi_1/dx\) with \(d\psi_2/dx\). If this is done, we get:

\[
\exp(jkL) \{\psi_0 \exp(-j\kappa h) + \psi_0' \exp(+j\kappa h)\} = \psi_0'' \exp(\gamma g) + \psi_0''' \exp(-\gamma g)
\]

\[
j \kappa \exp(jkL) \{\psi_0 \exp(-j\kappa h) - \psi_0' \exp(+j\kappa h)\} = \gamma \{\psi_0'' \exp(\gamma g) - \psi_0''' \exp(-\gamma g)\}
\]

11.4-11

Equations 11.4-10 and 11.4-11 represent four equations, which may be solved for the four unknowns \(\psi_0, \psi_0', \psi_0''\) and \(\psi_0'''\). Now, it can be shown that in any such system of simultaneous equations, a non-trivial solution only exists if the determinant of the matrix of the equation coefficients is zero. In this case, we find (after some manipulation) that this condition is satisfied when:

\[
\cos(kL) = \cos(\kappa h) \cosh(\gamma g) + \{(\gamma^2 - \kappa^2)/2\kappa\gamma\} \sin(\kappa h) \sinh(\gamma g)
\]

11.4-12

Since the electron energy \(\Xi\) depends on \(\kappa\) and \(\gamma\), Equation 11.4-12 is a form of dispersion equation, relating \(\Xi\) to \(k\). Hence, it should reveal any modifications to the dispersion characteristic imposed by the lattice. Assuming it is soluble, we might return to Equations 11.4-10 and 11.4-11 to find the values of \(\psi_0, \psi_0', \psi_0''\) and \(\psi_0'''\).

If this is done, it should be no surprise that the overall solution is as shown in Figure 11.4-1, where we have plotted the variation of \(|\psi|\) along the lattice. This repeats periodically, as expected from Equation 11.4-4. Not unnaturally, the 'strange waves' (or Bloch waves) peak in the neighbourhood of the ion cores, where electrons are most likely to be found.

As it turns out, Equation 11.4-12 is most easily interpreted graphically. However, before doing this, we shall make one simplification. We allow \(V_0\) to become very large and \(g\) to tend to zero, in such a way that the product \(V_0 g\) remains finite. In the process, the following limiting behaviour occurs: \(h \to L\) and \(\gamma g \to 0\), while \(\gamma^2 g\) remains constant and much greater than \(\kappa^2 g\). Equation 11.4-12 then tends to:

\[
\cos(kL) = \cos(\kappa L) + P \sin(\kappa L) / (\kappa L)
\]

11.4-13

Where \(P\) is a new parameter, given by:

\[
P = \gamma^2 g L/2 = mV_0 g L/h^2
\]

11.4-14
Equation 11.4-13 may now be interpreted as follows. The right-hand side can take a wide range of values, depending on the exact values of $P$ and $\kappa L$. However, the left-hand side is limited to the region between -1 and +1 (for real $k$). For a given value of $P$, we conclude that there will only be a restricted range of $\kappa L$ (and hence of $E$) for which a solution can be found. This is illustrated in Figure 11.4-2, which shows the right-hand side of Equation 11.4-13 plotted against $\kappa L$, for $P = 10$.

\[
\frac{\cos(\kappa L) + P \sin(\kappa L)}{\kappa L}
\]

Figure 11.4-2 A plot of the function $\cos(\kappa L) + P \sin(\kappa L) / (\kappa L)$ against $\kappa L$, for $P = 10$.

Only the shaded areas allow solutions for real $k$. For large $P$, these are centred on the zeros of the function $\sin(\kappa L) / (\kappa L)$, i.e. on $\kappa L \approx \nu \pi$. Breaks in the solution occur when the right-hand side of Equation 11.4-13 lies outside the limits of ±1, i.e. when $kL = \nu \pi$.

The corresponding $E$ - $k$ diagram is shown in Figure 11.4-3. As predicted, it is discontinuous, and there are ranges of energy with no solutions. The widths of these 'forbidden bands' depend on the value of $P$, but their positions do not; they lie at $k = \nu \pi / L$. The diagram combines some of the features of both free and bound electrons - it has an overall shape that is parabolic, but nonetheless only discrete bands of energy are allowed.

Figure 11.4-3 $E$-$k$ diagram for the Krönig-Penney potential, with $P = 10$. 
We can see why this must be, by considering Equation 11.4-13 in the limits of small and large $P$. For small $P$ (when the barriers are small), it approximates to \( \cos(kL) = \cos(\kappa L) \), which has the trivial solution \( \kappa = k \). The energy is then found from Equation 11.4-6 as:

\[
E = \hbar k^2 / 2m
\]

11.4-15

This is the parabolic \( E - k \) characteristic of a free electron, and is shown as a dashed line in Figure 11.4-3. Similarly, if $P$ is very large (when the barriers are much greater), Equation 11.4-13 tends to:

\[
P \sin(\kappa L) / (\kappa L) = 0
\]

11.4-16

This has the solution \( \kappa = \nu \pi / L \), with the corresponding allowed energy levels:

\[
E = \nu^2 \hbar^2 \pi^2 / (2mL^2)
\]

11.4-17

These should be compared with the solutions previously given for an isolated well (Equation 11.3-20); the two are clearly identical if the width of the well ($h$) is set equal to $L$.

**THE EFFECT OF A FINITE CRYSTAL LATTICE**

If the crystal is infinitely large, then the sections of dispersion characteristic outside the forbidden bands are continuous lines, and there are no further restrictions on the possible values of $E$ and $k$. Any real crystal will have finite dimensions, however.

We can model this through the introduction of additional boundary conditions. For our 1-D lattice, the length of an $M$-atom crystal will be $ML$. Since we would expect electrons to be confined inside the crystal, we might arbitrarily require:

\[
\psi(x) = 0 \quad \text{at } x = 0 \quad \text{and} \quad x = ML
\]

11.4-18

We have met this type of boundary condition many times before. It may be satisfied by assuming a solution as a sum of a forward- and a backward-travelling wave, which combine into a standing-wave pattern. Now, a value of $k$ satisfying Equation 11.4-13 for a given energy $E$ is also satisfied by $-k$, so the full dispersion diagram is symmetric about $k = 0$, with the left-hand half describing backward waves. Combining two waves of the same energy will give a standing wave of the form:

\[
\psi(x + NL) = \sin(kNL) \psi(x)
\]

11.4-19

Clearly, Equation 11.4-18 then requires that the allowed values of $k$ are restricted to:

\[
k = \nu \pi / ML
\]

11.4-20

Consequently, the dispersion characteristic cannot be continuous, but must consist of discrete points. In each band (which has width $\pi / L$) there will be $M$ states, each representing one standing-wave solution. Since $M$ is likely to be large, the energy levels are likely to be so close together that this granularity will be hard to detect; nonetheless, it is significant. The effect of incorporating $M$ atoms into a 1-D lattice is summarised in Figure 11.4-4, which shows that each level of an isolated atom is split into a band of $M$ states. Including spin, this figure rises to $2M$. 
Reduced dispersion characteristics

We now consider briefly one further way of representing the dispersion characteristic. As we have seen, forward- and backward-travelling waves are of the form $\psi(x + NL) = \exp(\pm jkNL) \psi(x)$. However, since these functions are periodic in $k$, they are actually indistinguishable from the very similar solutions $\psi(x + NL) = \exp(\pm jk'NL) \psi(x)$, provided $k' = k + 2\nu \pi / L$. This shows that a very compact diagram may be drawn for both forward and backward waves, by taking suitable elements from the extended diagram and shifting them to the left or right by multiples of $2\pi / L$. If this is done, the reduced dispersion characteristic shown in Figure 11.4-5 is obtained.

![Figure 11.4-5 Reduced E-k diagram for the Kröning-Penney potential.](image)

THREE-DIMENSIONAL CRYSTALS

Real crystals differ from this simple model in several respects. Firstly, they are three-dimensional; to account for this, we should allow a three-dimensional periodic potential, and consider the variation of $E$ with the three components of a vector $k$, describing propagation in three-dimensional space. Secondly, the ion cores need not be located on a regular...
cartesian grid - their positions are determined by the particular lattice structure of the crystal. Thirdly, the potential distribution near each ion will be more complicated than the one used above. Despite this, the broad conclusions remain qualitatively valid; the energy levels of a real crystal comprise a series of bands, the investigation of which has been a major task for solid state physicists. The results are normally represented as $\Phi$ - k diagrams for motion in a given direction, or as surfaces of constant energy in k-space.

As we shall see in the next section, the exclusion principle dictates that most optical and electronic effects will be determined by the outer bands, because low-energy bands are likely to be completely filled under normal conditions (so that no transitions may be made to them). The outermost and second outermost bands are known as the conduction and valence bands, respectively.

DIRECT-GAP MATERIALS

Furthermore, the important features are mostly determined by the shape of the $\Phi$ - k diagram near the bottom of the conduction band and the top of the valence band. One particular result, which cannot really be predicted at all by our current one-dimensional model, is that the value of k required for maximum energy in the valence band may or may not correspond to that for minimum energy in the conduction band. This is illustrated in the two possible $\Phi$ - k diagrams of Figure 11.4-6. In Figure 11.4-6a, the two bands lie one above the other at their point of closest approach. Crystals with this type of band-structure are known as direct-gap materials.

Naturally, we would expect the absorption and emission of radiation in crystals to involve electronic transitions. For example, in Figure 11.4-6a, absorption of a photon of energy $\hbar \omega > \Phi_g$ can provide sufficient energy to promote an electron from a position near the top of the valence band into the conduction band. The III-V semiconductor GaAs is one example of a direct-gap material, with $\Phi_g \approx 1.42$ eV. In this case, we would expect the wavelength of any absorbed or emitted light to be around 0.87 $\mu$m. Other materials of similar band structure can be made, with band-gaps chosen for the particular application. For example, the ternary alloy GaAs$_{1-x}$P$_x$ has an energy gap $\Phi_g \approx 1.42 + 1.24x$ eV in the range $0 \leq x \leq 0.45$ (after which the bandgap becomes indirect), so the corresponding optical wavelength may be tuned between 0.63 and 0.87 $\mu$m by changes in the alloy composition. Ga$_{1-x}$Al$_x$As has a similar linear energy gap variation up to $x = 0.35$. Using these materials, devices ranging from visible LEDs to near-infrared lasers may be constructed.

Figure 11.4-7 shows a contour map of the variation of $\Phi_g$ with composition for another important optoelectronic material, the quaternary alloy In$_{1-x}$Ga$_x$As$_{1-y}$P$_y$. As can be seen, the variation is quite complicated. However, most devices are made from layers of InGaAsP grown on InP substrates. These layers must have compositions that are lattice-matched to
InP for strain-free growth to be possible. As we will show in Chapter 13, the locus of all possible compounds meeting this condition is the approximate straight line linking InP to In\(_{0.53}\)Ga\(_{0.47}\)As. The overall materials system therefore allows the fabrication of devices with direct bandgaps lying between \(E_g = 1.35\) eV (InP) and \(E_g = 0.74\) eV (In\(_{0.53}\)Ga\(_{0.47}\)As). The corresponding photon wavelengths are 0.92 \(\mu\)m and 1.67 \(\mu\)m. Sources emitting at the important wavelengths of \(\lambda_0 = 1.3\) \(\mu\)m (where dispersion is minimum in silica fibres) and 1.55 \(\mu\)m (where loss is minimum) may thus be made using the InP/InGaAsP system.

![Variation of energy gap with alloy composition, for the InGaAsP materials system.](image)

**DESIGN EXAMPLE**

Optical absorption in a direct-gap material satisfies the principle of energy conservation, but Newton's laws also require conservation of momentum. While a photon carries a fixed momentum \(p_{\text{photon}} = \hbar \omega/c\) in vacuum, an electron can have a momentum \(p_{\text{electron}} = \hbar k\), whose exact value depends on the position of the electron in the \(E - k\) diagram. What do the relative sizes of these quantities imply? We can estimate the momentum of a typical photon as follows. For a photon energy of 1 eV, \(p_{\text{photon}}\) is approximately \(5 \times 10^{-28}\) N s in vacuum. In a material of refractive index \(n\), this rises by a factor of \(n\). Taking \(n \approx 4\), we have \(p \approx 2 \times 10^{-27}\) N s.

Now, the variation of the electron wavelength between the bottom and top of a band is \(\Delta \lambda = \hbar \Delta k = \hbar \pi/L\). If we assume the lattice constant \(L\) is \(\approx 2\) Å, then \(\Delta p = 1.6 \times 10^{-24}\) N s. Hence the momentum carried by a photon is \(\approx 1000\) times less than the momentum change required when transferring an electron across a band. It follows that while visible photons have enough energy to carry electrons over the energy gap, their momenta are insufficient to translate them significantly in the k-direction. Electron transitions due to optical interactions alone therefore follow essentially vertical paths on the \(E - k\) diagram.
INDIRECT-GAP MATERIALS

Other materials have different band structures. For example, in Figure 11.4-6b, there is an offset of Δk between the top of the valence band and the bottom of the conduction band. In this case, the crystal is said to be indirect-gap; the Group IV elements Si and Ge are typical indirect-gap materials. Clearly, a large momentum mismatch must be made up if optical absorption or emission is to occur. Where is this to come from? The answer is that it can be supplied through the agency of a phonon.

So far, we have concentrated on the behaviour of electrons in the lattice, which was itself assumed to be fixed. However, the lattice is simply an arrangement of atoms held together by bonds, which are only approximately rigid. Consequently, it can vibrate if supplied with energy. Since it possible to describe any such vibrations in terms of travelling sound waves, the lattice vibrations are wave-like, and we might expect that they may also be quantized.

This is indeed the case, and the elementary units of vibration are known as phonons. By analogy with photons, we can guess that a phonon carries an energy of $E_{\text{phonon}} = \hbar \omega$, where $\omega$ is the angular frequency of the vibration, and a momentum of $p_{\text{phonon}} = \hbar k = \hbar \omega v$, where $v$ is the velocity of sound in the material. A typical upper value for $\omega$ might be $10^{13}$ rad/s and for $v$, $5 \times 10^3$ m/s, so that phonon energies of (say) $10^{-3}$ eV would have momenta of $3 \times 10^{-26}$ Ns. Comparing these figures with those calculated for photons, we see that phonons typically carry less energy but more momentum. A combination of photon and phonon emission or absorption can therefore allow conservation of both energy and momentum.

Thus in Si and Ge, most interband electron transitions involving photons would also include a phonon. Since phonon absorption is relatively improbable at room temperature, such a transition does not have a high rate, so we should not expect significant light emission from indirect-gap materials. (However, in some instances, efficient emission can be obtained, by providing impurity levels within the band-gap, in which electrons can be trapped. The impurity site is capable of absorbing any momentum imbalance.

For example, in the GaAs$_{1-x}$P$_x$ system with $x > 0.45$, the band structure becomes indirect, but the addition of nitrogen can provide electron traps near the valence band edge, allowing effective emission down to the yellow part of the visible spectrum.) The band structures of a number of important optoelectronic materials are summarised in Table 11.4-1.

<table>
<thead>
<tr>
<th>Material</th>
<th>$E_g$ (eV) @ 300K</th>
<th>Band structure</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ge</td>
<td>0.66</td>
<td>Indirect</td>
<td>Semiconductor</td>
</tr>
<tr>
<td>Si</td>
<td>1.12</td>
<td>Indirect</td>
<td>Semiconductor</td>
</tr>
<tr>
<td>GaP</td>
<td>2.2</td>
<td>Indirect</td>
<td>Semiconductor</td>
</tr>
<tr>
<td>InP</td>
<td>1.35</td>
<td>Direct</td>
<td>Semiconductor</td>
</tr>
<tr>
<td>GaAs</td>
<td>1.42</td>
<td>Direct</td>
<td>Semiconductor</td>
</tr>
<tr>
<td>C (diamond)</td>
<td>5.47</td>
<td></td>
<td>Insulator</td>
</tr>
<tr>
<td>SiO$_2$</td>
<td>8</td>
<td></td>
<td>Insulator</td>
</tr>
</tbody>
</table>

Table 11.4-1 Characteristics of a number of important semiconducting and insulating materials.
11.5 EFFECTIVE MASS

Other important characteristics of electrons may be deduced from the \( E - k \) diagram. Recall that a free electron can be modelled as a wavepacket, travelling at the group velocity. Clearly, we would like to describe motion of electrons in a crystal in a similar way, but this time we have found solutions to the wave equation as a vast array of stationary states. However, standing-wave solutions are merely sets of counter-propagating waves, and these travelling waves may also be grouped together into a wave packet that can describe a partially localised electron. This time, motion of the packet amounts to a 'hopping' of the electron between the atoms.

Most importantly, there is now a restriction on the acceleration of an electron by an electric field. As the electron gains energy, the constituents of the packet must jump to higher levels. However, the exclusion principle implies that this is possible only if the levels are unoccupied. If all are filled, there can be no conduction. Herein lies the distinction between conductors and insulators, as we will see later. Even if appropriate levels are available, the process of conduction is more complicated than might be expected, as we now show.

Returning to the reduced \( E - k \) diagram of Figure 11.4-5, we note that since the band structure is periodic, a reasonable approximation to the shape of the \( v^\text{th} \) band is provided by the relation:

\[
E(k) = E_0 - \Delta E \cos(kL)
\]

11.5-1

where \( E_0 \) and \( \Delta E \) are constants. This function is shown in Figure 11.5-1a. Note that when \( k \) is small (i.e. near the bottom of the band), we can use a binomial expansion for the cosine, to get:

\[
E - E_{\text{min}} = \Delta E \frac{k^2 L^2}{2}
\]

11.5-2

where we have defined \( E_{\text{min}} \) as the minimum energy level of the band. Consequently, the \( E - k \) relation is approximately parabolic in this region.

The velocity of a wavepacket inside the crystal may again be found by differentiating the dispersion characteristic. In this case, \( v_g \) is given by:

\[
v_g = \frac{1}{\hbar} \frac{dE}{dk} = \frac{L \Delta E}{\hbar} \sin(kL)
\]

11.5-3

The resulting variation is shown in Figure 11.5-1b. Most interestingly, an increase in energy does not necessarily give an increase in \( v_g \). We can explain this curious fact as follows. By differentiating Equation 11.1-5, we can express the mass of a free electron as:

\[
m = \frac{\hbar^2}{(d^2 \omega/dk^2)}
\]

11.5-4

Thus \( m \) is simply related to the curvature of the dispersion characteristic. We can define a similar quantity \( m^* \) - known as the effective mass - for an electron in a crystal lattice. In this way, we can still treat the electron as a pseudo free particle, but allow for the interaction with the lattice through differences in the mass from that of a free electron. Differentiating Equation 11.5-1, we obtain:

\[
m^* = \frac{\hbar^2}{(L^2 \Delta E)} \sec(kL)
\]

11.5-5
The effective mass clearly varies through the band in an extraordinary way, as shown in Figure 11.5-1c. Near the bottom of the band (where $k \approx 0$), the $E - k$ characteristic is roughly parabolic, so the electron has a positive, constant mass. This can be found by approximating Equation 11.5-5 as:

$$m^* \approx \frac{\hbar^2}{L^2 \Delta E}$$

11.5-6

Note that this may be entirely different from the mass of a free electron; however, by combining Equations 11.5-2 and 11.5-6, we may still find a parabolic relation linking $E$, $k$ and $m^*$, given by:

$$E - E_{\text{min}} = \frac{\hbar^2 k^2}{2m^*} \quad (\text{for } E > E_{\text{min}})$$

11.5-7

This result shows that the $E - k$ characteristic near the bottom of the band may be written in the same form as for a free electron (remember that $E = \hbar^2 k^2 / 2m$ in this case), provided we shift the energy origin by $E_{\text{min}}$ and use the effective mass instead of the free mass.

Near to the top of the band (where $|k| \approx \pi/L$), the band curvature has changed sign, so the effective mass is now negative! This means that if a force is applied to an electron, it will
decelerate. This odd result can be understood by recognising that there can be a gain in potential energy, more than offsetting the apparent reduction in kinetic energy. Thus we conclude that electrons near the top of a band have a negative effective mass, while those near the bottom have positive mass.

Although the band structure of a three-dimensional crystal is considerably more complicated, it turns out that the conclusions above are still broadly valid. The effective mass may be found by differentiating the dispersion characteristic twice (although since this is now a function of \( k_x, k_y \) and \( k_z \) the result will depend on direction, so that \( m^* \) turns out to be a tensor quantity). However, it is generally true that the \( E - k \) relation is roughly parabolic both near the bottom of the conduction band and near the top of the valence band, and that the local curvature is of opposite sign. Consequently, the effective masses of electrons in the former case are positive, while those in the latter are negative.

Although this appears bizarre, we shall see in the next section that there is a simple interpretation of the negative-mass states in terms of a different type of charge carrier, known as a hole.

11.6 CARRIER STATISTICS

So far we have discussed solutions to the Schrödinger equation without considering the actual occupation of the states by electrons. We will now consider this point, for a three-dimensional cube of crystal. We start by deriving a suitable density of states function, as we did earlier for photons.

THE DENSITY OF STATES

To begin with, we return to the result previously obtained for a three-dimensional well, namely that for an electron confined in a cubic well of side \( h \) the allowed energy levels are given by:

\[
E = \nu^2 \frac{\pi^2 \hbar^2}{2m^*h^2}
\]

where \( \nu^2 = \nu_x^2 + \nu_y^2 + \nu_z^2 \) and \( \nu_x, \nu_y \) and \( \nu_z \) are integers. If \( h \) is large, the minimum difference between levels is small, so the energies again tend to a continuum. With the restriction that \( \nu_x, \nu_y \) and \( \nu_z \) must be positive, and including the effect of spin (which doubles the number of states), the number of states with energy less than \( E \) can be shown by our previous method to be \( \pi\nu^3/3 \). However, since:

\[
\nu^3 = C \frac{E}{\pi^{3/2}}
\]

where \( C = 2m^*/\pi^2 \hbar^2 \), it can easily be seen that:

\[
\pi\nu^3/3 = (\pi/3) \frac{E^{3/2}}{\pi^{3/2}}
\]

If we now define this quantity to be equal the integral of a new function \( s(E) \) (i.e. \( \int s(E) \, dE \), by analogy with \( s(\omega) \) for photons), then we may find \( s(E) \) by differentiation to be:

\[
s(E) = (\pi/2) \frac{E^{3/2}}{\sqrt{2}}
\]

The density of states may then be obtained by dividing Equation 11.6-4 by \( h^3 \) to get:
Now, we have obtained this result by assuming that electrons are confined inside a finite potential well. However, we really wish to find the density of states in a lattice. Most especially, we will be interested in the variation of $S(\mathcal{E})$ near the bottom of the conduction band and the top of the valence band. How should we modify Equation 11.6-5 to cope with this? Arguably, the result we have obtained has followed from combining the quantization effect of the well with the parabolic relation between $\mathcal{E}$ and $k$ for a free electron.

Since we showed in Equation 11.5-7 that the $\mathcal{E}$ - $k$ diagram is also parabolic near the bottom of the conduction band, we might expect a similar result to be obtainable merely by replacing $\mathcal{E}$ by $\mathcal{E} - \mathcal{E}_c$ (where $\mathcal{E}_c$ is the minimum energy of the conduction band) and $m$ by $m_e^*$ (where $m_e^*$ is the value of $m^*$ near $\mathcal{E}_c$). In this case, we get:

$$S(\mathcal{E} - \mathcal{E}_c) = (\pi/2) C''^{3/2} \sqrt{\mathcal{E} - \mathcal{E}_c} \text{ for } \mathcal{E} > \mathcal{E}_c , \text{ where } C'' = 2m_e^*/\pi^2 \hbar^2$$

11.6-6

The $\mathcal{E}$ - $k$ relation is also parabolic near the top of the valence band, although the parabola is now upside-down. Applying a similar argument, we can show that the density of states here is:

$$S(\mathcal{E}_v - \mathcal{E}) = (\pi/2) C'''^{3/2} \sqrt{\mathcal{E}_v - \mathcal{E}} \text{ for } \mathcal{E} < \mathcal{E}_v , \text{ where } C''' = 2m_h^*/\pi^2 \hbar^2$$

11.6-7

where $m_h^*$ is the modulus of $m^*$ near $\mathcal{E}_v$. Combining Equations 11.6-6 and 11.6-7, the variation of $S(\mathcal{E})$ throughout the conduction and valence bands is much as shown in Figure 11.6-1a.

![Figure 11.6-1 Typical variation of a) the density of states, and b) the Fermi-Dirac distribution function.](image-url)
THE FERMI-DIRAC DISTRIBUTION FUNCTION

We now turn our attention to the actual occupation of the levels. Once again, the details are determined by statistical thermodynamics. Given that the total energy of the electrons and their number are fixed, and that they satisfy the Pauli exclusion principle, it can be shown that the probability of occupation of a state of energy \( E \) is given by a function known as the Fermi-Dirac probability function (since it was originally derived independently by E. Fermi and P.A.M. Dirac in the 1920s). This represents the most likely distribution of electron energies, and has the form:

\[
F(E) = \frac{1}{1 + \exp\left\{ \frac{(E - E_F)}{kT} \right\}}
\]

where \( E_F \) (the Fermi energy) is the energy at which the probability of a state being filled is exactly 0.5. The Fermi-Dirac probability function is rather different from the Bose-Einstein function, and is shown in Figure 11.6-1b for two temperatures \( T_1 \) and \( T_2 \). As can be seen, increased temperature causes a greater spread of the function, so that the probability of occupation of higher-energy states rises. A reduction in temperature causes the opposite effect, so that lower states are filled preferentially. In the limit as \( T \to 0 \), we find that \( F(E) = 1 \) for \( E < E_F \), and \( F(E) = 0 \) for \( E > E_F \).

By and large, the electronic properties of most materials are determined by the absolute position of the Fermi level, and by the size of the energy gap \( E_g \) by comparison to \( kT \). We shall consider first what happens when \( E_F \) lies in the energy gap, when \( E_g \gg kT \) as shown in Figure 11.6-1. Since the Fermi-Dirac function tails off rapidly into the conduction band, we should find only a few electrons in this range. In pure or intrinsic material, these must have originated in the valence band, but thermal energy provides a means for them to move up into the conduction band. Together with these filled states, there is a vast array of empty ones. Electrons can easily move into these states, so they can gain energy in an electric field. Some electrical conductivity is therefore to be expected.

Similarly, since the Fermi-Dirac function is close to unity in the valence band, we would expect most states in this band to be filled. In intrinsic material, however, some empty states must be generated as electrons are promoted to the conduction band. The presence of even a few such states can allow nearby electrons to move into them, and hence contribute to conduction. In this case, the process of conduction is best viewed rather differently. Since the availability of an empty state allows a neighbouring particle - a negatively charged, negative-mass electron - to move into it, the empty state and the electron can effectively change places. Because only a small number of spaces are available in a great sea of electrons, it is simpler to treat the spaces themselves as the particles. These positively-charged pseudo-particles are known as holes; since they move in the opposite direction to electrons, they may be described by a positive mass. This leads to great simplifications.

To summarise, we would expect that a small number of electrons in the conduction band and holes in the valence band will be able to conduct electricity. Furthermore, an increase in temperature will extend the tail of the Fermi-Dirac distribution, and hence the number of electrons and holes available for conduction, and a shift of \( E_F \) from the middle of the band-gap will markedly increase the number of carriers in one band at the expense of the other. Materials with this kind of energy band diagram are classed as either semiconductors or insulators. The distinction is essentially one of degree. If \( E_g \) is so large that the numbers of carriers in the two bands are very small, then the material is an insulator. If the energy gap is moderate, it is a semiconductor. Table 11.4-1 illustrates this distinction.
Finally, we note that in most metals, the conduction band is partly filled even at low temperatures, so the conductivity is always high. Despite this advantage, the limited conductivity of a semiconductor is more than compensated for by the degree of control that can be exercised through the positioning of the Fermi level. This is achieved by doping the material with impurity atoms. Since the properties of the semiconductor are then largely determined by the dopants, it is called extrinsic material.

For example, crystalline silicon is formed from Group IV atoms, each of which has four outer-shell or valence electrons and combines in a tetrahedral covalent bond with its nearest neighbours. Group V atoms (e.g. arsenic or phosphorous) are of a similar size, but have five valence electrons. If such an atom is substituted for silicon in the lattice, four of its five electrons will be used up in nearest-neighbour bonding. Very little energy (typically, < 0.05 eV) is required to detach the remaining one from its parent atom, since it is only weakly bound. If this can be supplied, the fifth electron will be available for conduction. We can represent the process by defining additional localised impurity states in the energy diagram, located slightly below the bottom of the conduction band. Since the dopants provide electrons in this case, they are known as donors, and the level of the impurity states is defined as $E_d$ (as in Figure 11.6-2a). Less energy is required to promote electrons into the conduction band from these states than from the valence band, and, in the process, the number of conduction electrons is increased, without altering the number of holes. Doping with donors has the additional effect of lifting the Fermi level up towards $E_c$, as we shall see later.

Group III atoms (such as gallium or boron) have only three valence electrons. Consequently, if these are substituted for silicon, they lack the correct number of electrons for complete bonding. Given a small amount of energy, an electron may be attracted from a neighbouring silicon atom, leaving a hole in the silicon bond network. We may represent this process by defining other impurity states at the acceptor level $E_a$, just above $E_v$ as shown in Figure 11.6-2b. Because holes may be created from these states very easily, the effect is to alter the hole density without changing the number of electrons. A further consequence is that the Fermi level is pulled down towards $E_v$.

Dopants may also be added to III-V semiconductors, such as GaAs and InP. Clearly, a Group VI atom will provide an additional electron if it replaces a Group V atom in the lattice, so it can act as a donor. Similarly, Group II dopants can replace Group III atoms and act as acceptors. However, Group IV dopants may be donors or acceptors, depending on which atom they replace.

Thus, as we move $E_F$ around in the energy gap (or even out of it) by suitable doping, we can alter the conductivity of semiconductors, and produce material with either predominantly
'normal' electron conduction (known as \textbf{n-type} material), or mainly 'abnormal' electron or hole conduction (\textbf{p-type} material). We have already seen one application for this technique in Chapters 3 and 9; a change in the number of free carriers can be used to generate a change in refractive index, via the plasma contribution to the dielectric constant. However, many other highly important properties follow from further application of the same basic principle.

\section*{11.7 INTRINSIC AND EXTRINSIC SEMICONDUCTORS}

The equilibrium densities of electrons and holes can be calculated very simply for intrinsic material. To find the total density \( n \) of electrons in the conduction band, we merely multiply the density of available states (i.e. \( S(\mathcal{E}) \)) by their probability of being filled (\( F(\mathcal{E}) \)), and integrate the result from the bottom of the conduction band to the top. Thus we may write:

\[ n = \int_{\mathcal{E}_c}^{\mathcal{E}_t} F(\mathcal{E}) \, S(\mathcal{E}) \, d\mathcal{E} \]

\textbf{11.7-1}

where \( \mathcal{E}_t \) is the level at the top of the conduction band. Returning to Figure 11.6-1, however, we note that it is often true that \( \mathcal{E} - \mathcal{E}_F \gg kT \) within the conduction band, so that \( F(\mathcal{E}) \) declines roughly exponentially with \( \mathcal{E} \) in this range. We may therefore write:

\[ F(\mathcal{E}) \approx \exp\{-(\mathcal{E} - \mathcal{E}_F)/kT\} \]

\textbf{11.7-2}

Since this function tends to zero very quickly as \( \mathcal{E} \) increases, the upper limit \( \mathcal{E}_t \) of the integral above may be allowed to tend to infinity without appreciably changing the answer. Using the density of states function given in Equation 11.6-6, the integral may then be approximated by:

\[ n \approx \int_{\mathcal{E}_c}^{\infty} \exp\{-(\mathcal{E} - \mathcal{E}_F)/kT\} \left(\frac{\pi}{2}\right) C''^{3/2} \sqrt{(\mathcal{E} - \mathcal{E}_c)} \, d\mathcal{E} \]

\textbf{11.7-3}

This is a standard integral, which may be evaluated analytically to give:

\[ n \approx N_c \exp\{-(\mathcal{E}_c - \mathcal{E}_F)/kT\} \]

\textbf{11.7-4}

where \( N_c \) is a term varying only slowly with temperature, given by:

\[ N_c = 2(2\pi m_e^*kT/h^2)^{3/2} \]

\textbf{11.7-5}

The net result is that \( n \) depends (for the most part) exponentially on temperature, and on the position of the Fermi level relative to the base of the conduction band. The first feature is as we would expect, since an increase in temperature means that more carriers will have sufficient energy to move from the valence band into the conduction band. We will return to the second shortly.

Similarly, to find the total density \( p \) of holes in the valence band, we merely multiply the density of available states by their probability of not being filled (\( 1 - F(\mathcal{E}) \)), and integrate the result from the bottom of the valence band to the top. Thus we may write:

\[ p = \int_{\mathcal{E}_v}^{\mathcal{E}_p} \{1 - F(\mathcal{E})\} \, S(\mathcal{E}) \, d\mathcal{E} \]

\textbf{11.7-6}

where \( \mathcal{E}_v \) is the energy at the bottom of the valence band. A suitable approximation to the Fermi-Dirac function may again be used. Returning to Figure 11.6-1, we note that \( \mathcal{E} - \mathcal{E}_p \ll \)}
kT throughout the valence band, so that \( F(\varepsilon) \) is close to unity, and \( 1 - F(\varepsilon) \) decreases roughly exponentially with \( -\varepsilon \) in this range. We may therefore write:

\[
1 - F(\varepsilon) \approx \exp\left\{-(\varepsilon - \varepsilon_c)/kT\right\}
\]

11.7-7

Since this tends to zero very quickly as \( \varepsilon \) decreases, the lower limit \( \varepsilon_b \) of the integral may be allowed to tend to minus infinity, without much effect. Using the density of states function given in Equation 11.6-7, and evaluating the integral by standard methods, we get:

\[
p \approx N_v \exp\left\{-(\varepsilon_c - \varepsilon_v)/kT\right\}
\]

11.7-8

where the term \( N_v \) is also slowly-varying with temperature, and given by:

\[
N_v = 2(2\pi m_h^*/\hbar^2)^{3/2}
\]

11.7-9

The result is similar to that previously obtained for \( n \), but this time we find that the density of holes is exponentially dependent on the position of the Fermi level relative to the top of the valence band.

CARRIER GENERATION AND RECOMBINATION IN INTRINSIC MATERIAL

The generation of carrier pairs should yield equal electron and hole densities, so that \( n = p \). We may use this to deduce the position of the Fermi level. Substituting for \( n \) and \( p \) from Equations 11.7-4 and 11.7-8, we find after a little rearrangement that \( \varepsilon_F \) is given by:

\[
\varepsilon_F = \frac{\varepsilon_c + \varepsilon_v}{2} + \frac{3kT}{4} \log_e \left(\frac{m_{h}^*/m_{e}^*}\right)
\]

11.7-10

Since \( m_{e}^* \) and \( m_{h}^* \) are of the same order, the Fermi level lies about half way between the conduction and valence bands, as anticipated. We may also calculate the quantity \( n_i = n = p \) (the intrinsic carrier concentration). We first form the product \( n \times p \) using Equations 11.7-4 and 11.7-8, as:

\[
n \times p = N_c N_v \exp\left\{-(\varepsilon_c - \varepsilon_v)/kT\right\} = N_c N_v \exp\left\{-\varepsilon_g/kT\right\}
\]

11.7-11

As can be seen, this result is independent of the position of the Fermi level, a point we will return to shortly. \( n_i \) may then be found as the square-root of the above, namely:

\[
n_i = \sqrt{N_c N_v} \exp\left\{-\varepsilon_g/2kT\right\}
\]

11.7-12

The intrinsic carrier concentration therefore decreases exponentially with \( \varepsilon_g \). This explains the poor conductivity of insulators, which have large energy gaps. Why, however, should \( n_i \) depend only on temperature? To answer this, we must examine the dynamics of carrier-pair generation. Since this is a thermal effect, we would expect its rate to depend only on \( T \). We might define this as \( G(T) \).

However, in equilibrium, it must be exactly balanced by an inverse process called recombination, in which a conduction electron falls back into the valence band, annihilating an electron and a hole. Now, the probability of an electron finding a hole available for recombination must be proportional to the hole density, i.e. to \( p \). Since the electron density is itself \( n \), the recombination rate \( R \) must be proportional to the product of \( n \) and \( p \). In equilibrium, we must have \( R = G(T) \), which implies that \( R \) can only be a function of temperature. Consequently, the product \( n \times p = n_i^2 \) is constant at a given value of \( T \).
EXTRINSIC MATERIAL

We now consider the properties of extrinsic material, starting with the electron density in n-type material. Let us assume there is a density $N_d$ of donor atoms in an otherwise pure semiconductor. At a given temperature, we might expect a fraction $N_d^+$ of these to be ionised, and so have made their electrons available for conduction. As a result, the electron density will increase. The details of what actually occurs depend on the value of $T$. At low temperatures, $N_d^+$ is small, so that $n \approx n_i$. At higher temperatures, there is a range where the donors are fully ionised, so $N_d^+ \approx N_d$. Normally, the dopant density is high enough that $N_d \gg n_i$ at room temperature. In this case, the donated electrons dominate the electron density, and we may write $n \approx N_d$. Electrons are then the **majority carriers**; their density is essentially equal to that of the dopant atoms, and independent of temperature. At higher temperatures still, the rate of thermal carrier generation increases sufficiently that $n$ tends to $n_i$ again.

In the normal regime, where $n \approx N_d$, we may use Equation 11.7-4 to find an approximate position for the Fermi level, as:

$$E_{F_n} \approx E_c - kT \log_e \left( \frac{N_c}{N_d} \right)$$

11.7-13

Consequently, in n-type material, the Fermi level moves close to the base of the conduction band, as we have previously stated. (Note that if $E_{F_n}$ is very near to the band edge, or even in the band itself, Equation 11.7-4 is no longer accurate. This is the case in heavily doped material.)

We must also consider the effect of doping on the hole density. Now, to first-order, the generation rate $G$ will be unchanged by the introduction of dopant atoms, because carrier-pair generation is a thermal effect. Consequently, the recombination rate $R$ must also be unaffected, so that the expression $np = n_i^2$ must still be valid. In the normal regime, where $n \approx N_d$, we would therefore expect the hole density to decrease, so that the product $np$ is held constant. The holes are therefore **minority carriers**, since their density is reduced by doping to a value $p = n_i^2/n \approx n_i^2/N_d$, which is much lower than $n_i$. In n-type material, therefore, conduction occurs predominantly by electrons.

Results for p-type material (where doping is through acceptor atoms) may be found by analogy. In the normal regime, the acceptors are fully ionised, and the density of holes induced in the valence band by the doping greatly outweighs the effects of thermal carrier-pair generation. In this case, holes are the majority carriers, and we find that $p \approx N_a$, where $N_a$ is the acceptor density. Using Equation 11.7-8, we can show that the position of the Fermi level becomes:

$$E_{F_p} \approx E_v + kT \log_e \left( \frac{N_v}{N_a} \right)$$

11.7-14

The Fermi level then shifts to a position near the top of the valence band, also as described earlier.

MOTION OF CHARGE CARRIERS

Having established the carrier densities, we must examine how they may move. Clearly, an electric field will accelerate an electron. However, the acceleration will not continue indefinitely; collisions at random intervals will cause the electron to lose its kinetic energy,
reducing its velocity to zero. The instantaneous speed of the electron will therefore be random, but its average value - known as the drift velocity, \( v_d \) - in the direction of the field will be non-zero. The motion of \( n \) electrons per unit volume results in a current density \( J_e \) of:

\[
J_e = nev_d
\]

Clearly, \( v_d \) should increase with the applied field, so we may write:

\[
v_d = \mu_e E
\]

where \( \mu_e \) is a new parameter, the electron mobility. The current density \( J_e \) is therefore also proportional to the field. A similar motion of holes may occur; because holes move in the opposite direction, but carry the opposite charge, the analogous expression in this case is \( J_h = p_e \mu_h E \), where \( \mu_h \) is the hole mobility. The total current density is the sum of these two components, namely:

\[
J = \{ ne\mu_e + p_e \mu_h \} E
\]

Now, Maxwell's equations imply that a one-dimensional field \( E \) will yield a current density \( J = \sigma E \). From the above, we can see that this is substantially true, provided we define the conductivity \( \sigma \) as:

\[
\sigma = ne\mu_e + p_e \mu_h
\]

This is not the end of the story, however, since there is a further mechanism for motion of free carriers, that is not included in the version of Maxwell's equations given in Chapter 2: diffusion.

In any system of mobile particles (e.g. a gas, or a collection of free carriers), random thermal motion occurs independently of any other effect. If concentration gradients exist, this will tend to remove them, so that the concentration everywhere tends to equalise. The necessary mathematics is common to many branches of physics, and may be derived by considering a simple field-free system involving one-dimensional diffusion. Figure 11.7-1 shows two adjacent layers of a mixture of mobile carriers (say, electrons); each has unit cross-sectional area and thickness \( dx \), but the left-hand layer contains a density \( n \) of electrons, while the right-hand one has a different density \( n' = n + \frac{\partial n}{\partial x} \) dx.

Due to the random motion, we would expect a fraction of the electrons in the left-hand region to move to the right and cross the boundary between the layers. Now, the number moving into the right-hand region per unit area and time will be proportional to \( n \), so we
may write it as \( D_e n \), where \( D_e \) is a constant known as the electron **diffusion coefficient**. Simultaneously, electrons from the right-hand region will move in the opposite direction. The probability of moving in either direction must be the same, so the number of electrons crossing the boundary from right to left is \( D_e \{ n + \partial n/\partial x \, dx \} \). The two processes yield a net motion to the right of \( D_e n - D_e \{ n + \partial n/\partial x \, dx \} \) electrons. These move a distance \( dx \) in unit time, so this is equivalent to a flux density \( F \) of electrons, where:

\[
F = -\Delta \partial n/\partial x
\]

Because electrons carry a negative charge, this is equivalent to a **diffusion current density** of:

\[
J = D_e e \partial n/\partial x
\]  

Equation 11.7-20 shows that a non-uniform electron density will result in a flow of current in the direction of the concentration gradient. A similar equation may be written for holes; since these carry a positive charge, this time we get \( J = -D_h e \partial p/\partial x \), where \( D_h \) is the diffusion coefficient for holes. Thus, including both drift and diffusion components, the electron and hole current densities are:

\[
J_e = n e \mu_e E + D_e e \partial n/\partial x \quad \text{and} \quad J_h = p e \mu_h E - D_h e \partial p/\partial x
\]

In the steady state, the partial derivatives are replaced by total ones. The total current density is then the sum of \( J_e \) and \( J_h \). However, before we can use Equation 11.7-21, we must determine the actual carrier distributions. In general, the distribution of (say) electrons will depend on \( x \) and \( t \). We can find the variation \( n(x, t) \) by setting up a continuity equation, starting by neglecting the effect of recombination.

Figure 11.7-2 shows another view of our one-dimensional system, where a layer of thickness \( dx \) has an electron flux density \( F \) flowing in from the left, and \( F' = F + \partial F/\partial x \, dx \) flowing out to the right.

![Figure 11.7-2 Geometry for derivation of the continuity equation.](image)

The effect of the difference between \( F \) and \( F' \) is a net outflow of \( F - \{ F + \partial F/\partial x \, dx \} = - \partial F/\partial x \, dx \) electrons per unit area per unit time. Since these are lost from a layer of thickness \( dx \), the local electron concentration must drop, at a rate \( \partial n/\partial t \), where:

\[
\partial n/\partial t = - \partial F/\partial x
\]

Equation 11.7-22 -the **continuity equation** - shows that a flux gradient will cause a decrease in the local electron density, which tends to equalise the value of \( n \) everywhere.
Assuming that $D_e$ is independent of $n$, we may combine Equations 11.7-19 and 11.7-22 to eliminate $F$. We then get:

$$\frac{\partial n}{\partial t} = D_e \frac{\partial^2 n}{\partial x^2}$$

Equation 11.7-23 - known as the **diffusion equation** - may be solved to find the electron density, subject to given boundary conditions. We shall see how this is done in the next Chapter.

We now consider the additional effect of recombination. Let us suppose that the electron concentration is uniform, and higher than the equilibrium value $n_E$ for some reason. Due to recombination, $n$ will tend to decrease. As we will show later, the rate of decrease will be proportional to the excess electron density, so we may write:

$$\frac{\partial n}{\partial t} \bigg|_{\text{recomb}} = -(n - n_E)/\tau_e$$

11.7-24

where $\tau_e$ is a constant known as the **electron lifetime**. This has the general solution:

$$n(t) - n_E = \{n(0) - n_E\} \exp(-t/\tau_e)$$

11.7-25

where $n(0)$ is the initial value of $n$. As a result, the excess electron concentration decays exponentially with time. A synthesis of the effects of recombination and diffusion is simple; we merely modify Equation 11.7-23 to include the right-hand side of Equation 11.7-24.

### 11.8 DETAILED BALANCE

Having previously described the optical and electronic characteristics of semiconductors in isolation, we shall now consider combined optoelectronic interactions. We shall start by returning to the black-body radiation law, to examine its origins in the elementary processes occurring within the material. The processes we must consider are absorption, spontaneous emission and stimulated emission; these were previously shown in Figures 11.3-5 and 11.3-6, but we shall now consider their operation in detail. For simplicity, we shall consider transitions between single states in two levels of a direct-gap semiconductor. The lower state might lie in the valence band and have energy $E_1$, while the upper one (with energy $E_2$) might lie in the conduction band, as in Figure 11.8-1. For convenience, the Fermi-Dirac distribution function is also shown in this figure.

First, we consider spontaneous emission. Here, an electron from the upper state drops down to the lower state, emitting a photon of angular frequency $\omega$ (where $\hbar \omega = E_2 - E_1$) en route. Now, the likelihood of this transition occurring must be proportional to the product of the probabilities of finding an electron in the upper state in the first place, i.e. $F(E_2)$, and of finding the lower state unoccupied (namely $\{1 - F(E_1)\}$). If it does actually take place, the optical energy density $\rho(\omega)$ will increase. We may therefore write the rate of increase of $\rho(\omega)$ due to this transition alone as:

$$\frac{d\rho(\omega)}{dt} \bigg|_{\text{spon}} = A \cdot F(E_2) \cdot \{1 - F(E_1)\}$$

11.8-1

where $A$ is a constant.

Next, we consider absorption. Here, an electron from the lower state is promoted into the upper one, through the absorption of a photon. The likelihood of this occurring must depend on the probabilities of finding an electron in the lower state, i.e. $F(E_1)$, and of finding the upper state unoccupied, namely $\{1 - F(E_2)\}$, and also on the energy density.
If the transition occurs, \( \rho(\omega) \) will decrease this time, so the rate of change of energy density due to absorption can be written as:

\[
dp(\omega)/dt \bigg|_{\text{abs}} = -B_{12} \rho(\omega) F(E_1) \{1 - F(E_2)\}
\]

where \( B_{12} \) is a constant.

Finally, we consider stimulated emission, where the effect of an incident photon is to cause an electron to drop from the upper state to the lower one, releasing an additional photon. In this case, we have:

\[
dp(\omega)/dt \bigg|_{\text{stim}} = B_{21} \rho(\omega) F(E_2) \{1 - F(E_1)\}
\]

where \( B_{21} \) is a further constant.

The total rate of change of energy density \( dp(\omega)/dt \) must be given by the sum of all these values. In equilibrium, this must be zero, so that:

\[
A F(E_2) \{1 - F(E_1)\} - B_{12} \rho(\omega) F(E_1) \{1 - F(E_2)\} + B_{21} \rho(\omega) F(E_2) \{1 - F(E_1)\} = 0
\]

Substituting for the Fermi-Dirac distribution function from Equation 11.6-8, and performing a small amount of rearrangement, we may then extract the value of \( \rho(\omega) \) as:

\[
\rho(\omega) = A / [B_{12} \exp((\hbar\omega)/kT) - B_{21}]
\]

At this point, we pause to note our achievement: we have recovered the equilibrium energy density at angular frequency \( \omega \) by considering the detailed balance of three processes which tend either to increase or to decrease the density of photons of this frequency. In fact, we could perform a similar analysis for transitions between any two states, and we would obtain a similar result. Why should this be? The answer must lie in the universal black body radiation law. Returning to Equation 11.2-13, we have already shown that the energy density \( \rho(\omega) \) must be:

Figure 11.8-1 Two particular states in the valence and conduction bands of a semiconductor.
Comparison between Equations 11.8-5 and 11.8-6 shows that this will be the case, if \( B_{12} = B_{21} = B \) and \( A/B = \frac{\hbar n^3 \omega^3}{\pi^2 c^3} \). We have therefore deduced two important truths: the coefficients associated with absorption and stimulated emission are identical, while those for spontaneous and stimulated emission have a definite ratio. Historically, the A and B coefficients are known as the Einstein coefficients (after Albert Einstein, who first suggested the possibility of stimulated emission).

11.9 RATE EQUATIONS

In optoelectronics, we will normally be concerned with operation away from thermal equilibrium. For example, we might increase the photon density above its equilibrium value by shining a beam of light on a semiconductor, or we might increase the carrier density by injecting extra electrons. To analyse the results, we must make some simplifications. We should really describe the transfer of electrons between bands in terms of the probabilities of occupation of a particular pair of states, integrating over all possible states. Since this is a complicated procedure, we shall use a much simpler model, in which we regard the system as having only two states. Among other things, this implies that all photons involved have roughly the same energy. This model is so far from the truth that we might expect any of its predictions to be valueless. In fact, if the conditions are carefully chosen, the predictions are consistent with observation and are simple enough to be useful.

As before, the rate of change of the concentration \( \phi \) of photons of energy \( \hbar \omega = \hbar \omega \approx E_g \) has three contributions. We begin by considering the case where there is no optical feedback, and the electron and hole densities are low. In these circumstances, we can neglect stimulated emission entirely. The rate of spontaneous emission will obviously depend on the electron density in the conduction band. However, it will also depend on the density of holes in the valence band, because this band is so full that the availability of unoccupied states for electrons to move into is a rate-limiting factor. Absorption depends on the photon density only, since there are a very large number of electrons in the valence band and many spaces in the conduction band. The rate of change of photon density is therefore:

\[
\frac{d\phi}{dt} = anp - b\phi
\]

where \( a \) and \( b \) are constants related to (but not identical to) the Einstein A and B coefficients used in the previous Section. Now, the generation of each photon consumes one electron and one hole, so we may write the rate of change of electron density as:

\[
\frac{dn}{dt} = -anp + b\phi
\]

Naturally, we could write an identical equation for holes, but we shall not bother: the essence of the problem is already fully described. The problem with these equations is that they are nonlinear. However, they can be linearised through suitable approximations, as we now show.

In thermal equilibrium, the electron, hole and photon densities are stabilized at their equilibrium values \( n_e, p_h \) and \( \phi, \) and all derivatives are zero. We may therefore write:

\[
an_e p_h - b\phi = 0
\]
We now assume that all quantities are perturbed from their equilibrium values, so that \( n = n_E + \Delta n, \ p = p_E + \Delta p \) and \( \phi = \phi_E + \Delta \phi \). In this case, Equation 11.9-2 becomes:

\[
\frac{d(n_E + \Delta n)}{dt} \approx -a(n_E + \Delta n)(p_E + \Delta p) + b(\phi_E + \Delta \phi)
\]

We may now remove a number of terms using Equation 11.9-3. At this point, we also note that the requirement for charge neutrality implies that \( \Delta n \) must equal \( \Delta p \). Assuming additionally that \( \Delta n \) and \( \Delta p \) are small, Equation 11.9-4 may be reduced to:

\[
\frac{d(\Delta n)}{dt} \approx -a\Delta n(n_E + p_E) + b\Delta \phi
\]

11.9-5

This implies that an excess photon density will increase the electron density through absorption. However, this process is counteracted by spontaneous emission, which will tend to remove any such increase by recombination. The change due to spontaneous emission alone can be written:

\[
\left. \frac{d(\Delta n)}{dt} \right|_{\text{spon}} = -\frac{\Delta n}{\tau_{rr}}
\]

11.9-6

where \( \tau_{rr} \) is a constant known as the **radiative recombination lifetime**, defined as:

\[
\tau_{rr} = \frac{1}{a(n_E + p_E)}
\]

11.9-7

**DESIGN EXAMPLE**

We may estimate the radiative recombination lifetime in p-type silicon with a doping level of \( N_a \approx 10^{24} \text{ m}^{-3} \) as follows. For Si, the relevant recombination coefficient is \( a \approx 2 \times 10^{-21} \text{ m}^3/\text{s} \). The recombination time is \( \tau_{rr} = 1/a(n_E + p_E) \); this may be approximated by \( 1/a p_E \approx 1/a N_a \) since \( p_E \gg n_E \). Thus, \( \tau_{rr} \approx 1/(2 \times 10^{-21} \times 10^{24}) = 0.5 \times 10^3 \text{ s} \). This is so long, that the corresponding rate \( r_{rr} = 1/\tau_{rr} \) is extremely small. By comparison, the recombination coefficient in GaAs is \( a \approx 7 \times 10^{-16} \text{ m}^3/\text{s} \), a factor of \( \approx 10^5 \) larger.

**NON-RADIATIVE RECOMBINATION**

Now, it is possible for recombination to involve the generation of phonons rather than photons. This is termed **non-radiative recombination**, and has a characteristic lifetime \( \tau_{nr} \). Since the rates of the two processes are additive, we might therefore define an overall recombination rate as:

\[
\frac{d(\Delta n)}{dt} \bigg|_{\text{recomb}} = -\frac{\Delta n}{\tau_e}
\]

11.9-8

where \( 1/\tau_e \) - the overall electron lifetime used in the previous Section - is given by:

\[
1/\tau_e = 1/\tau_{rr} + 1/\tau_{nr}
\]

11.9-9

With this new notation, the electron **rate equation** may be written in the form:

\[
\frac{d(\Delta n)}{dt} \approx -\frac{\Delta n}{\tau_e} + b\Delta \phi
\]

11.9-10

Allowing the generation of photons to occur only through radiative recombination, a similar argument would lead to the corresponding equation for photons:
This shows that any excess of photons will tend to decrease through absorption, and increase through spontaneous emission, as might be expected.

CONCENTRATION GRADIENTS

So far, we have assumed uniform electron and photon densities, and we must now make allowance for concentration gradients. This is very simple - in fact, we have already covered the modifications needed for electrons in Section 11.7. We must include another term, that accounts for the extra rate of change of $\Delta n$ due to gradients in electron flux. For a one-dimensional flux in the x-direction, this term is $\partial \Delta n / \partial t = -\partial F / \partial x$. Since $F = -D_e \partial \Delta n / \partial x$, $\partial F / \partial x = -D_e \partial^2 \Delta n / \partial x^2$ and Equation 11.9-10 becomes:

$$\partial \Delta n / \partial t = D_e \partial^2 \Delta n / \partial x^2 - \Delta n / \tau_e + b \Delta \phi$$

Equation 11.9-12

This places the electron rate equation in a form similar to Equations 11.7-23 and 11.7-24. However, since the electron diffusion current is $J = D_e \partial \Delta n / \partial x$, we may also write Equation 11.9-12 as:

$$\partial \Delta n / \partial t = (1/e) \partial J / \partial x - \Delta n / \tau_e + b \Delta \phi$$

Equation 11.9-13

This version will prove more appropriate for the analysis of light-emitting devices in Chapter 12.

We now consider a similar modification to the photon rate equation. For photon motion in a different direction (say, the z-direction) the excess photon flux is $\Delta \Phi = v_g \Delta \phi$, where $v_g$ is the group velocity for photons. The extra term is now $-\partial (\Delta \Phi) / \partial z = -v_g \partial (\Delta \phi) / \partial z$, and Equation 11.9-11 becomes:

$$\partial (\Delta \phi) / \partial t \approx -v_g \partial (\Delta \phi) / \partial z + \Delta n / \tau_{rr} - b \Delta \phi$$

Equation 11.9-14

Equations 11.9-12 and 11.9-14 must really be solved simultaneously, to provide a full description of the interaction between electrons and photons. However, considerable insight may be gained from approximate solutions for several limiting cases.

(I) OPTICAL ABSORPTION

We start by making an approximation to the photon rate equation, that is appropriate for describing the absorption of a travelling wave of photons. If radiative recombination can be neglected, and a steady-state exists, so that $\partial (\Delta \phi) / \partial t = 0$, Equation 11.9-14 reduces to:

$$d(\Delta \phi)/dz = -(b/v_g) \Delta \phi$$

Equation 11.9-15

The rate of change of $\Delta \phi$ with distance therefore depends linearly on $\Delta \phi$. Now, as each photon carries energy $\hbar \omega \approx \hbar \nu_g$, and travels at velocity $v_g$, the optical power density is $P \approx \hbar \nu_g \Delta \phi$. Since $P$ is proportional to $\Delta \phi$, it should decay with distance as:

$$dP/dz \approx -(b/v_g)P$$

Equation 11.9-16
We have already come across this behaviour before: in an absorbing medium, the power carried by an optical wave travelling in the z-direction varies with distance as \( P(z) = P(0) \exp(-\alpha_p z) \), where \( \alpha_p \) is the power absorption coefficient (equal to \( 2\alpha \), where \( \alpha \) is the amplitude absorption coefficient of Chapter 2). In this case, \( \frac{dP}{dz} = -\alpha_p P \). By comparison with Equation 11.9-16, we may extract the power absorption coefficient as \( \alpha_p \approx \frac{b}{v_g} \). We have therefore found a satisfying link between a macroscopically observable quantity (\( \alpha_p \)) and microscopic events (interband transitions).

(II) PHOTOGENERATION OF CARRIERS

We now make an approximation to the electron rate equation that can describe the photogeneration of carriers. If diffusion currents may be neglected, and a steady-state again exists, so that \( \partial(\Delta n)/\partial t = 0 \), Equation 11.9-13 reduces to:

\[
\Delta n \approx \alpha_p v_g \tau_e \Delta \phi \approx \alpha_p \tau_e \Delta \Phi
\]

11.9-17

This implies that absorption of photons will give rise to an increase in \( n \) (and similarly in \( p \)). We might use this mechanism in a photodetector, which works by measuring the corresponding increase in conductivity. This can be done by applying a voltage to contacts on either side of a slab of material, as shown in Figure 11.9-1. Any change in the current through the circuit is then linearly proportional to the intensity of the illuminating beam. The output may be converted into a voltage, by measuring the voltage dropped across a load resistor \( R_L \).

![Figure 11.9-1 A photoconductive detector.](image)

This device is called a photoconductive detector. While it is extremely simple, it suffers from two important disadvantages. Firstly, the conductivity will not be zero, when the optical beam is switched off, due to the inevitable presence of thermally generated carriers. Consequently, there will be a DC voltage across \( R_L \). Although this may be blocked by the capacitor \( C \), thermal carriers will also be responsible for noise in the output. Secondly, the value of \( \Delta n \) depends on \( \tau_e \), so for high sensitivity we require a long lifetime. However, a large value of \( \tau_e \) implies that the photo-generated carriers will persist for some time after the beam is switched off (remember that we showed in Equation 11.7-25 that the decay of \( \Delta n \) from an initial value follows \( \Delta n(t) = \Delta n(0) \exp(-t/\tau_e) \)). It is therefore difficult to combine high sensitivity with a high-speed response.

As we will see later, it is possible to construct different types of detector, which do not have these limitations. These are based on p-n junction diodes, and are essential for the success of modern lightwave communications systems.
(III) ELECTROLUMINESCENCE

We now make an approximation that is more appropriate to the electrical generation of light through spontaneous emission (known as electroluminescence), which will be useful in describing the operation of LEDs. Assuming this time that absorption is unimportant, and that photon concentration gradients are negligible, the photon rate equation may be approximated by:

$$\frac{\partial(\Delta \phi)}{\partial t} = \frac{\Delta n}{\tau_{rr}}$$

Making a similar approximation to the electron rate equation we get:

$$\frac{\partial(\Delta n)}{\partial t} = \frac{1}{e} \frac{\partial I}{\partial x} - \frac{\Delta n}{\tau_e}$$

These equations show that the existence of a diffusion current will lead to a rise in $\Delta n$, which will in turn result in an increase in $\Delta \phi$ through spontaneous emission.

(IV) OPTICAL GAIN

Finally, we will extend our model to include stimulated emission; this version of the equations will be appropriate for the description of semiconductor laser operation. We will now ignore spontaneous emission, since stimulated emission will dominate this process. Now, the rate of stimulated emission will simply depend on the product of $n$ and $p$ and the photon density. In this new situation, however, the rate of absorption must be reconsidered, since $n$ and $p$ must be very high, and we cannot assume that there will be large numbers of electrons in the valence band or empty levels in the conduction band. We can only say that there is some number $N_{val}$ of the former and $N_{con}$ of the latter, and the actual values of these quantities will depend on the degree of excitation of the system.

We may therefore modify the rate equation for a travelling photon wave (Equation 11.9-14) to get:

$$\frac{\partial \phi}{\partial t} \approx -v_g \frac{\partial \phi}{\partial z} - b' \phi \{ N_{val} N_{con} - np \}$$

where $b'$ is a new constant. Note that $\phi$, $n$ and $p$ appear in the equation, rather than $\Delta \phi$, $\Delta n$ and $\Delta p$, since we are no longer considering small perturbations. Normally, $np$ is much less than $N_{val} N_{con}$, so that net absorption occurs. However, Equation 11.9-20 implies that optical gain will be possible, if we can somehow increase $np$ above $N_{val} N_{con}$. This might be achieved by taking (say) a p-type material, which already contains a large hole density, and somehow injecting a large number of electrons. Under these circumstances, there will be many more electrons in the conduction band than in the valence band, so the electron population is inverted from its normal equilibrium distribution.

There must be some particular level of excitation at which the loss of photons by absorption is exactly balanced by gain from stimulated emission. Under these circumstances, the material is said to be transparent. If we define the electron concentration required for transparency to be $n_0$, there will be a net gain of photons by stimulated emission if $n$ is increased above $n_0$. In this case, we may write:

$$\frac{\partial \phi}{\partial t} \approx -v_g \frac{\partial \phi}{\partial z} + G \phi (n - n_0)$$
where $G$ is the **gain constant** for the material. Typical values for $n_e$ and $G$ are $10^{24} \text{ m}^3$ and $10^{12} \text{ m}^3/\text{s}$, respectively. At this point, the amplification of a travelling optical wave will be feasible. It should be noted that the generation of photons by stimulated emission is not isotropic; the stimulated photons travel in essentially the same direction as the incident photons.

Further modification to the equations (e.g. to include all three optical processes at once) are obviously possible. This is actually required in the case of a semiconductor laser - at low injection levels, spontaneous emission is the dominant process, but stimulated emission takes over at higher levels. However, we shall postpone this complication until Chapter 12.
PROBLEMS

11.1. Show that the energy density of radiation emitted by a black body in an interval between the wavelengths $\lambda_0$ and $\lambda_0 + d\lambda$ is given by $\rho(\lambda_0) \, d\lambda = \left(\frac{8\pi \hbar c}{\lambda_0^5}\right) \{\exp(\frac{\hbar c}{\lambda_0 kT}) - 1\} \, d\lambda$. Sketch the variation of $\rho(\lambda_0)$ with $\lambda_0$, for two different temperatures - say, 1000 K and 6000 K. Show that $\rho$ peaks at a wavelength defined by Wien's law $\lambda_{max} T = \text{constant}$, and calculate the value of the constant.

[2.9 x 10^{-3} \, \text{m K}]

11.2. What is the value of accelerating voltage required to obtain an electron wavelength of 0.05 nm in vacuum? What is the corresponding electron momentum?

[602 V; 1.32 x 10^{-23} \, \text{Ns}]

11.3. An electron of energy $E$ is incident on the potential step shown below. By assuming a suitable solution for the wavefunction on either side of the step, calculate the probability that the electron will be reflected by the step, for the two cases of (a) $E > V$, and (b) $E < V$.

11.4. Calculate the wavelength at which the materials in Table 1.4-1 should first start to absorb light. Which material would you use to fabricate a) a photoconductive detector, and b) a transparent window, for operation at $\lambda_0 = 1.55 \, \mu\text{m}$. Compare your choice for a) with the ternary alloy Ga_{0.47}In_{0.53}As ($E_g = 0.73 \, \text{eV}$, direct-gap).

11.5. Calculate the position of the Fermi level in intrinsic InP, assuming that $E_g = 1.35 \, \text{eV}$ at 300 K, and that $m^*/m = 0.077$, $m^*/m = 0.64$, where $m$ is the mass of a free electron.

[0.716 \, \text{eV above the valence band edge}]

11.6. Using the data given in the table below, compare the intrinsic carrier concentration $n_i$ in Ge, Si and GaAs at a temperature of 300K. Assuming that $N_c$ and $N_v$ are roughly independent of temperature, estimate the rise in temperature needed to double $n_i$ in Ge.

<table>
<thead>
<tr>
<th>Material</th>
<th>$N_c$ (m^{-3})</th>
<th>$N_v$ (m^{-3})</th>
<th>$E_g$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ge</td>
<td>$1 \times 10^{25}$</td>
<td>$2.8 \times 10^{25}$</td>
<td>0.66</td>
</tr>
<tr>
<td>Si</td>
<td>$6 \times 10^{24}$</td>
<td>$1.04 \times 10^{25}$</td>
<td>1.12</td>
</tr>
<tr>
<td>GaAs</td>
<td>$4.7 \times 10^{23}$</td>
<td>$7 \times 10^{24}$</td>
<td>1.42</td>
</tr>
</tbody>
</table>

[2.3 x 10^{19} \, \text{m}^{-3}; 6.95 x 10^{15} \, \text{m}^{-3}; 2.25 x 10^{12} \, \text{m}^{-3}; 16.3 \, \text{K}]

R.R.A.Syms and J.R.Cozens Optical Guided Waves and Devices 42
11.7. Which of the following can act as (a) donors and (b) acceptors in GaAs: Be, C, Cd, Mg, S, Se, Si, Sn, Te, Zn? A sample of GaAs is doped with $10^{18}$ Zn atoms per m$^3$. Is the material p-type or n-type? Using the data from Question 11.7, find the position of the Fermi level and the majority and minority carrier densities, assuming complete ionization of the dopants. 

[0.408 eV above the valence band edge; $10^{18}$ m$^{-3}$; $5 \times 10^6$ m$^{-3}$]

11.8. Explain why the likelihood of occupation of an acceptor level may be found using the Fermi-Dirac probability function as $F(E_a) = 1/[1 + \exp((E_a - E_F)/kT)]$. Using this expression, check the validity of assuming complete ionization of the dopant atoms in Zn-doped GaAs at room temperature. You may take it that the impurity level for Zn lies 0.031 eV above the valence band edge. How would the density of ionized dopant atoms be calculated in Sn-doped GaAs?

11.9. Show that if stimulated emission is not included in the calculation, the black-body radiation law modifies to $\rho(\omega) \, d\omega = \left(\frac{\hbar \omega^3}{\pi^2 c^3}\right) \exp(-\hbar \omega/kT) \, d\omega$. When would this represent a good approximation? Derive a different approximation, valid for low photon energies.

11.10. Assuming that $n_0 \approx 10^{24}$ m$^{-3}$, and $G = 10^{-12}$ m$^3$/s, estimate the electron density required to achieve an optical power gain of $10^6$ m$^{-1}$ in a material of refractive index 3.5. 

[8.7 \times 10^{25}$ m$^{-3}$]
SUGGESTIONS FOR FURTHER READING

Kittel C. "Introduction to solid state physics" John Wiley and Sons, New York, 3rd Edn., (1966); Chs. 9 and 10.
Pais A. "Subtle is the Lord ... The science and life of Albert Einstein" Oxford University Press, Oxford (1982); Chs. 19, 21 and 23.
Yariv A. "Quantum electronics" John Wiley and Sons, New York, 3rd Edn., (1989); Chs. 1, 2 and 11.
OPTOELECTRONIC DEVICES

12.1 THE P-N JUNCTION DIODE

This Chapter is devoted to the wide variety of optoelectronic devices that are made in semiconductor materials. Many (although not all) involve the absorption and emission of light, and most are based around the surprisingly versatile structure mentioned at the end of Chapter 11, the p-n junction diode. We shall begin by considering the electronic aspects of this device.

In its simplest form, a p-n junction diode can be constructed as a homojunction, using an abrupt change between p- and n-type regions in a single piece of semiconductor. Figure 12.1-1 illustrates the formation of a diode, in a one-dimensional 'thought experiment' based on the ideas of the previous Chapter. The p- and n-regions might initially be considered to be separated slabs of material. When these are placed in intimate contact, the difference in hole concentration between the p- and n-type regions should result in the diffusion of holes from the former into the latter (Figure 12.1-1a). In the same way, electrons ought to diffuse in the reverse direction, from the n-type material into the p-type. If the two carrier types meet near the interface, they will recombine. The effect should then be that a layer - known as the depletion region - will be almost entirely denuded of carriers (Figure 12.1-1b).

![Diagram of p-n junction diode](image)

Figure 12.1-1 Formation of a p-n junction between p- and n-type material: a) counterdiffusion of majority carriers, and b) the fully formed junction.

A logical question to ask is, what determines the width of the depletion region? For example, is it possible that the counter-diffusion process would continue indefinitely, until all the free carriers have been used up, so that the depletion region extends right through both layers of material? The short answer is that the process is self-limiting. As we shall see, carrier recombination in the depletion region sets up a space charge; this in turn induces an electric field, which acts as an obstacle to further motion of carriers, so that the depletion region quickly stabilises as a thin layer. However, in practice, the dynamics of junction formation are irrelevant to the operation of the device, and we need only consider the situation after equilibrium has been established.

THE DISTRIBUTION OF CHARGE AND ELECTRIC FIELD INSIDE THE DEPLETION LAYER

Due to the lack of free carriers, the depletion region must be locally charged, since the charges of the dopant ions that are embedded in the lattice are no longer compensated by those of the mobile carriers. The sign of this 'exposed charge' varies. On the p-side, for
example, it is negative, due to exposure of the charge of the ionised acceptors. In contrast, the charging is positive on the n-side, because the positive charges of donor ions are now exposed. To a first approximation, the charge distribution within each region is uniform (although we shall review this later). Thus, assuming total ionization of the dopant atoms, a space extending from (say) $x = -x_{dp}$ to $x = 0$ in the p-type material may be assumed to have a local charge density $\rho = -eN_a$, while in the n-type material the space between $x = 0$ and $x = x_{dn}$ has a charge density $eN_d$ (Figure 12.1-2a). Here $x_{dn}$ and $x_{dp}$ are the lengths of the p- and n-sides of the depletion layer; these are related to the total width $w$ of the region by $w = x_{dp} + x_{dn}$.

![Figure 12.1-2 a) charge density and b) electric field, for a p-n junction in equilibrium.](image)

Now, when taken as a whole, the complete structure must be electrically neutral, so the exposed charges must balance each other. We may therefore put:

$$x_{dp}N_a = x_{dn}N_d$$

12.1-1

This allows us to write the width of the depletion layer as:

$$w = x_{dn}\left\{1 + N_d/N_a\right\}$$

12.1-2

Note that the 'split' of the depletion layer between the p- and n-sides may be highly unequal, if $N_a \neq N_d$. For example, the p-side might be much more heavily doped than the n-side; in this case (known as a p'-n junction), we find that $x_{dp} \ll x_{dn}$, so the depletion layer extends further into the n-side.

We can use Figure 12.1-2a to find the electric field resulting from the charge distribution. The calculation is based Gauss’ law, which requires that the electric flux flowing out of a closed surface equals the charge contained inside it. Taking the enclosed volume to be a parallel-sided slab of unit cross-sectional area lying between two planes at $-x_{dp}$ and $x$, the enclosed charge is given by:
This must give rise to an equal flux density, which (by symmetry) has only an x-component $D_x$. As a result, the corresponding electric field $E_x = \frac{D_x}{\varepsilon}$ is given by:

$$E_x = -(eN_a / \varepsilon) (x + x_{dp}) \quad (\text{for} \quad -x_{dp} \leq x \leq 0)$$

$$E_x = -(eN_d / \varepsilon) (x_{dn} - x) \quad (\text{for} \quad 0 \leq x \leq x_{dn})$$

where $\varepsilon$ is the dielectric constant of the semiconductor. This electric field variation is shown plotted in Figure 12.1-2b. As can be seen, the field varies linearly on each side of the depletion region. On the p-side, it falls from $E_x = 0$ at $x = -x_{dp}$ to $E_x = -(eN_a / \varepsilon) x_{dp}$ at $x = 0$; similarly, on the n-side, it rises from this minimum value to zero at $x = x_{dn}$. Now, the existence of an electric field within the depletion region suggests that there will be a voltage between different points. We might therefore find the p.d. between (for example) the left-hand edge of the depletion region and an arbitrary point $x$ by direct integration, i.e. as:

$$V(x) = -\int_{-x_{dp}}^{x} E_x \, dx$$

Substituting from Equation 12.1-4, we then obtain:

$$V(x) = (eN_a / \varepsilon) \left\{ \frac{x^2}{2} + xx_{dp} + \frac{x_{dp}^2}{2} \right\} \quad (\text{for} \quad -x_{dp} \leq x \leq 0)$$

$$V(x) = (eN_d / \varepsilon) \left\{ xx_{dn} - \frac{x^2}{2} \right\} \quad (\text{for} \quad 0 \leq x \leq x_{dn})$$

This distribution is shown schematically in Figure 12.1-3. Throughout the depletion region, $V(x)$ varies parabolically, following a smooth S-shaped curve from zero at $-x_{dp}$ to a maximum (which we shall refer to as the built-in voltage, $V_{bi}$) at $x_{dp}$. Since the electric field is zero at either end of the depletion layer, $dV/dx$ is also zero at these points.
Conversely, we can find the width \( w \) in terms of the built-in voltage by inverting Equation 12.1-8, as:

\[
w = \sqrt{(2\epsilon V_{bi}/e) (N_a + N_d) / N_a N_d}
\]

At this point, we have found a closed-form expression for the depletion layer width, which contains just one unknown parameter, \( V_{bi} \). We shall now consider the factors that determine \( V_{bi} \) itself.

THE BAND DIAGRAM FOR A P-N JUNCTION

We start by returning to the results of Chapter 11, where we drew energy band diagrams for isolated regions of n- and p-type material. How should these be modified for the composite p-n junction? We would expect some features - the relative spacings of the conduction band, the Fermi level and the valence band, for example - to be preserved at a reasonable distance from the depletion layer. Inside the depletion region, the band structure will clearly be modified. However, we may make one deduction: the Fermi level (for which the occupation probability is 0.5) must be the same in all regions, because this level is derived purely from considerations of thermal equilibrium. Armed with these two rules, we can guess that the band diagram for the p-n junction must be constructed from the two diagrams of Figure 11.6-2 as shown in Figure 12.1-4.

We start by considering the situation outside the depletion region. Here the bands are flat, and the main features are spaced as before, but the absolute values of the energy levels in the two halves of the structure are shifted to make the Fermi levels coincide. After this shift, the conduction band level is clearly different in the p- and n-regions. We will define the values of \( E_c \) in these two regions as \( E_{cp} \) and \( E_{cn} \). Now, remember that the position of the Fermi level in n-type material was previously defined relative to the conduction band, in terms of the density of donor atoms. Using Equation 11.7-13, we may therefore define the Fermi level \( E_{Fn} \) as:

\[
E_{Fn} \approx E_{cn} - kT \log_e (N_c/N_d)
\]

Similarly, the position of the Fermi level in p-type material can be defined relative to the valence band in terms of the acceptor density. However, it will be more convenient to relate \( E_{fp} \) to \( E_{cp} \) instead. We can do this by combining Equations 11.7-12 and 11.7-14, which gives:
We now have two results for the Fermi level, one for each side of the junction. According to our previous argument, they should be equal. Equating them allows us to find the difference between the conduction band levels on the p- and n-sides, as:

\[ E_{cp} - E_{cn} = kT \log_e \left( \frac{N_a N_d}{n_i^2} \right) \]

12.1-12

Now, a difference in the energy of a conduction electron on the two halves of the junction represents a potential barrier, which acts to prevent the transit of electrons across the depletion layer. There is a similar barrier to hole motion, which can be calculated from the difference between the valence band levels on the p- and n-sides. Overall, the energy barrier is equivalent to a built-in voltage across the structure of \( V_{bi} = \frac{E_{cp} - E_{cn}}{e} \). \( V_{bi} \) is therefore given by:

\[ V_{bi} = \frac{(kT/e) \log_e \left( \frac{N_a N_d}{n_i^2} \right)}{e} \]

12.1-13

Equation 12.1-13 is the missing link, which will allow us to calculate the width of the depletion layer.

DESIGN EXAMPLE

We can find the parameters of a typical p-n junction as follows. Assuming that the material is silicon (for which \( n_i = 1.4 \times 10^{16} \) m\(^{-3}\) at \( T = 300 \) K), and that the donor and acceptor densities in the p- and n-regions have the typical values of \( N_a = 10^{24} \) m\(^{-3}\), \( N_d = 10^{22} \) m\(^{-3}\), respectively (so this is a p\(^+\)-n junction), Equation 12.1-13 implies that the built-in voltage is 

\[ V_{bi} = \frac{1.38 \times 10^{-23} \times 300}{1.6 \times 10^{-19}} \log_e \left( \frac{10^{24} \times 10^{22}}{2 \times 10^{32}} \right) = 0.82 \text{ V} \]

The depletion layer width is then given by Equation 12.1-9 (assuming that \( \varepsilon = 11.7 \varepsilon_0 \) for Si) as

\[ w = \sqrt{\frac{2 \times 11.7 \times 8.85 \times 10^{-12} \times 0.82 / 1.6 \times 10^{-19} \times (10^{24} + 10^{22}) / (10^{24} 	imes 10^{22})}{(2 \times 11.7 \times 8.85 \times 10^{-12} \times 0.82 / 1.6 \times 10^{-19})}} \text{ m} = 0.33 \mu\text{m} \]

Most of this layer extends into the n-region, since \( N_a \gg N_d \). The salient features of a p-n junction are therefore a built-in voltage of the order of a volt, dropped across a layer less than a micron wide. This is equivalent to an electric field of \( \approx 2.5 \times 10^6 \) V/m, a high value.

THE BAND DIAGRAM INSIDE THE DEPLETION REGION

We may exploit the relation between voltage and the energy of a conduction electron once again, to fill in the missing features of the band diagram inside the depletion region. We have already seen that the voltage \( V(x) \) relative to the left-hand edge of the depletion layer varies parabolically. Consequently, the conduction band energy level will also be spatially varying, following:

\[ E_c(x) = E_{cp} - eV(x) \]

12.1-14

The conduction band therefore bends parabolically across the depletion layer, as in Figure 12.1-4. The valence band varies similarly, maintaining a constant energy gap \( E_g \) between the two levels.
Knowledge of this band structure may be used to calculate the free carrier distribution throughout the depletion layer. For example, if we substitute Equation 12.1-14 into Equation 11.7-4, we can find the variation of the conduction electron density as:

\[
n(x) = N_c \exp\{-[\varepsilon_c(x) - \varepsilon_F]/kT\}
\]

A typical variation of \(n(x)\) is shown in Figure 12.1-5. Note that \(n(x)\) does not abruptly fall to zero at the edge of the depletion layer, as we have assumed so far; instead, the variation is smooth. However, the drop from the equilibrium value in the n-type material outside the depletion layer (labelled \(n_n\)) is sufficiently rapid for our assumed charge distribution (Figure 12.1-2a) to be a good approximation to the truth.

\[
\text{Figure 12.1-5 Carrier distribution for a p-n junction in equilibrium.}
\]

Our model is therefore reasonably self-consistent. The corresponding hole density variation \(p(x)\) is also shown. The equilibrium hole density \(p_p\) in the p-type material may be vastly different from \(n_n\) if the junction is asymmetric (i.e., if \(N_a \neq N_d\)).

Note also that the value of \(n\) at the depletion layer edge in the p-region (which we might call \(n_p\)) is not identically zero, although it will generally be fairly small. From Equation 12.1-15, we can see that \(n_p = n(x_{dp}) = \exp\{-(\varepsilon_{cn} - \varepsilon_F)/kT\}\), so that \(n_p = n(-x_{dp})\) is given by:

\[
n_p = n_n \exp(-eV_{bi}/kT)
\]

The value of \(p\) at the depletion layer edge in the n-region (termed \(p_n\)) is found in a similar way as:

\[
p_n = p_p \exp(-eV_{bi}/kT)
\]

where \(p_p = \exp\{-(\varepsilon_p - \varepsilon_{vp})/kT\}\), and \(\varepsilon_{vp}\) is the valence band level in the p-region (Figure 12.1-4). Given that \(V_{bi}\) is significantly greater than 25 mV (i.e. kT/e at 300 K), we find that \(n_p, p_n \ll n_n, p_p\).

We can use these distributions to examine the details of the equilibrium inside the depletion layer, for example for electron motion. In Chapter 11, we saw that the current density for electrons is composed of two components (due to drift and diffusion, respectively). If the two contributions are calculated, it can be shown that they balance exactly at any point inside the depletion layer, and that the net current density is zero (although each component may individually be very large).

A similar dynamic balance may be found for holes, so that equilibrium is established for both types of charge carrier. This contrasts with the situation outside the depletion layer;
here there is no field, and no concentration gradient either, so that the two current components are zero independently for both species.

THE BAND DIAGRAM OF A BIASED P-N JUNCTION

Suppose that electrical contacts are now attached to the two ends of the diode, so that external voltages can be applied. We shall assume that these are metal, and of a particular type that forms an ohmic contact with the semiconductor. In this case, the metal can supply or receive any number of electrons, so the carrier density at the interface will be the equilibrium value. For example, aluminium forms an ohmic contact on p- or heavily doped n-type Si. Ohmic contacts may be also made to intrinsic or n-type Si, via an intermediate layer of n’-Si. However, other combinations (e.g., Al on n - GaAs) form a rectifying contact called a Schottky barrier, which is itself a form of diode.

Since the depletion layer has a very low carrier concentration, we would expect the bulk of any applied voltage to be dropped across this highly resistive layer. If we make the n-type side more negative than the p-type by V volts - a condition known as forward bias - this will tend to reduce the built-in barrier from $eV_{bi}$ to $e(V_{bi} - V)$. As a result, the band diagram will alter as shown in Figure 12.1-6.

On the other hand, if the n-side is made more positive than the p-side - when the junction is said to be in reverse bias - the applied voltage will tend to increase the size of the built-in potential barrier. In this case, the band diagram will modify as shown in Figure 12.1-7.

![Figure 12.1-6 Energy band structure of a p-n junction in forward bias.](image)

![Figure 12.1-7 Energy band structure of a p-n junction in reverse bias.](image)
CURRENT-VOLTAGE CHARACTERISTICS

The reduction in the barrier height in a forward-biassed junction will naturally make it easier for electrons to travel across the depletion region from the n-side to the p-side, and for holes to move in the reverse direction. The arrival of these carriers at the depletion layer edges will then cause a change in the minority carrier densities. The new values $n_p'$ and $p_n'$ may be found by substituting $V_{bs} - V$ for $V_{bs}$ in Equations 12.1-16 and 12.1-17. If this is done, it is easy to see that:

$$n_p' = n_p \exp(eV/kT) \quad \text{and} \quad p_n' = p_n \exp(eV/kT)$$  \hspace{1cm} 12.1-18

Clearly, $n_p'$ will be much greater than the equilibrium value $n_p$ (and similarly $p_n' \gg p_n$) for values of $V$ greater than about 25 mV. Even a small forward bias will therefore produce a considerable increase in minority carrier density near the depletion layer edges. This is called minority carrier injection. Now, any such increase in $n_p$ (for example) must follow from an actual transport of electrons from the n- to the p-side. However, it is important to recognise that the densities of the majority carriers are generally so much greater than those of the minority carriers that the fractional change in $n_n$ is very small. Its value (and that of $p_p$) may thus be considered unchanged to a good approximation.

The fate of the injected carriers depends on the geometry. If the diode is short enough, recombination may be neglected, and the carriers simply diffuse to the contacts. In this case, we must return to the results of Chapter 11, where we showed that the equations governing carrier diffusion had the form $\partial n, p/\partial t = D_e, h \partial^2 n, p/\partial x^2$. In the steady-state, $\partial n, p/\partial t = 0$, so the carrier densities must vary linearly with distance. The solution is then as shown in Figure 12.1-8; on the p-side, the electron concentration falls linearly from $n_p'$ at the depletion layer edge ($x = -x_{dp}$) to the equilibrium value $n_p$ at the contact ($x = -x_p$). Similarly, on the n-side, the hole concentration falls from $p_n'$ to $p_n$ between $x = x_{dn}$ and $x = x_n$.

Despite the absence of appreciable fields outside the depletion region, some current must still flow to the contacts, because the concentration gradients give rise to diffusion currents. The total current density is the sum of contributions from electron and hole diffusion, namely:

$$J = D_e e \frac{dn}{dx} - D_h e \frac{dp}{dx} = D_e e \{n_p' - n_p\}/L_p + D_h e \{p_n' - p_n\}/L_n$$  \hspace{1cm} 12.1-19

where the lengths $L_p$ and $L_n$ are given by $L_p = x_p - x_{dp}$ and $L_n = x_n - x_{dn}$. Note that if the doping is unequal, the two components in Equation 12.1-19 may differ vastly in size. For example, in a p+-n junction, the majority hole density $p_p$ is much greater than the...
corresponding electron density $n_e$, so the minority hole densities $p_n'$ and $p_n$ are greater than $n_e'$ and $n_p$. The current is then carried mainly by holes. Substituting for $n_e'$ and $p_n'$, we find the following relation between $J$ and $V$:

$$J = J_s \{ \exp(eV/kT) - 1 \}$$  \hspace{1cm} 12.1-20

where $J_s$ (known as the **reverse saturation current density**) is given by:

$$J_s = D_e n_p/L_p + D_h p_n/L_n$$  \hspace{1cm} 12.1-21

This is the well-known exponential characteristic of a diode, shown in Figure 12.1-9. For a device of cross-sectional area $A$, the current is simply $I = J \times A$, so the $I - V$ characteristic has a similar shape.

When $V$ is zero, the current is also zero. For a small positive voltage, however, the forward current can be extremely high. In contrast, the reverse current rapidly saturates at $I_r = J_s \times A$ (which is typically small) when $V$ is negative. As a result, the diode acts as a rectifier, presenting an extremely high impedance to reverse voltages and a low impedance to forward ones. Note that the behaviour of a real diode departs from the ideal at high reverse voltages, a feature we will return to shortly.

![Figure 12.1-9 J-V characteristics for a p-n junction diode.](image)

**THE EFFECT OF RECOMBINATION**

We now examine how this characteristic is modified, if the distance to the contacts is large enough for recombination to be important (which is the case in light-emitting devices). Once again, we start by returning to the results of Chapter 11. In the steady state, the electron density (for example) on the p-side is governed by:

$$D_e \frac{\partial^2 n}{\partial x^2} - \frac{(n - n_p)/\tau_e = 0}{\tau_e}$$  \hspace{1cm} 12.1-22

where we have written $n_n$ for the equilibrium level $n_{E}$. This is a standard second-order differential equation, which can be solved exactly, subject to the boundary conditions of $n = n_e'$ at the edge of the depletion layer ($x = -x_{dp}$) and $n = n_p$ at the contact ($x = -x_p$). Provided $x_{dp}$ is sufficiently large, however, a reasonable approximation to the solution is given by:
n(x) - n_p = \{n_p' - n_p\} \exp\{(x + x_{dp}) / \sqrt{(D_e \tau_e)}\)

Equation 12.1-23 shows that the excess electron density decays exponentially away from the depletion layer edge, falling to 1/e of its initial value in a length \(L_e = \sqrt{(D_e \tau_e)}\). This is known as the **diffusion length** of electrons in the p-type material. The excess hole density on the n-side varies in a similar way, decaying to 1/e of its initial value in a corresponding length \(L_h = \sqrt{(D_h \tau_h)}\).

Now, the value of \(J_e\) at \(x = -x_{dp}\) is the total injected electron current density. Performing the necessary differentiation, we find that:

\[J_e = D_e e\{n_p' - n_p\} / L_e\]

This is qualitatively very similar to the equivalent term in Equation 12.1-19, apart from the replacement of \(L_p\) by \(L_e\). Consequently, we may deduce that our later result for the \(J - V\) characteristic (Equations 12.1-20 and 12.1-21) will still be valid, merely provided we replace \(L_p\) by \(L_e\) and \(L_n\) by \(L_h\). A diode therefore still behaves as a rectifier, even when recombination is significant.

**THE EFFECT OF VOLTAGE ON THE DEPLETION LAYER WIDTH**

We now consider how the width of the depletion layer is modified by the application of a voltage. To find the new width \(w(V)\), we merely substitute \(V_{bi} - V\) for \(V\) in Equation 12.1-9, which gives:

\[w(V) = \sqrt{1 - V / V_{bi}} w(0)\]

where \(w(0)\) is the width of the unbiased junction. The effect of a negative value of \(V\) is therefore to increase the depletion layer width, while a positive value has the opposite effect.

Now, according to Equation 12.1-25, \(w\) can become negative for voltages greater than \(V_{bi}\); this is clearly unrealistic, and merely shows that our original approximation concerning the charge distribution inside the depletion layer has broken down at this point.

**AVALANCHE BREAKDOWN**

At high reverse voltages, our approximations will also break down (although in a slightly different manner). Consider the band diagram shown in Figure 12.1-10. An electron moving to the right will travel down the sloping region of the conduction band in a series of 'hops', with each abrupt decrease in energy corresponding to a collision with the lattice.

![Figure 12.1-10](image-url) Carrier multiplication by avalanche breakdown under high reverse bias.
If the electric field is high enough, the electron may gain a considerable amount of kinetic energy between collisions. Consequently, it is possible for the energy given up at a collision to be greater than $E_g$. In this case, the energy is sufficient to ionise a lattice atom, by promoting a valence electron to the conduction band. If this occurs, the number of carriers will increase, since there will now be two electrons moving to the right, and one hole moving to the left. After further acceleration by the field, the electrons may perform further ionisations, and so on. Under high reverse bias, a shower of new carriers may therefore be created. This leads to a sudden rise in the reverse current - a behaviour known as **avalanche breakdown** - at the characteristic breakdown voltage $V_{br}$ shown in Figure 12.1-9. This effect is exploited in the avalanche photodiode, which we will describe later.

**HETEROJUNCTION DIODES**

Up until now, we have assumed that the materials on either side of the junction are of essentially the same type, with any differences introduced through doping. This need not be the case - for example, we have already mentioned the metal-semiconductor junction. Alternatively, different semiconductors may be used. Clearly, there are certain constraints; it must be feasible to grow the layered structure as a perfect crystal, which requires the two materials to be lattice-matched. Despite this restriction, it is entirely possible to fabricate diodes with different band-gaps on either side of the junction. In this case, the structure is known as a **heterojunction**.

Heterojunctions are of great importance in optoelectronics, since they allow two important new features to be built in to the basic diode structure. Firstly, the difference in refractive index between the two materials may be used to provide optical confinement, as in the heterostructure waveguides described in Chapter 9. Secondly, the modification to the energy band diagram introduced by the variation in $E_g$ may be used to provide different potential barriers to the motion of electrons and holes, allowing each type of carrier to be confined independently to a particular region of space. As a result, the performance of light-emitting devices may be improved enormously.

The energy band diagram of a typical heterojunction is shown in Figure 12.1-11, under equilibrium conditions. Here the energy gap $E_{g1}$ on the p-side is greater than the value $E_{g2}$ on the n-side, so the structure is known as a P-n anisotype heterojunction (the upper-case letter denoting the material with the larger energy gap). Outside the depletion layer, the diagram is constructed as before, using the continuity of the Fermi level to determine the absolute levels of the conduction and valence bands. Inside the depletion layer itself, however, the construction is slightly different.

The analysis required to determine the band variation may be adapted from that used for the homojunction. The exposed charge distribution of Figure 12.1-2a may again be used; however, the electric field variation shown in Figure 12.1-2b must be modified, to allow for the different dielectric constants $\varepsilon_1$ and $\varepsilon_2$ in the two materials. The former value must be used in the upper line in Equation 12.1-4, the latter in the lower one. Integration of the electric field to find the potential variation $V(x)$ across the depletion layer may then be carried out much as before. If this is done, a slightly different proportion of the built-in voltage $V_{bi}$ is found to be dropped across each side of the junction.

Due to the difference in energy gap between the two sides of the junction, it is no longer possible to construct the conduction and valence bands by assuming that they follow the potential variation $V(x)$. In fact, as we shall see shortly, both bands must be discontinuous. Instead, we must rely on the assumption that the vacuum level - the level above the conduction band - is continuous and follows $V(x)$. Generally, this band is displaced by an
amount $\chi$ from the conduction band level $E_c$, where $\chi$ is a quantity known as the electron affinity (the energy required to liberate an electron from the conduction band) and $\phi$ above the Fermi level, where $\phi$ is the work function.

![Energy band structure for a P-n heterojunction in equilibrium.](image)

To complete the diagram, the vacuum level is drawn in on either side of the junction. The difference in height of this level then determines the built-in voltage $V_{bi}$, allowing the width of the depletion layer to be found as before. Within the depletion layer, the variation of the vacuum level follows $V(x)$. The conduction band is then drawn in at a height $\chi_1$ below the vacuum level on the p-side, and $\chi_2$ below on the n-side, where $\chi_1$ and $\chi_2$ are the electron affinities in the two materials.

This construction leads to an abrupt discontinuity $\Delta E_c = \chi_2 - \chi_1$ in the conduction band level at the boundary between the semiconductors. The valence band is then drawn in at a height $E_{g1}$ below the conduction band level on the p-side, and $E_{g2}$ below on the n-side, leading to a discontinuity $\Delta E_v$ at the interface. The net result is that the potential barrier faced by electrons has height $eV_{bi} + \Delta E_c$, while that for holes has height $eV_{bi} - \Delta E_v$, making it harder for electrons to cross the junction, and easier for holes. The same state of affairs exists in the P-p isotype heterojunction, formed between two different p-type semiconductors. Other combinations may be envisaged where the reverse is true, so the holes face a higher barrier; this is the case in both p-N and n-N structures.

**PRACTICAL P-N JUNCTION DIODES**

p-n junction diodes may be fabricated as multilayer structures, for example by growing a layer of p-type material on an n-type substrate using an epitaxial process. In this case, the resulting junction is said to be abrupt, and its main features are as described above. Alternatively, diodes may be fabricated by diffusion, for example by heavily doping a region near the surface of an n-type substrate with acceptors. In this case, the properties of the semiconductor do not change rapidly between the p- and n-type regions, but are instead vary more gradually, following the distribution of dopant beneath the surface. In this type of graded junction, the charge distribution in the depletion region varies more smoothly than we have assumed, and the electric field distribution must also be modified accordingly. Nonetheless, the operating principles described above are still broadly correct.
12.2 ELECTRO-OPTIC SEMICONDUCTOR DEVICES

We now return to a problem we touched on briefly in Chapters 9 and 10, that of constructing electro-optic waveguide devices in semiconducting materials. Here, the difficulty is that the presence of carriers prevents the development of a high electric field across the guide. The p-n junction provides a solution to the problem. For example, Figure 12.2-1a shows a cross-section of a typical phase modulator, built from an n-type InP homostructure strip-loaded waveguide on a <100> n' substrate. The guide is capped with a p' layer, thus forming a p'-n homojunction at the top of the guide, with the bulk of the depletion region located inside the guide itself.

One ohmic contact is placed on top of the p' layer, while the other is connected to the substrate. When the junction is reverse biased, the highly resistive depletion region grows downwards, so that all of the applied voltage is eventually dropped across the guide layer. The field is then sufficient to obtain a usable index change through the relatively weak $r_{41}$ electro-optic coefficient, thus inducing an optical phase change. Figure 12.2-1b is an electron microscope photograph of a similar device, based on an InP/InGaAsP heterostructure.

Electro-optic directional coupler switches may be constructed in a similar way, using two closely spaced strip-loaded guides. For example, Figure 12.2-2 shows a cross-section through a typical directional coupler, this time constructed using $n$-on-$n'$ GaAs homostructure guides.

Figure 12.2-1 a) Cross-section of an InP integrated optic phase modulator; b) SEM photograph of an actual device (photograph courtesy P.M.Rogers, British Telecom Research Laboratories).

Figure 12.2-2 Cross-section of a GaAs integrated optic directional coupler (after H.Inoue et al., Appl. Opt. 25, 1484-1490 (1986)).
Here, a much simpler electrode structure is used, with Al electrodes deposited directly on top of the two guides. As a result, these contacts are not ohmic (although the substrate contact is), but are instead of the rectifying Schottky barrier type mentioned earlier. These also act as diodes, so that operation in reverse bias is much as described above. To switch the coupler, a negative voltage is applied to just one of the Schottky electrodes, thus inducing an electric field which desynchronises the guides.

Electrically switchable devices may even be constructed using semiconductors with weak or negligible electro-optic coefficients. We already know from Chapters 3 and 9 that the refractive index changes used in homostructure guides are obtained through variations in doping, which cause changes in the free carrier concentration. A similar index change can be created dynamically, by the electrically-controlled injection of minority carriers, this time using a forward-biased p-n junction.

For example, Figure 12.2-3 shows the cross-section of an X-switch formed in silicon (which does not exhibit a linear electro-optic effect at all). Again, a p⁺-n junction is constructed at the top of the guide. When the junction is forward-biased, holes are injected into the guiding region, thus generating a refractive index change. Optical absorption is avoided by operating at a wavelength outside that required for band-to-band transitions, e.g. $\lambda_0 = 1.3 \mu m$. Similar devices may be made in InGaAsP, and the index change caused by carrier injection is actually an order of magnitude larger than that available through the electro-optic effect. However, switching speeds are limited by the lifetime of the injected carriers, which must disappear through recombination when the forward current is switched off.

![Figure 12.2-3 Cross-section of a Si-based carrier-injection switch (after J.P. Lorenzo et al., Appl. Phys. Lett. 51, 6-8 (1987)).](image)

### 12.3 PHOTODIODES

We shall now consider the p-n junction in the alternative role of a light detector, or photodiode. In Chapter 11, we noted that absorption in a bulk semiconductor can induce band-to-band transitions, causing a measurable increase in conductivity. However, the basic weaknesses of photoconductive detection are a high intrinsic level of noise (because of the presence of thermally excited carriers), and a trade-off between speed and sensitivity (due to recombination of carriers before they can be collected). The solution is to arrange for the absorption to occur in a region devoid of thermal carriers, and then use a high electric field to sweep the photo-generated carriers apart before they can recombine. This can be done inside the depletion layer of a p-n junction. Figure 12.3-1 shows the excitation of an electron from the valence band into the conduction band in this region, by a photon of energy $hf > E_c - E_v$. Clearly, the built-in field will tend to drive the photo-generated electron...
out of the depletion region towards the n-side, and the corresponding hole towards the p-side.

![Figure 12.3-1 Generation of an electron-hole pair by a photon inside the depletion region of a p-n junction.](image_url)

Two types of photodiode are in common use. Most are surface-entry devices, where the light is input through (for example) the p+ side of a p+n junction. Ideally, the light should pass into the depletion layer without suffering significant absorption in the p+ region, because carrier pairs generated outside the depletion layer will not be separated. Once inside this layer, the absorption should be total, because any light transmitted into the n-region will also be wasted. The depletion layer should therefore be as thick as possible. Side-entry devices, where the light is passed directly into the depletion layer, are another possibility. In this case, the useful absorbing region may be extremely long. However, the depletion layer should again be thick, or the light will spill out on either side.

Once out of the depletion layer, the carriers diffuse to the contacts. As a result, there is an increase in the reverse current, so that the J - V characteristic is modified in the presence of light as shown in Figure 12.3-2. Photodiodes are normally operated at a fixed reverse bias voltage, when they act as a current source. In this case, a reverse current (known as the dark current) flows in the absence of illumination and contributes to noise. For a large reverse bias, this current will be Is.

![Figure 12.3-2 J-V characteristic for a p-n junction diode in the presence of a light flux.](image_url)
We may estimate the responsivity of a photodiode as follows. Assuming that the optical power falling on the detecting area is $P$, the number of photons arriving per second is $P/\hbar \nu = P\lambda_0/\hbar c$. However, not all of these will be absorbed, and not all the carriers generated by those that are will actually reach the contacts. We may model this by defining a **quantum efficiency** $\eta$, representing the number of useful carrier-pairs generated per photon. Clearly, the maximum possible value of $\eta$ is 100%. In this case, every photon generates one electron (which is collected on the n-side) and one hole (which arrives on the p-side). The net effect is that one unit of charge crosses the device from side-to-side. The total charge crossing in unit time, which is the **photocurrent** $I_p$, is then:

$$I_p = \eta e \lambda_0 / \hbar c$$

so the photocurrent is linearly proportional to the optical power, a most useful result which holds good over a wide dynamic range. The **responsivity** $R = I_p/P$ is therefore given by:

$$R = \eta e \lambda_0 / \hbar c$$

Assuming uniform quantum efficiency, the responsivity should then be linearly proportional to the wavelength $\lambda_0$. However, this is only the case for photons with sufficient energy to promote an interband transition. Less energetic photons will not contribute to the photocurrent at all, so the responsivity must fall to zero for $\lambda_g > \hbar c/\Eg$ (or $\lambda_g > 1.24/\Eg$, where $\lambda_g$ is in $\mu$m and $\Eg$ is in eV).

**DESIGN EXAMPLE**

We may calculate the theoretical responsivity of a silicon p-n junction photodiode as follows. $\Eg$ for Si is 1.14 eV at 300 K, so the optimum wavelength is $\lambda_g = 1.24/1.14 = 1.09 \mu$m. Assuming a uniform quantum efficiency of 100%, $R_{\text{max}} = 1.6 \times 10^{-19} \times 1.09 \times 10^{-6} / (6.62 \times 10^{-34} \times 3 \times 10^8) = 0.88 \text{ A/W}$. The theoretical responsivity is then as shown by the dashed line in Figure 12.3-3.

![Figure 12.3-3 Ideal (dashed line) and actual (full line) responsivity of a silicon p-n junction photodiode.](image-url)
In practice, the quantum efficiency of a surface-entry device is variable, and peaks at about 80%. At short wavelengths, $\eta$ is reduced, because the attenuation is so strong that the light is absorbed before it reaches the depletion layer, and at long wavelengths it falls because the light is transmitted right through this layer. Furthermore, since Si has an indirect band-gap, $\eta$ is reduced further near $\lambda_g$, because a simultaneous interaction with a high-energy phonon is required for momentum conservation in the band-to-band transition, a relatively unlikely event. The characteristics of a real Si photodiode are therefore as given by the full line in Figure 12.3-3, which shows a useful responsivity only up to $\lambda_0 \approx 0.85 \, \mu m$. For high sensitivity at the near-infrared wavelengths used in second-generation fibre communication systems, materials with smaller bandgaps are required. Ge (useful up to $\lambda_0 \approx 1.55 \, \mu m$) or GaInAs (up to 1.67 $\mu m$) are both suitable.

**PIN PHOTODIODES**

The basic limitation of a p-n junction photodiode - that the depletion layer is so thin that radiation of long wavelength is only weakly absorbed - is overcome in the p-i-n structure shown in Figure 12.3-4a (known as a **PIN photodiode**). Here, a region of intrinsic or lightly doped material is introduced between two heavily doped p- and n-type regions. Because the doping is so low in this region, the depletion layer can then be arranged to extend right through it under a modest reverse bias, yielding the electric field variation of Figure 12.3-4b. The effective depletion layer width may therefore be fixed at a value far greater than the 'natural' one, approximately the width of the intrinsic region.

Figure 12.3-4c shows a cross-section of a typical GaInAs PIN photodiode. The n- region is a mesa of Ga$_{0.47}$In$_{0.53}$As (a particular composition that is lattice-matched to InP). This is grown epitaxially on an n' InP substrate, and the p' region is then made by zinc diffusion. The structure is therefore a hybrid, part homo- and part hetero-junction. In this device, light enters through the InP substrate, which has an energy gap of 1.35 eV and so is transparent to wavelengths longer than about 0.92 $\mu m$. This avoids the absorption of any light at the important 1.3 and 1.55 $\mu m$ wavelengths outside the depletion layer. The quantum efficiency is then almost uniform between 1.0 and 1.6 $\mu m$, just below $\lambda_g$ for GaInAs.

![PIN Photodiode Diagram](image-url)
AVALANCHE PHOTODIODES

It is possible to make use of the high-field characteristics of the p-n junction to construct a photodiode with internal gain. In the avalanche photodiode (or APD), impact ionisation is used as a method of multiplying the photocurrent before it enters the circuit of any electrical amplifier, thus providing an increase in receiver sensitivity. Figure 12.3-5 shows the reach-through avalanche photodiode (RAPD), which illustrates the basic principle involved.

The RAPD is a four-layer device, with a p⁺-i-p⁻ structure (Figure 12.3-5a). Under reverse bias, most of the applied voltage is dropped across the p-n⁻ junction, although the exact fraction depends on the doping levels. As the voltage is increased, the depletion layer associated with this junction widens, until it just reaches through to the intrinsic region. The electric field distribution is then as shown in Figure 12.3-5b. In the depletion region, the field is extremely high; normally, the peak field is held at ≈ 10% below the value at which avalanche breakdown occurs. In the long intrinsic region, however, the electric field is only moderate. Light enters the structure through the p⁺ layer, and travels through to the intrinsic region, where it is absorbed. The photogenerated carriers drift in the moderate electric field, with electrons travelling towards the p-n⁻ junction. Carrier multiplication then takes place inside the high-field region by impact ionization. As a result, there is a large increase in the photocurrent, up to 1000 times, which provides the desired improvement in detector sensitivity. However, since the avalanche process is a statistical one, there is a corresponding increase in the noise level.

WAVEGUIDE-INTEGRATED PHOTODIODES

The integration of photodetectors with waveguides poses an inherent compatibility problem - the former are intended to act as efficient, localised absorbers of light, while the latter should (ideally) be lossless. A number of strategies may be used to circumvent the difficulty. Firstly, the waveguide and the detector may be constructed on different levels of a multilayer, allowing the properties of each level to be tailored individually. Light may then be transferred between the two at the appropriate point, either by directional coupling or by radiation. In the latter case, a standard component such as a grating coupler may be used, or the guide may be induced to cut off. For example, Figure 12.3-6 shows the cross-section of a structure combining a passive guide made in two glassy layers with a p-n junction photodiode formed by a p-type diffusion into an n-type silicon substrate. To couple light
into the detector, the height of the SiO₂ buffer layer is gradually reduced until the high-index substrate approaches the guide sufficiently to induce cutoff. The light is then radiated onto the detector.

Figure 12.3-6  An Si p-n junction integrated with an overlay guide (after S. Ura et al., IEEE J. Lightwave Tech. LT-4, 913-918 (1986); © IEEE 1986).

Alternatively, waveguides and photodetectors may be constructed on the same level of an integrated structure, and butt-coupled together. Clearly, this requires material properties to be tailored locally. For example, Figure 12.3-7 shows an InP/InGaAsP double-heterostructure channel guide integrated with an GaInAs PIN photodiode. Here, the waveguide is formed first, by conventional methods. A trench with dimensions ≈ 10 µm x 20 µm is then etched right through the guide to the n⁺-InP substrate. This trench is filled with n GaInAs by liquid phase epitaxy (LPE) to form the intrinsic region of the diode, and the upper layer is then Zn-doped to form the p' region. The net result is that the channel guide is coupled directly into the depletion region of a vertically oriented PIN photodiode.

Figure 12.3-7  a) Cross-section of an InGaAs PIN photodiode integrated with an InGaAsP waveguide (after W.Döldissen et al., Elect. Lett. 25, 35-37 (1989)), and b) SEM view in the region of LPE overgrowth (photograph courtesy W. Döldissen, Heinrich-Hertz-Institut).

12.4 THE LIGHT-EMITTING DIODE

We now consider another application of the p-n junction, as a light source. In this incarnation, it is known as a light-emitting diode (or LED). Now, in Chapter 11, we described electroluminescence as a mechanism for generating photons in a direct-gap semiconductor. However, an isolated lump of material will not emit significant quantities of light, because in thermal equilibrium at room temperature the number of downward electron transitions is extremely small. To improve the optical output, we must move the material far from equilibrium, so that the rate of spontaneous emission is considerably increased. This might be done by taking (for example) a p-type material, which already contains a large hole density, and pouring electrons into it. This can be done in a forward-biased p-n junction.
CHAPTER TWELVE

Figure 12.4-1 Spontaneous emission by recombination of injected minority carriers in a forward biased p-n junction.

Figure 12.4-1 illustrates the process; electrons are injected into the p-type material, where they combine with the holes already present. At the same time, holes are injected into the n-type material; however, the junction is normally highly asymmetric, so that light emission takes place mainly from one side.

THE EMISSION SPECTRUM OF AN LED

Once again, we recall that spontaneous emission generates photons that travel in random directions, so the emission is isotropic; as we will show later, this greatly reduces the external efficiency of an LED. The emission is also unpolarized. Furthermore, we note that the output does not consist only of light at the wavelength \( \lambda_g = \frac{hc}{E_g} \), as a two-state model would imply. Because of the wide distribution of carrier energies within the conduction and valence band, the output is incoherent, with a spectrum consisting of a broad range of wavelengths (Figure 12.4-2). The electron distribution normally peaks near an energy \( \approx \frac{1}{2} kT \) above the edge of the conduction band, and extends for several \( kT \) (say, \( \approx \frac{5}{2} kT \)) above \( E_c \). Similarly, the hole distribution peaks at around \( \frac{1}{2} kT \) below the valence band edge, extending for about \( \frac{5}{2} kT \) below \( E_v \). Consequently, the possible photon energies lie in the range \( E_g \) to \( E_g + 5kT \), and the most likely energy is \( \approx E_g + kT \).

The optical wavelength corresponding to an energy \( E_g + \alpha kT \) (where \( \alpha \) is an arbitrary constant) is \( \lambda_0 = \frac{hc}{E_g + \alpha kT} \). Since \( kT \) is typically much less than \( E_g \), we may put:

\[
\lambda_0 \approx \left( \frac{hc}{E_g} \right) \left\{ 1 - \frac{\alpha kT}{E_g} \right\}
\]

Figure 12.4-2 Representative emission spectrum of an LED.
If $kT$ and $E_g$ are measured in eV, and $\lambda_0$ in $\mu$m, Equation 12.4-1 may be written in the alternative form $\lambda_0 \approx (1.24/E_g) \{ 1 - \alpha kT/E_g \}$. The output from an LED therefore peaks at a wavelength $\lambda_p$, given by:

$$\lambda_p \approx (1.24/E_g) \{ 1 - kT/E_g \}$$

Similarly, the half-power bandwidth of the emission spectrum is $\Delta \lambda$, where:

$$\Delta \lambda \approx 3.1 kT/E_g^2$$

Now, $E_g$ for GaAs is 1.42 eV. For a GaAs LED at room temperature, for example, we therefore find that $\lambda_p \approx (1.24/1.42) \{ 1 - 0.0258/1.42 \} = 0.857 \mu$m. Similarly, $\Delta \lambda = 3.1 \times 0.0258 / 1.42^2 = 0.04 \mu$m, or 40 nm. For materials with smaller bandgaps, designed to emit at longer wavelengths, $\Delta \lambda$ is even larger. For example, an InGaAsP LED operating near $\lambda_0 = 1.3 \mu$m would have a spectral half-width approaching 100 nm. Consequently, the LED cannot be used as a source for high-speed, long-distance optical communications; the dispersion caused by such an extended spectrum would be far too large. LED transmitters are therefore restricted to short-haul applications.

**THE FREQUENCY RESPONSE OF AN LED**

We now consider the efficiency of the electroluminescence process inside an LED, which we might term the *internal efficiency* $\eta$. As it turns out, $\eta$ is frequency-dependent. We can illustrate this by calculating the optical output of an LED for a time-varying drive current, assuming a two-state model for simplicity. We start by returning the rate equations of Chapter 11. Considering spontaneous emission to be the dominant process, and the injected carriers to be electrons, the first relevant equation is Equation 11.9-19. Since the system will be significantly perturbed, however, we must replace terms of the form $\Delta n$ by $n$, and so on.

Furthermore, for simplicity, we shall consider the LED to be a lumped-element device, consisting of a uniform recombination region of cross-sectional area $A$ and depth $\Delta x$ (i.e., a finite volume $v = A \Delta x$). In this case, we may rewrite Equation 11.9-19 as:

$$\frac{dn}{dt} + n/\tau_e = J/e \Delta x = I/ev$$

where $I = J A$ is the total current injected into the recombination region. Now, in the general case, the current may consists of a DC bias, with a superimposed AC modulation. We might therefore put:

$$I = I' + I'' \exp(j \omega t)$$

Where $I'$ is the DC bias, and $I''$ and $\omega$ are the amplitude and angular frequency of the AC modulation, respectively. We assume that $I' > I''$, so that the LED is always under forward bias. We may take the solution for the electron density in a similar form, as:

$$n = n' + n'' \exp(j \omega t)$$

Substituting these expressions into Equation 12.4-4, we get:

$$n' + n''(1 + j \omega \tau_e) \exp(j \omega t) = \{ I' + I'' \exp(j \omega t) \} \tau_e/ev$$

Equating the DC and AC terms in Equation 12.4-7, it is clear that we must have:
\[ n' = \frac{\Gamma}{\tau_{ee}} \text{ and } n'' = \left( \frac{\Gamma''}{\tau_{ee}} \right) \left\{ \tau_{ee} / (1 + j\omega\tau_{ee}) \right\} \]

12.4-8

We now return to the corresponding photon rate equation. For spontaneous emission due to electron injection, the rate of change of photons may be taken as:

\[ \frac{d\phi}{dt} = \frac{n}{\tau_{rr}} - \frac{\phi}{\tau_{p}} \]

12.4-9

Here we have included an additional term \( -\phi/\tau_{p} \), which describes rate of loss of photons from the active volume (since it is now finite). This must be proportional to \( \phi \) itself, and the proportionality constant \( \tau_{p} \) is known as the photon lifetime. For an active volume whose longest dimension is \( L \), \( \tau_{p} \) is of the order of \( L/v_{g} \), the time taken for a photon to escape from the region. For typical device dimensions of 100 \( \mu \)m, \( \tau_{p} \) is \( \approx 10^{-12} \) seconds. Since \( 1/\tau_{p} \) is then so large, we may neglect \( d\phi/dt \) by comparison, and write:

\[ \frac{\phi}{\tau_{p}} = \frac{n}{\tau_{rr}} \]

12.4-10

For a recombination volume \( v \), the total number of photons escaping per second is \( v\phi/\tau_{e} \). This must represent the external photon flux \( \Phi \) generated by the diode. Since each photon carries energy \( h\nu = e\lambda_{g} \approx eE_{g} \), the optical power \( P \) emitted is:

\[ P \approx \frac{hc\Phi}{\lambda_{g}} = \left( \frac{hc\lambda_{g}}{\lambda_{g}} \right) \frac{vn}{\tau_{rr}} \]

12.4-11

We now assume a solution for \( P \) in the form of a DC emission plus an AC modulation term, i.e. as:

\[ P = P' + P'' \exp(j\omega t) \]

12.4-12

Substituting this expression into Equation 12.4-11, together with our previous solution for \( n \), we get:

\[ P' + P'' \exp(j\omega t) = \left( \frac{hc\lambda_{g}}{\lambda_{g}} \right) \left[ (\Gamma'_{ee}/\tau_{ee}) + (\Gamma''_{ee}/\tau_{ee}) \right] \left\{ \tau_{ee} / (1 + j\omega\tau_{ee}) \right\} \exp(j\omega t) \]

12.4-13

Equating the DC and AC components separately in Equation 12.4-13, we then obtain:

\[ P' = \eta'\frac{hc\lambda_{g}}{\lambda_{g}} \text{ and } P'' = \eta''\frac{hc\lambda_{g}}{\lambda_{g}} \]

12.4-14

where we have defined two new quantities, the DC and AC internal efficiencies \( \eta' \) and \( \eta'' \), as:

\[ \eta' = \tau_{ee}/\tau_{rr} \text{ and } \eta'' = \tau_{ee}/(\tau_{ee} (1 + j\omega\tau_{ee})) \]

12.4-15

Equation 12.4-14 shows that the optical output from an LED is linearly proportional to current, for both DC and AC injection. However, the internal efficiency in each case is slightly different. We shall now examine the implications of this result, by considering the likely values of \( \eta' \) and \( \eta'' \).

**DC INTERNAL EFFICIENCY**

Bearing in mind that \( 1/\tau_{e} = 1/\tau_{rr} + 1/\tau_{in} \), the DC efficiency \( \eta' \) has the value:

\[ \eta' = r_{in} / \{ r_{in} + r_{ex} \} \]

12.4-16
where $r_{rr} = 1/\tau_{rr}$ and $r_{nr} = 1/\tau_{nr}$ are the rates of radiative and non-radiative transitions, respectively. Consequently, the DC efficiency depends on the fraction of all downward transitions that are radiative, a physically convincing result. In simple LEDs, $\eta'$ may be as high as 50%. Substituting for the constants, the DC output power may be written in the simple form $P' = 1.24 \eta' I/\lambda_g$, where $\lambda_g$ is measured in $\mu$m.

For a GaAs LED emitting at $\lambda_g \approx 850$ nm, with $\eta' \approx 50\%$, roughly 0.73 mW of optical power should thus be generated per mA of drive current. However, this figure is lowered by absorption, and by total-internal-reflection at the surfaces of the device, as we shall show later.

**AC INTERNAL EFFICIENCY**

Combining Equations 12.4-15, we note that the AC internal efficiency may also be written as:

$$\eta'' = \eta' / (1 + j\omega \eta' \tau_{rr})$$

Since $\eta''$ is complex, it is clear that the optical modulation cannot be in-phase with the drive current modulation, but must instead lag behind. The magnitude of $\eta''$ is given by:

$$|\eta''| = \eta' / \sqrt{1 + \omega^2 \eta'^2 \tau_{rr}^2}$$

At low frequencies, the AC internal efficiency tends to the DC value, but for angular frequencies greater than the break-point $\omega_c = 1/\eta' \tau_{rr}$, $|\eta''|$ is approximately:

$$|\eta''| \approx 1/\omega \tau_{rr}$$

so that $|\eta''|$ is independent of $\eta'$ at high frequencies.

Figure 12.4-3 shows a logarithmic plot of $|\eta''|$ versus $\omega$, for three values of $\eta'$. As can be seen, there is a trade-off between the DC efficiency and the position of the break point.

Figure 12.4-3 Variation of the AC internal efficiency $\eta''$ of an LED, for three values of DC internal efficiency.
This illustrates a basic limitation of LEDs: they may be either fast, but inefficient, or slow but efficient. However, it is not possible to obtain the best of both worlds. DC efficiency may be optimised by maximising $\eta'$, by heavily doping the recombination region. The frequency response is clearly limited by $\tau_{rr}$, and in typical devices the break-point occurs at $f_c = \omega_c / 2\pi = 10 - 20$ MHz. LEDs may therefore only be used as sources for optical links of moderate bit-rate.

THE EXTERNAL EFFICIENCY OF AN LED

Despite the reasonable internal efficiency of an LED, it is an unfortunate truth that very little of this theoretical output is available for use. To illustrate this point we show in Figure 12.4-4a a typical surface-emitting LED; this is a layered p-n structure, in which the optical output is taken from the upper surface. Because the light generation is isotropic, half of the light is emitted downwards, and has no chance of escaping from this surface. Of the half that is emitted upwards, only a fraction escapes, due to the small critical angle at the semiconductor/air interface.

![Figure 12.4-4](image)

Figure 12.4-4  a) A typical surface emitting LED; b) geometry for calculation of external efficiency.

The proportion of the light that escapes into the air can be found as follows. We first define a source of radiation as the point L inside the LED and the point of escape from the surface of a given ray emitted by L as S, as shown in Figure 12.4-4b. If the angle made by this ray with the surface normal is $\theta$, and the distance between L and S is r, the fraction of the radiation emitted by L lying in the solid angle between $\theta$ and $\theta + d\theta$ is $2\pi r \sin(\theta) r d\theta / 4\pi r^2 = (1/2) \sin(\theta) d\theta$. The proportion of this fraction that actually crosses the interface is then given by $T(\theta) = 1 - \Gamma(\theta)^2$, where $\Gamma(\theta)$ is the Fresnel reflection coefficient for incidence at angle $\theta$. The external efficiency $\eta_e$ of the LED is then found as the integral of the product of the above two quantities, over the angular range for upward-travelling radiation, namely $0 \leq \theta \leq \pi/2$. We may therefore write:

$$\eta_e = \int_0^{\pi/2} \{T(\theta)/2\} \sin(\theta) d\theta$$

Now, $T(\theta)$ is a slowly-varying function, falling from a maximum value $T(0) = 4n/(n + 1)^2$ at normal incidence (where n is the refractive index of the semiconductor) to $T(\theta_c) = 0$ at the critical angle. A reasonable approximation to Equation 12.4-20 is then:

$$\eta_e = T(0)/2 \int_0^{\theta_c} \sin(\theta) d\theta = T(0)/2 \{1 - \cos(\theta_c)\}$$
Now, for typical LED materials, $n$ is large (3.5 for GaAs), so $\theta_c = \sin^{-1}(1/n) \approx 1/n$ and $\left\{1 - \cos(\theta_c)\right\} \approx \theta_c^2/2 \approx 1/2n^2$. With these approximations, we may write the external efficiency as:

$$\eta_e \approx \frac{1}{n(n + 1)^2}$$

12.4-22

For GaAs, $\eta_e \approx 0.0141$, so that only about 1.4% of the light generated inside the LED can actually escape from the surface. Consequently, the DC external output is reduced to $P_e = P'\eta_e$. For the GaAs LED discussed earlier, the usable output would be $0.73 \times 0.0141 \text{ mW} \approx 10 \mu\text{W}$ per mA drive current.

In fact, a fraction of the light is absorbed while travelling between L and S, reducing this figure still further. There are two partial solutions to the problem. Firstly, a reflector can be added to the lower surface, so that some of the light emitted downwards may be exploited. Secondly, the upper surface of the LED may be antireflection coated, which provides an improvement in $T(0)$. Thirdly, the surface may be lensed (either by etching the surface into a hemisphere or by attaching an external microlens), so that a larger proportion of the light is incident on the interface at an angle less than $\theta_c$. This solution is used extensively in display LEDs, but is inappropriate for lightwave systems, where the aim is to couple the LED output into an optical fibre. In this case, the power actually delivered into the fibre will be lower than the estimates obtained above by a factor $\eta_c$, the coupling efficiency.

Calculation of $\eta_c$ for the general case of an optical fibre butt-coupled to the surface of an LED is a relatively complicated problem. However, for an emission area that is small compared with the fibre core, and arranged on-axis as shown in Figure 12.4-5, we can obtain a useful approximation to $\eta_c$.

![Figure 12.4-5](image)

Light emitted by the LED travels in all directions, ranging from normal to the surface to parallel. It can be shown that the distribution of output intensity with external angle $\theta_e$ is Lambertian, following:

$$I(\theta_e) = I(0) \cos(\theta_e)$$

12.4-23

We may use this distribution perform a calculation very similar to the previous one. Firstly, we note that the fraction of the output lying in the solid angle between $\theta_e$ and $\theta_e + d\theta_e$ is:
\[
dI = I(\theta_e) \sin(\theta_e) \, d\theta_e / \int_{0}^{\pi/2} I(\theta_e) \sin(\theta_e) \, d\theta_e
\]

The proportion of this fraction that crosses the air/fibre interface is then given by \( T'(\theta_e) = 1 - \Gamma'(\theta_e)^2 \), where \( \Gamma'(\theta_e) \) is the Fresnel reflection coefficient for incidence at an angle \( \theta_e \).

Now, of the part that is transmitted into the fibre, only the fraction that is total-internal-reflected at the core/cladding interface will actually be guided. In Chapter 8, we showed that this condition led to a maximum acceptance angle for a multimode step-index fibre of:

\[
\theta_{\text{emax}} = \sin^{-1}\left(\sqrt{n_1^2 - n_2^2}\right) = \sin^{-1}(\text{NA})
\]

where NA is the numerical aperture of the fibre. Consequently, \( \eta_c \) may be found by multiplying Equation 12.4-23 by \( T(\theta_e) \) and integrating from \( \theta_e = 0 \) to \( \theta_e = \theta_{\text{emax}} \). Since \( \theta_{\text{emax}} \) is typically small, and \( T(\theta_e) \) is slowly-varying, \( \eta_c \) may be approximated by:

\[
\eta_c = T'(0) \int_{0}^{\theta_{\text{emax}}} I(\theta_e) \sin(\theta_e) \, d\theta_e / \int_{0}^{\pi/2} I(\theta_e) \sin(\theta_e) \, d\theta_e
\]

Where \( T'(0) = 4n'/(n'+1)^2 \), and \( n' \) is the refractive index of the fibre core. At this point, we may substitute for \( I(\theta_e) \) and perform the necessary integration. This yields the extremely simple result:

\[
\eta_c = T'(0) \sin^2(\theta_{\text{emax}}) = T'(0) \, \text{NA}^2
\]

Thus, for a step-index fibre with \( n' \approx 1.5 \) and \( \text{NA} \approx 0.2 \), we obtain a coupling efficiency of \( \eta_c \approx 0.04 \). This figure may be improved by contacting the fibre directly to the LED, or by using a lens between the two (if the emitting area is smaller than the fibre core). One alternative device (the edge emitting LED, which we will describe shortly) also offers an advantage. However, the fundamental cause of low external efficiency is the undirected nature of spontaneous emission. A major improvement may only be obtained by adopting the alternative mechanism of light production, stimulated emission.

**DOUBLE HETEROSTRUCTURE LEDS**

In place of the simple p-n junction, a more complicated structure is often used. Typically, this might be a P-n-N double heterostructure, containing three layers: a p-type layer with a wide energy gap, a narrow-gap n-type layer, and a wide-gap n-type layer. Now, we have already mentioned the way in which a single heterojunction may be used to provide potential barriers of different heights for electrons and holes; the double heterostructure may be used to provide a high barrier at different positions for electrons and holes. In this way, the recombination region may be limited to a defined region of space, resulting in an increase in the radiative recombination efficiency. We illustrate this in Figure 12.4-6, which shows the energy band diagram for a P-n-N structure.

Here the potential barrier faced by electrons is higher at the P-n interface, while that faced by holes is higher at the n-N interface. Under forward bias, electrons injected into the n-type layer from the right are prevented from diffusing away to the left by the first of these barriers. Similarly, holes injected into the n-type layer from the left are stopped from leaking away to the right by the second barrier. The net effect is that carriers may be poured into the central layer, which acts as a localised recombination region. Two further advantages stem from this construction. The first is that the larger band-gaps of the P- and N-type layers can make them effectively transparent to radiation generated in the narrow-gap recombination process.
The second is that the refractive index of the narrow-gap material at the centre of the structure may be higher than that of the material on either side. It is therefore possible to combine localisation of the recombination region with confinement of an optical field. This feature is exploited extensively in the double heterostructure laser, which will be described in the next Section.

Figure 12.4-6 Energy band structure for a P-n-N double heterostructure in equilibrium.

Figure 12.4-7 shows a typical surface-emitting double-heterostructure LED, the Burrus-type LED (named after its inventor). This is a high-efficiency device, suitable for use with multimode fibre. To fabricate the device, a double heterostructure is first grown on a substrate, and a SiO₂ isolation layer is then deposited on the lower layer of the heterojunction. This layer is etched, to expose the heterojunction over a small window. Metallization layers are then added to the upper and lower surfaces; clearly, the lower of these makes contact with the heterojunction only in the region of the window. A deep well, aligned with this window, is then etched right through the substrate to the upper layer of the heterojunction, and a multimode optical fibre is epoxied into the well in contact with the LED surface. When a forward current is passed through the LED, it flows mainly through the region immediately below the fibre. The recombination region (or active region) is therefore confined vertically by the double heterojunction and laterally by the distribution of current flow.

Figure 12.4-7 A Burrus-type double heterojunction surface-emitting LED.
THE EDGE-EMITTING LED

Figure 12.4-8 shows an alternative geometry, the edge-emitting or superluminescent LED. This is also based on a double heterostructure, but now the emission is taken from the edge of the junction rather than its surface. Generally, the current is forced to flow through a narrow strip down the device centreline by the introduction of a current-blocking silica layer. As a result, the active volume is constrained vertically by the double heterostructure and laterally by the current flow. Most importantly, light generated by spontaneous emission is also confined vertically to a certain extent; the refractive index difference at the top and bottom of the active layer results in total internal reflection. This tends to channel a large fraction of the light towards the emission windows at relatively shallow angles. While the radiation pattern is still Lambertian in the horizontal plane (with a half-power beam width of $\approx 120^\circ$), it is therefore considerably compressed in the vertical plane (into a width of $\approx 30^\circ$).

The large distance travelled before the light escapes does result in an increased level of absorption (although this may be minimised by making the active layer thin, and forcing most of the light to travel in the wide-bandgap confining layers). Consequently, the external efficiency is lower than in a surface-emitting LED. However, this is more than compensated for by the increased directivity of the output, which allows much higher coupling efficiency to a fibre. In turn, this makes a reduction in the internal efficiency possible, thus improving the high-frequency response. Antireflection coatings may again be used to improve the transmission through the emitting facet. Edge-emitting LEDs may therefore have bandwidths of $\approx 5x$ their surface-emitting counterparts, while still coupling $\approx 5x$ the power into a low-NA fibre.

Despite all this, the performance improvement offered is only moderate, and a major difference can only be obtained using the alternative mechanism of light production, stimulated emission. Surprisingly, the structure required is very similar, but it has entirely different optical characteristics. We shall now consider this device - the semiconductor laser - in detail.
12.5 THE SEMICONDUCTOR LASER I: BASIC OPERATION

To explain the difference between the semiconductor laser and the LED, we must return to the results of Chapter 11. When absorption and stimulated emission are the dominant processes, the rate equation for a travelling photon wave in our two-level model may be taken as:

$$\frac{\partial \phi}{\partial t} = -v_g \frac{\partial \phi}{\partial z} + G\phi(n - n_0)$$  \hspace{1cm} 12.5-1

In the steady-state, $\frac{\partial \phi}{\partial t} = 0$, so that:

$$\frac{\partial \phi}{\partial z} = g_p \phi$$  \hspace{1cm} 12.5-2

where the power gain is $g_p = G(n - n_0)/v_g$. In this case, we can find the variation of $\phi$ with $z$ as:

$$\phi(z) = \phi(0) \exp(g_p z)$$  \hspace{1cm} 12.5-3

Equation 12.5-3 describes an exponentially growing wave. This type of amplification ought to offer a method of light generation, if we can harness it. Clearly, we require $n > n_0$ for $g_p$ to be positive, so that net gain occurs. To achieve this, we must find some means to raise the electron density far above the equilibrium level. The injection of electrons across a forward-biased diode immediately suggests itself as a suitable method.

Figure 12.5-1 shows a three-dimensional sketch of the band-structure of a $p-n^+$ junction diode under heavy forward bias. Here the forward voltage $V$ is so large that the relative levels of the conduction band on either side of the junction are the opposite of those in equilibrium. In this case, we would expect a veritable flood of electrons to be injected into the $p$-side. An optical wave travelling parallel to and sufficiently close to the depletion layer will then move through a region with greatly increased electron concentration.

Unfortunately, our previous expression for $n_p'$ is inaccurate at the high forward bias required to invert the population. However, by a more accurate calculation, it can be shown that the net gain is obtained when the difference between the quasi Fermi-levels $E_{Fp}$ and $E_{Fn}$ is greater than the bandgap $E_g$.

Now, the injected electrons will diffuse away from the depletion layer edge, so the gain must fall off in the same direction. Arguably, the excess electron concentration will fall to zero within about a diffusion length. Since no amplification will occur outside this region, the optical wave should be confined near the depletion layer to take full advantage of the gain. One approach would be to force it to propagate in a waveguide that passes through the gain region. We might then ask, how should the guide be designed for optimum gain?

To answer this, we must find the effective power gain due to a distribution $g_p(x)$ that varies over the guide cross-section. This can be done using the methods of Chapter 9, where we calculated the effect of a nonuniform index perturbation $\Delta n$ on a guided mode in a phase modulator. There we found that the change in effective index was found from the overlap between the perturbation and the normalised transverse field, as $\Delta n_{\text{eff}} = |\Delta n \cdot |E|^2|$. It is easy to show that the effective power gain is also found as an overlap, this time between the transverse field and the gain distribution, i.e. as:
\[
g_{\text{eff}} = \langle g_p, |E|^2 \rangle
\]

Consequently, we may deduce that the most important condition for high effective gain is a strong overlap between the modal field and the gain distribution.

Figure 12.5-1 Amplification of a travelling optical wave near the depletion layer of a forward biased p-n junction.

Early semiconductor lasers were made from p-n homojunctions. These contain an in-built planar guide, formed by the increase in refractive index that occurs naturally in the depletion region due to the absence of free carriers. However, since the index change obtained this way is very small (\( \Delta n/n \approx 0.1 - 1\% \)), such a guide provides a very weak confinement of the optical field. The lack of any kind of potential barrier also allows the diffusion of injected carriers away from the junction. Consequently, homostructure lasers have a low effective gain and are of historical interest only.

A much more successful geometry is the **double heterostructure laser**, which uses compositional variations to optimise the optical and carrier confinement. For example, Figure 12.5-2a shows the band diagram for a P-n-N heterostructure under strong forward bias. Note that there is a potential well in the n-region, in both the conduction band and the valence band. As mentioned earlier, electrons injected into the upper well from the N-side are highly localised, and available for recombination with holes injected into the lower well from the P-side.

Figure 12.5-2b shows the corresponding refractive index distribution across the junction; the wide band-gap P- and N-type regions have a lower index than the n-region, so a roughly symmetric slab guide (with \( \Delta n/n \approx 5\% \)) is created exactly in the high-gain region.

The modal field (Figure 12.5-2c) is then well confined, and has a strong overlap with the gain. Furthermore, the confinement is highly stable, since it is hardly altered by changes in the injection level. Modern semiconductor lasers are therefore almost exclusively built around variants of the double heterostructure. For example, in the **separate confinement heterostructure** (SCH), optical field confinement and localisation of the recombination
region are optimised separately, using a quadruple heterojunction. The two inner junctions limit the active region, while the two outer ones confine the optical field.

![Figure 12.5-2](image)

**Figure 12.5-2** a) Energy bands, b) refractive index and c) modal field in a double heterostructure laser.

The wide band-gap P- and N-type regions have a lower index than the n-region, so a roughly symmetric slab guide (with $\Delta n/n \approx 5\%$) is created exactly in the high-gain region. The modal field (Figure 12.5-2c) is then well confined, and has a strong overlap with the gain. Furthermore, the confinement is highly stable, since it is hardly altered by changes in the injection level. Modern semiconductor lasers are therefore almost exclusively built around variants of the double heterostructure. For example, in the **separate confinement heterostructure** (SCH), optical field confinement and localisation of the recombination region are optimised separately, using a quadruple heterojunction. The two inner junctions limit the active region, while the two outer ones confine the optical field.

**DESIGN EXAMPLE**

We may estimate the effective power gain in a double heterostructure laser as follows. In the central active layer, we may take the power gain to be constant, and equal to $g_{p1}$. However, there may also be additional sources of loss (free carrier absorption, for example), which we might define by a power absorption coefficient $\alpha_{p1}$. In the unpumped outer layers, there will be no gain, only loss defined by $\alpha_{p2}$ and $\alpha_{p3}$ in the left- and right-hand layers, respectively. For the co-ordinates of Figure 12.5-2, the overlap integral in Equation 12.5-4 may then be evaluated as:

$$g_{\text{eff}} = \Gamma_1(g_{p1} - \alpha_{p1}) - \Gamma_2\alpha_{p2} - \Gamma_3\alpha_{p3}$$

12.5-5
where the overlap terms $\Gamma_1$, $\Gamma_2$ and $\Gamma_3$ are defined as:

$$\Gamma_1 = \int_0^h |E|^2 \, dx \quad , \quad \Gamma_2 = \int_0^\infty |E|^2 \, dx \quad , \quad \Gamma_3 = \int_\infty^h |E|^2 \, dx$$

and $\Gamma_1 + \Gamma_2 + \Gamma_3 = 1$. To maximise the effective gain, $\Gamma_1$ should be made much greater than $\Gamma_2$ and $\Gamma_3$, by confining as large a fraction of the optical field as possible inside the central pumped region. For a more complicated channel guide geometry, we might write by analogy:

$$g_{peff} = \Gamma (g_p - \alpha_{pcore}) - (1 - \Gamma)\alpha_{pclad}$$

where $\Gamma$ is the overlap of the guided mode with the active region - known as the confinement factor, and $\alpha_{pcore}$ and $\alpha_{pclad}$ are the losses in the core and cladding, respectively.

### PRACTICAL LASER GEOMETRIES

An optical amplifier may be turned into an oscillator by providing positive feedback. We have already described how this is done in Chapter 4; all that is needed is to add mirrors to the gain block, so that the structure becomes a Fabry-Perot cavity. In the semiconductor laser, the mirrors may be fabricated very simply, by cleavage down crystal planes orthogonal to the junction plane, as shown in Figure 12.5-3. Naturally, their reflectivity is fixed; for a material of refractive index 3.5, surrounded by air, the amplitude reflectivity is $R_1 = R_2 = (3.5 - 1) / (3.5 + 1) \approx 0.56$. This figure is high enough to obtain a suitable level of feedback, while still allowing a good fraction of the light generated to escape from the cavity. Gain is provided by the injection of current in a direction normal to the junction. Because a planar structure provides confinement of the light only in one direction, we would expect light to travel between the mirrors as a sheet beam, growing in amplitude as it propagates.

![Figure 12.5-3](image-url)  
**Figure 12.5-3** Basic geometry of a planar double heterostructure laser.

In practise, the use of a simple layered structure is undesirable. Firstly, both the drive current and the heat generated by non-radiative transitions will be large, since they are proportional to the area of the junction. Secondly, the finite lateral extent of the guide implies that it will actually function as a very broad, multimoded channel guide. The optical field inside the cavity will then consist of a random mixture of modes, which may lead to a highly nonuniform intensity distribution across the beam. Thirdly, broad-area emission will result in very low coupling efficiency into a fibre. Almost always, extra refinements are used to reduce the active volume and introduce transverse confinement, so that the light propagates in a single-moded channel guide.
The primary problem with semiconductor lasers is therefore to combine two-dimensional confinement of the optical wave with localisation of the recombination region. There are many possible solutions. One is to restrict the flow of current laterally, for example by using an SiO\textsubscript{2} isolation layer with a stripe contact window, as shown earlier in Figure 12.4-8. This geometry is known as the oxide stripe laser. Alternatively, a full-area contact might be used, but the semiconductor might be made highly resistive in all except a narrow stripe region, for example by proton bombardment. Both methods restrict the gain region laterally. However, the lateral optical confinement is only moderate, since this is introduced by a local increase in gain rather than by an index variation. Such lasers are therefore often described as gain-guided. Due to the weak confinement, the guide may support several transverse modes, and the relative importance of each mode may be highly dependent on the level of injection. For stable operation on a single transverse mode, stronger lateral confinement must be established using refractive index changes. This requires a more complicated structure.

Figure 12.5-4 shows the simplest example of a suitable geometry, the ridge waveguide laser. In this example, the laser is designed to emit long-wavelength IR radiation (at $\lambda_0 \approx 1.5$ $\mu$m), so the InP/InGaAsP materials system has been used. The starting point is a planar InP/InGaAsP-InP double heterostructure, but the upper InP layer is partially etched into a rib (typically $\approx 5$ $\mu$m wide) to provide lateral confinement of the mode by strip-loading. Note that the current is also automatically confined by the finite width of the rib, although it may start to spread as it passes into the active layer.

![Ridge Waveguide Laser Diagram](image)

Figure 12.5-4 A ridge waveguide laser (after A.W. Nelson et al., Br. Telecom J. 4, 85-103 (1986)).

Figure 12.5-5 shows a more complicated example, the buried heterostructure laser. The laser shown is designed to emit at a shorter wavelength ($\lambda_0 \approx 0.83$ $\mu$m), and is therefore based on the GaAs/GaAlAs materials system. The starting point is a GaAlAs-GaAs-GaAlAs double heterostructure, which is again fashioned into a ridge. This time, however, the ridge is etched completely down to the substrate, and buried on either side by additional high-resistivity GaAlAs. Lateral optical confinement is not lost, however, since the burying layer has a lower refractive index than the active region. Contact is then made to the top of the heterostructure via an oxide stripe window, and the current is constrained to flow through the active region by the high resistivity of the surrounding material.

![Buried Heterostructure Laser Diagram](image)
Figure 12.5-5 A GaAs/GaAlAs buried heterostructure laser

Still better current confinement is possible using more complicated structures. For example, Figure 12.5-6 shows the double-channel planar buried heterostructure laser. This begins life as a planar InP-InGaAsP-InP double heterostructure, which has two deep channels etched through it to the substrate. Four additional layers are then added, by further epitaxial growth. In order, these consist of p-InP, n-InP, p-InP and p''-InGaAs. The important feature is that by utilising a peculiarity of liquid-phase epitaxial growth known as selective area epitaxy, the n-InP layer may be constrained to grow everywhere except over the central ridge. As a result, the structure contains several in-built, reverse-biased p-n junctions, which provide far better current blocking than high-resistivity material. Not only do they force the current to flow into the top of the ridge, but they also prevent it from spreading out of the sides of the active region.

Figure 12.5-6 Cross section of an InP/InGaAsP double-channel planar buried heterostructure laser (after A.W. Nelson et al., Br. Telecom J. 4, 85-103 (1986)).

CONDITIONS FOR LASER OPERATION

We now consider the basic analysis of semiconductor lasers. From Chapter 4, we know that two conditions are required for laser operation. The first is that a longitudinal resonance is satisfied, so that the phase-change accumulated in a round-trip up and down the cavity is a whole number of multiples of $2\pi$. Only then will the feedback be positive, so that the amplitude of the optical field inside the cavity is high. For a single-mode guide with effective index $n_{\text{eff}}$, this condition may be written as:

$$2k_0 n_{\text{eff}} L = 2\nu\pi$$  \hspace{1cm} 12.5-8

However, it should be noted that the effective index is normally dependent on the level of injection, due to the free carrier contribution to the dielectric constant. The resonant wavelengths are spaced by:
\[ \Delta \lambda = \lambda_v - \lambda_{v+1} = \lambda_v^2 / 2L_{\text{eff}} \]

12.5-9

\( \Delta \lambda \) is therefore inversely proportional to the cavity length. The second condition for laser operation is that the round-trip gain must exceed unity, so that:

\[ R_1R_2 \exp(2g_{\text{eff}}L) > 1 \]

12.5-10

The effective power gain at the threshold of lasing is therefore:

\[ g_{\text{peff}} = 2g_{\text{eff}} = 1/L \log_e (1/R_1R_2) \]

12.5-11

Consequently, the shorter the cavity, the higher the threshold power gain. For a cavity of length 250 \( \mu \)m, with end-mirrors of reflectivity \( R_1 = R_2 = 0.56 \), we obtain \( g_{\text{peff}} \approx 4,640 \) m\(^{-1}\).

Note that the gain that must be provided in the active region is even higher, since from Equation 12.5-7 we obtain:

\[ g_{\text{pt}} = \alpha_{\text{core}} + \left( (1 - \Gamma) / \Gamma \right) \alpha_{\text{clad}} + \left( 1/\Gamma L \right) \log_e (1/R_1R_2) \]

12.5-12

In our earlier discussion of LEDs, we described how the distribution of electron and hole energies leads to a finite spectrum for spontaneous emission. For the same reason, gain can be obtained in a semiconductor laser over a finite range known as the gain bandwidth. Normally, the variation of gain with wavelength is written in the form:

\[ g_p(\lambda) = a(n - n_0) - b(\lambda - \lambda_p)^2 \]

12.5-13

where \( a \) and \( b \) are constants, and \( \lambda_p \) is the wavelength for peak gain. With this in mind, we would expect the output of a laser with a relatively long cavity to consist of a series of closely-spaced lines, as shown in Figure 12.5-7a. Each corresponds to a particular longitudinal mode, when Equation 12.5-8 is satisfied for a given value of \( \nu \). The gain bandwidth of (for example) an InGaAsP laser has a half-width of \( \approx 30 \) nm. For the typical parameters of \( \lambda_c = 1.5 \) \( \mu \)m, \( L = 250 \) \( \mu \)m and \( n_{\text{eff}} = 3.2 \) we obtain \( \Delta \lambda \approx 1.4 \) nm, so many lines must lie within the gain bandwidth. We cannot comment on the relative powers in each of the lines yet (except to say that there will be more power in those nearest to the gain peak) since the details of the spectrum depend on the level of injection. However, an extended spectrum is generally undesirable in fibre communications systems.

![Figure 12.5-7](image)

Figure 12.5-7  Representative output spectra for a) and b) a long-cavity laser, with one and two transverse modes, respectively; c) a short-cavity laser supporting a single transverse mode.

Furthermore, we note that a laser with a broader stripe that supports several transverse modes would have a more complicated spectrum. Since each transverse mode will have a different effective index, each will have its own associated spectrum of longitudinal resonances. However, the effective gain of higher-order modes is likely to be lower, since
they are more weakly confined than the fundamental mode. Consequently, the spectrum of a two-moded laser might be as shown in Figure 12.5-7b, where the smaller lines correspond to the second-order mode. Since the lowest-order mode is generally the most useful, single-transverse-mode operation is highly desirable. In this case, and with a much shorter cavity, we might expect only one longitudinal resonance to fall inside the gain bandwidth, so that the output is effectively single-line (Figure 12.5-7c). However, this requires an extremely small value of $L (\approx 50 \mu m)$.

**QUANTUM WELL AND MULTIQUANTUM WELL LASERS**

Band-gap engineering offers further scope for the control of laser properties. For example, Figure 12.5-8a shows the equilibrium band diagram of an undoped GaAlAs-GaAs-GaAlAs double heterostructure. Due to the difference in bandgap between the materials, there are potential wells in both the conduction band and the valence band. In conventional lasers, these are exploited purely for carrier confinement. However, as the width $h$ of the central layer is reduced, new effects start to occur. In Chapter 11, we analysed the confinement of an otherwise free electron in a one-dimensional well, and showed that the result was a quantization of the electron energy. For an infinite well of width $h$, with a minimum potential energy of $V = 0$, the allowed energies were:

$$E = \nu^2 (\pi^2 \hbar^2 / 2m h^2) + \hbar^2 \beta^2 / 2m$$  \hspace{1cm} (where $\nu = 1, 2 \ldots$)

Of these two contributions, the first arises from the quantization effect of the well, while the second is a kinetic energy term due to motion in the two remaining allowed directions. By analogy, we might expect the energies of conduction electrons in the well of Figure 12.5-8 to be of the form:

$$E_{c1}' = E_{c1} + \nu^2 (\pi^2 \hbar^2 / 2m_e^* h^2) + \hbar^2 \beta^2 / 2m_e^*$$  \hspace{1cm} (where $\nu = 1, 2 \ldots$)

while the corresponding hole energies might be:

$$E_{v1}' = E_{v1} - \mu^2 (\pi^2 \hbar^2 / 2m_h^* h^2) - \hbar^2 \beta^2 / 2m_h^*$$  \hspace{1cm} (where $\mu = 1, 2 \ldots$)

where $E_{c1}$ and $E_{v1}$ are the extremities of the conduction and valence bands in the GaAs layer, and $m_e^*$ and $m_h^*$ are the electron and hole effective masses. As a result, the minimum difference between conduction and valence levels is no longer $E_{g1} = E_{c1} - E_{v1}$, but:

$$E_{g1}' = E_{g1} + (\pi^2 \hbar^2 / 2h^2) (1/m_e^* + 1/m_h^*)$$  \hspace{1cm} 12.5-17

This energy difference corresponds to a transition between stationary electrons and holes in the levels defined by $\nu = 1, \mu = 1$. This is equivalent to the creation of an artificial semiconductor, with a bandgap different from that of either of the materials forming the well. Note that $E_{g1}'$ is controlled essentially by the value of $h$. Consequently, the emission wavelength of a laser constructed in this way should be similarly variable.

In fact, the wavelength may be tuned over almost the entire range set by the bandgaps of the two constituent materials, GaAs and Ga$_{1-x}$Al$_x$As. The tuning range is therefore limited mainly by the transition of GaAlAs to an indirect bandgap material at $x \approx 0.35$. Using quantum wells made from InP and InGaAs (which are both direct-gap), the wavelength may be tuned over the staggering range 1.07 - 1.55 $\mu m$, by varying the well width between 10 and 110 $\AA$. 

R.R.A.Syms and J.R.Cozens  Optical Guided Waves and Devices  36
The quantum size effect offers other advantages besides wavelength tunability. For example, it is possible to repeat the calculation to find the density of states \( S'(E) \) in the conduction band inside a single well. If this is done, we obtain:

\[
S'(E) = \frac{\nu}{\pi \hbar} \text{ for } E_{\nu} \leq E \leq E_{\nu+1}
\]

where \( E_{\nu} = E_{c1} + \frac{\nu^2(\pi^2 \hbar r^2/2m_e^*)}{\hbar^2} \). Consequently, \( S'(E) \) is discontinuous, as shown in Figure 12.5-8b. Although this is very different from our previous parabolic result for \( S(E) \) - which is also the density of states for an infinite well width - the most important point is that \( S'(E) \) intersects \( S(E) \) at each energy \( E_{\nu} \), but otherwise lies inside it. We may therefore deduce that there must be a smaller total density of states in the conduction band in a quantum well (actually, this is obvious, since the effect of quantization is to disbar many possible states). Now, remember that we must invert the electron population to obtain optical gain. With fewer states to invert, the electron concentration required for transparency must be lower, so a quantum well laser must have a lower threshold than a conventional one. However, its peak gain must be reduced, since the conduction band will be filled more easily.

If the well is large enough, a double heterostructure laser can be made using a single quantum well as the active layer. However, as the well size is reduced, it becomes impossible to provide suitable optical confinement - the resulting waveguide is too narrow, so that the optical field extends considerably outside the gain region. It is then normal to construct the active layer as a multilayer stack, consisting (for example) of many alternating layers of InP and InGaAs, and to provide the optical confinement separately. In this case, the structure is known as a multi-quantum well (MQW) laser.

### 12.6 THE SEMICONDUCTOR LASER II: STEADY-STATE ANALYSIS

Assuming that the laser operates on a single transverse mode, and that the output is concentrated in a single line, we may model the operation of a laser using relatively minor
modifications of the two-state, lumped-element rate equations derived earlier for LEDs. Although the optical field inside the cavity will actually consist of a standing-wave pattern, we shall ignore this and simply assume an average value for the photon density $\phi$. This will be high at the resonant wavelength (which we take to be $\lambda_g$), and zero otherwise. For simplicity, we shall ignore optical loss, and assume a confinement factor of $\Gamma = 1$. We may then take the photon rate equation for a cavity of volume $v$ in the form:

$$\frac{d\phi}{dt} = \frac{\beta n}{\tau_{rr}} + G\phi(n - n_0) - \frac{\phi}{\tau_p}$$  \hspace{1cm} 12.6-1

Here we have included terms describing spontaneous emission ($\frac{\beta n}{\tau_{rr}}$), stimulated emission and absorption ($G\phi(n - n_0)$), and loss of photons from the cavity ($\frac{\phi}{\tau_p}$). The factor $\beta$ has been introduced into the first of these to describe the fraction of the isotropically generated spontaneous emission which is actually coupled into the guided mode; typically, this is very small ($\approx 10^{-3}$), due to the low NA of a single-mode guide. Similarly, the electron rate equation may be taken as:

$$\frac{dn}{dt} = \frac{I}{e\tau_e} - \frac{n}{\tau_e} - G\phi(n - n_0)$$  \hspace{1cm} 12.6-2

Since Equations 12.6-1 and 12.6-2 are non-linear, we will solve them by means of approximations that are valid in the two important regimes of operation. In the steady state, the left-hand side of each of these equations is zero, so we may then combine them to eliminate $G\phi(n - n_0)$. This yields:

$$\frac{\phi}{\tau_p} = \frac{\beta n}{\tau_{rr}} + \left(\frac{I}{e\tau_e} - \frac{n}{\tau_e}\right)$$  \hspace{1cm} 12.6-3

Now, the light flux out of the cavity is $\Phi = v\frac{\phi}{\tau_p}$ as before. As we have ignored optical loss, this flux must all emerge from the cavity end-mirrors. If each photon carries energy $hc/\lambda_g$, the optical power output is then $P = \frac{hc\Phi}{\lambda_g}$, or:

$$P = (hc/\lambda_g)\left(\frac{\beta n}{\tau_{rr}} + \left(\frac{I}{e\tau_e} - \frac{n}{\tau_e}\right)\right)$$  \hspace{1cm} 12.6-4

Note that Equation 12.6-4 represents the total output, so that $P/2$ emerges from each of the mirrors.

**OPERATION BELOW THRESHOLD**

For a low photon density, spontaneous emission is the dominant process, so we may set the term $\left(\frac{I}{e\tau_e} - \frac{n}{\tau_e}\right)$ (which describes stimulated emission) equal to zero. In this case, we have $n = I\tau_e/e$, so that:

$$P = \frac{\eta hci}{e\lambda_g}$$  \hspace{1cm} 12.6-5

where $\eta = \frac{\beta \tau_e}{\tau_{rr}}$. For low currents, the output is therefore proportional to current, as in the left-hand part of Figure 12.6-1a. The laser acts as an LED in this regime, with a reduced internal efficiency.

**THE THRESHOLD CONDITION**

Now, the existence of spontaneously emitted light in the laser cavity provides a source of optical noise, from which laser oscillation may start. The noise will travel up and down the cavity, and, if there is net gain, it will be amplified as it propagates. Very quickly, the photon density will become high, so that stimulated emission becomes the dominant process. In the
steady state, assuming that spontaneously emitted light is now negligible, we may approximate Equation 12.6-1 by:

$$G \phi (n - n_0) - \phi / \tau_p = 0$$  
12.6-6

We may then obtain the electron density during lasing as:

$$n = n_0 + 1/G \tau_p$$  
12.6-7

As a result, the electron concentration is entirely independent of $\phi$. Equation 12.6-7 must also represent the electron density at threshold, $n_t$, and therefore describes the **threshold condition**.

![Figure 12.6-1](image)

**Figure 12.6-1** a) Optical power - current characteristic of a semiconductor laser, b) dependence of threshold current on temperature, and c) kink in characteristic due to mode-hopping.

**DESIGN EXAMPLE**

We may the result above to estimate the photon lifetime. Since we have already established the threshold condition, we may find $n_t$ by an alternative route. Setting $\Gamma \approx 1$ and $\alpha_{\text{core}} = \alpha_{\text{clad}} \approx 0$ in Equation 12.5-12, the threshold gain is $g_{pt} = 1/L \log(1/R_1 R_2)$. Since $g_{pt} = G(n_t - n_0)/v_g$, we may find the threshold electron density as:

$$n_t = n_0 + (v_g / GL) \log_e (1/R_1 R_2)$$  
12.6-8

Comparing Equations 12.6-7 and 12.6-8 we may then obtain the photon lifetime $\tau_p$ as:

$$\tau_p = L / \{ v_g \log_e (1/R_1 R_2) \}$$  
12.6-9

The photon lifetime is therefore independent of the level of injection, and depends only on the cavity length and mirror reflectivity. For a semiconductor of refractive index $n$, we may put $v_g = c/n$. Assuming that $n = 3.5$, the end-mirror reflectivity is $R_1 = R_2 = 0.56$. If the cavity length is 250 $\mu$m, the photon lifetime is $\tau_p = 250 \times 10^{-6} \times 3.5 / \{ 3 \times 10^8 \times \log_e (1/0.562) \} = 2.5 \times 10^{-12}$ sec, or 2.5 psec.
OPERATION ABOVE THRESHOLD

We now consider operation above threshold. Assuming negligible spontaneous emission, Equation 12.6-4 approximates to:

\[ P = \left( \frac{hc\nu}{\lambda g} \right) (I/ev - n/\tau_e) \]  

Since we have already shown that the electron concentration is clamped above threshold, \( n/\tau_e \) must be constant. We might define it to be equal to \( I_t/ev \), where \( I_t \) is a new parameter known as the **threshold current**. It is therefore proportional to the active volume \( v \) and the threshold electron density \( n_t \), and inversely proportional to \( \tau_e \). Its exact value is highly dependent on the laser type; typical values are the range 10 - 100 mA. Using this definition, Equation 12.6-10 may be written as:

\[ P = \frac{hc(I - I_t)}{e\lambda g} \]  

The output of a semiconductor laser is therefore a discontinuous function of the drive current, being proportional to \( I - I_t \) in the lasing regime as shown in the right-hand part of Figure 12.6-1a. The peak output power is normally limited either by mirror damage (caused by the high optical intensity passing through the end-facets) or by thermal breakdown (caused by the high current density passing through the junction). Output powers in the range 1 - 10 mW are routinely available at all the major wavelengths (near 0.85 \( \mu \)m for GaAs/GaAlAs lasers, and 1.3 and 1.5 \( \mu \)m for InP/InGaAsP).

Generally, \( I_t \) rises with temperature, because the spread of energies in the conduction and valence band increases, reducing the density of carriers that can generate a photon of a given wavelength. The recombination time \( \tau_e \) is also decreased. Consequently, the power-current characteristic of a laser at three values of \( T \) might be as shown in Figure 12.6-1b. In fact, operation at a greatly reduced temperature was often used to lower the threshold current to an acceptable level in early lasers, which had poor confinement factors. \( \tau_e \) also decreases as the active region ages, so \( I_t \) rises with time.

MULTILINE OPERATION

The two-state model we have used cannot describe multiline operation. However, we may obtain a feel for the likely behaviour, as follows. If the spontaneous emission term is retained in Equation 12.6-1, a steady-state solution for the photon density may be found as:

\[ \phi = \frac{\beta n/\tau_e}{\left\{ 1/\tau_p - G(n - n_0) \right\}} \]  

This shows that the light in the cavity may be considered to arise from spontaneously-emitted noise, 'amplified' by the factor \( F = 1 / \left\{ 1/\tau_p - G(n - n_0) \right\} \). Since the denominator of this expression will be very small, \( F \) must be very large. In a more complicated model, including several longitudinal modes, we might assign a different gain to each, following Equation 12.5-13. We might then write:

\[ F_v = 1 / \left\{ 1/\tau_p - [G(n - n_0) + b' (\lambda - \lambda_p)^2] \right\} \]

for the amplification factor for mode \( v \), where \( \lambda_v \) is the wavelength of the mode and \( b' = b \nu \). The relative sizes of each \( F_v \) then depends on the closeness of each term \( G(n - n_0) + b' (\lambda - \lambda_p)^2 \) to \( 1/\tau_p \). For low currents (i.e., when \( n - n_0 \) is small), all the amplification factors are small, and roughly equal, so that multiline operation occurs. As \( n - n_0 \) increases, every value
of $F_\nu$ will increase, but the line whose wavelength is closest to $\lambda_p$ will have a proportionately larger amplification factor.

We can then guess that the mode spectrum will tend to alter as shown in Figure 12.6-2, where the power is gradually concentrated into one dominant line as the current rises. This is highly desirable, since it reduces the effect of dispersion when a laser is used as a source for optical communications. Note that an increase in drive current may also make it possible for the transverse lasing mode to change, for example, from the lowest-order mode at low current to the second-lowest at higher current. This behaviour is known as **mode-hopping**. It may be observed as a kink on the laser output characteristic, as shown in Figure 12.6-1c, and is generally undesirable.

![Figure 12.6-2 Typical variation of the mode spectrum of a semiconductor laser.](image)

**OPTICAL CHARACTERISTICS**

We now consider briefly a number of other characteristics of a semiconductor laser. Firstly, in order to obtain good overlap between the optical field and the gain distribution, the height $h$ of the active region of a double heterostructure laser supporting a single transverse mode is normally very small, of the order of 0.1 - 0.5 $\mu$m, while the corresponding width $w$ is typically $\approx$ 1 - 5 $\mu$m. The core of the laser waveguide is therefore rectangular, with an aspect ratio of $\approx$ 10 : 1. The mode supported by the guide is similarly distorted, being highly compressed perpendicular to the plane of the layering, as shown in the near-field pattern of Figure 12.6-3.

To a good approximation, each facet behaves as a point emitter, so that the far-field radiation pattern is roughly that of a paraxial spherical wave. Now, we saw in Chapter 4 that the spread of a bounded beam is inversely dependent on its width. Due to the shape of the near-field pattern, we would expect the angular spread to be greatest in the plane perpendicular to the junction (also as shown in Figure 12.6-3). A typical value for the full-angle, half-maximum spread in intensity in this plane is $\theta_\perp \approx$ 30 - 50°; the corresponding spread $\theta_\parallel$ parallel to the junction is much smaller, in the range 10 - 15°, but both figures depend strongly on the construction of the laser. The output may be collimated using a spherical lens, simply by placing the emitting area at the focal point. The resulting beam will roughly parallel, but (due to the differences in beam divergence) its cross-section will be elliptical. Approximate circular symmetry may be restored by passing the beam through an
anamorphic expansion telescope that magnifies in one direction only. However, the presence of gain in the active region causes a slight astigmatism of the modal phase-front. If accurate collimation is required, this must be corrected by passing the beam through an additional cylindrical lens.

Figure 12.6-3 Near- and far-field radiation pattern from a buried heterostructure laser.

Because of the increased directivity and coherence of the light generated by stimulated emission, much higher coupling efficiency to a low-NA single-mode fibre can be obtained with a laser than with an LED. The coupling efficiency may be found by evaluating the overlap between the modal fields of the laser and fibre, much as was done in Chapter 8. Since the laser mode is typically rather small, an improvement in efficiency may be obtained by using a lens. This magnifies the laser mode slightly, so that it makes a better match to the fibre mode.

Finally, we note that a nominally single-moded guide may support two orthogonal polarization modes. However, due to differences in end-mirror reflectivity, the mode polarized parallel to the junction plane experiences greater round-trip gain, and will lase preferentially. Consequently, the stimulated emission is linearly polarized in this direction. However, spontaneous emission will add unpolarized light to the output, which will also have a component perpendicular to the junction plane. This lowers the degree of polarization, i.e. the ratio of orthogonally polarized components. This is limited, and will vary with the output power. A typical value for a laser driven well past threshold is greater than 50 : 1.

12.7 THE SEMICONDUCTOR LASER III: MODULATION

We now consider the behaviour of a semiconductor laser when it is modulated by a sinusoidal time-varying drive current, to demonstrate the improvement in its high-frequency response over that of an LED. Now, it can be shown that the time taken to build up the photon density in the cavity from zero is relatively long, due to the slow response of the
device at low current (when it operates as an LED). For high-speed operation, a laser is therefore normally biased past threshold, using a DC bias current. We shall therefore assume that the drive current is of the form $I = I' + I'' \exp(j\omega t)$, where $I'$ is the DC bias, and $I''$ and $\omega$ are the amplitude and angular frequency of the superimposed AC modulation, respectively. Note that the bias is chosen so that $I' > I_t$ and $I' \gg I''$. We shall also define the other quantities of interest in a similar way, as the sum of a DC bias and an AC modulation term. We therefore put $\phi = \phi' + \phi'' \exp(j\omega t)$, $n = n' + n'' \exp(j\omega t)$ and $P = P' + P'' \exp(j\omega t)$.

We now substitute these expressions into Equations 12.6-1 and 12.6-2. To simplify the results, we perform the following operations: (1) we neglect spontaneous emission as small by comparison with stimulated emission, and (2) we eliminate as many terms as possible, using the relations between the DC terms $\phi'$, $n'$, $I'$ and $P'$ derived in the previous section. The photon rate equation then reduces to:

$$j\omega \phi'' = G \phi'n''$$  \hspace{1cm} 12.7-1

Similarly, the electron rate equation becomes:

$$j\omega n'' = I''/ev - n''/\tau_e - \phi''/\tau_p - G \phi'n''$$  \hspace{1cm} 12.7-2

Equations 12.7-1 and 12.7-2 may then be combined, to yield the small-signal optical modulation amplitude $P''$ in terms of the corresponding current modulation term $I''$, as:

$$P''(\omega) = (hcI''/e\lambda_g) (G\phi'/\tau_p) / \{-\omega^2 + j\omega[1/\tau_e + G\phi'] + G\phi'/\tau_p\}$$  \hspace{1cm} 12.7-3

Because this expression is rather complicated, we now introduce a number of new parameters. Firstly, we define the modulation amplitude at zero frequency as:

$$P''(0) = hcI''/e\lambda_g$$  \hspace{1cm} 12.7-4

Using this, we define the complex modulation ratio as $m(\omega) = P''(\omega) / P''(0)$. Finally, we introduce a resonant frequency $\omega_r$ and a damping factor $\zeta$, defined as:

$$\omega_r = \sqrt{(G\phi'/\tau_p)} \quad \text{and} \quad \zeta = \{1/\tau_e + \tau_p \omega_r^2\} / 2\omega_r$$  \hspace{1cm} 12.7-5

With this new notation, the modulation ratio may be written as:

$$m(\omega) = \omega_r^2 / \{(\omega_r^2 - \omega^2) + 2j\zeta\omega\omega_r\}$$  \hspace{1cm} 12.7-6

The small-signal frequency response of a laser is therefore that of a damped simple harmonic oscillator. This result might well have been anticipated; since the rate equations amount to two coupled first-order differential equations, the transient response is effectively described by two uncoupled second-order equations, each in the form of the governing equation of an oscillator. Since $m(\omega)$ is complex, the optical modulation must lag behind the current once again. The modulus of $m(\omega)$ is of most interest, since this describes the depth of the optical modulation. This is given by:

$$|m(\omega)| = \omega_r^2 / \sqrt{((\omega_r^2 - \omega^2)^2 + (2\zeta\omega\omega_r)^2)}$$  \hspace{1cm} 12.7-7

Figure 12.7-1 shows a logarithmic plot of $|m(\omega)|$ versus $\omega$, for typical values of $\omega_r$ and $\zeta$. At low angular frequencies, $|m(\omega)|$ is unity. As the frequency rises, there is a peak in the response, due to a resonance between the electron and photon populations. This is known as
a **relaxation resonance**; for low damping, it occurs at a frequency close to $\omega_r$. Finally, at high frequencies, the response falls off roughly as $|m(\omega)| = \omega_r^2/\omega^2$. The bandwidth is therefore set almost entirely by the resonant frequency $\omega_r$. Using some simple manipulations, it is possible to define $\omega_r$ alternatively as:

$$\omega_r = \sqrt{G \lambda g P' / h c}$$  \hspace{1cm} 12.7-8

The resonant frequency is therefore variable, rising as the square root of the DC power level. In order to obtain the greatest modulation bandwidth, the laser should therefore be driven as hard as possible.

**Figure 12.7-1** Small-signal frequency response of a semiconductor laser.

**DESIGN EXAMPLE**

We may estimate $\omega_r$ and $\zeta$ for a GaAs laser as follows. Assuming a cavity length of 250 $\mu$m, and an active area cross-section of 0.1 $\mu$m x 3 $\mu$m, a gain constant of $10^{-12}$ m$^3$/s, and an output power level of 3 mW at $\lambda_g = 820$ nm, we obtain $\omega_r = \sqrt{10^{-12} \times 0.82 \times 10^6 \times 3 \times 10^3 / (6.62 \times 10^{-34} \times 3 \times 10^8 \times 250 \times 0.1 \times 3 \times 10^{-18})} = 1.29 \times 10^{10}$ rad/sec, corresponding to a bandwidth of $\approx$ 2 GHz.

This is entirely typical; the modulation bandwidth of a laser is typically $\approx$ 100x greater than that of an LED, making the laser the preferred source for high bit-rate communications systems. For modulation at still higher speeds, an external modulator must be used. Now, we have previously estimated the photon lifetime for a 250 $\mu$m - long GaAs laser as $\tau_p \approx 2.5$ psec. Assuming a recombination time of $\tau_e \approx 1$ nsec, the damping factor is $\zeta = \{10^9 + 2.5 \times 10^{-12} \times (1.29 \times 10^{10})^2 / (2 \times 1.29 \times 10^{10}) \approx 0.05$. This is also typical; the relaxation oscillations are underdamped at normal powers, leading to the sharp peak in the response near $\omega_r$ in Figure 12.7-1.

**12.8 DBR AND DFB LASERS**

Although it is blessed with the important advantage of simplicity, the Fabry-Perot semiconductor laser suffers from several undesirable characteristics. The first is that its output is often multi-line, a major drawback for optical communication systems. In this Section, we shall consider a number of solutions to the problem. The two most successful
approaches involve reflection gratings, of the type described in Chapter 10; these are known as the distributed Bragg reflector (or DBR) and the distributed feedback (or DFB) laser, respectively. Figure 12.8-1 is a comparison between the two structures.

Figure 12.8-1 a) DBR and b) DFB lasers.

In the DBR laser, either or both of the cleaved end-mirrors are replaced by a corrugated reflection grating. As a result, the DBR laser has several sections: a central active region, which provides the gain, and passive grating reflectors, which provide feedback. Now, we would expect high reflectivity from the grating only for longitudinal modes whose wavelengths approximately satisfy the Bragg condition, \(2n_{\text{eff}} \nu \Lambda = \lambda\). When this is not satisfied, the reflectivity is quickly reduced to zero, drastically reducing the photon lifetime in the process. We have already shown that the relative importance of different longitudinal modes is governed by amplification factors \(F_{\nu}\). Including a mode-dependent photon lifetime \(\tau_{\nu}\), these might be written as \(F_{\nu} = 1 / \{1/\tau_{\nu} - [G(n - n_0) + b'(\lambda_{\nu} - \lambda_p)^2]\}\). Since they depend inversely on the difference between two comparable quantities, even a small increase in \(1/\tau_{\nu}\) can have a significant effect. Consequently, the introduction of grating feedback causes a vastly greater purification of the spectrum than the grating filter response might suggest. In fact, the mode selectivity is governed almost entirely by the curvature of the response near the Bragg wavelength, and the output of a DBR laser may consist only of a single line.

Now, the wavelength of each longitudinal mode is governed by the optical path length between the reflectors. Since the effective index in the active region will depend on the level of pumping, the positions of the longitudinal modes relative to the peak in grating reflectivity must alter with current. To obtain optimum reflectivity for one desired mode, independent of variations in drive current, a slightly more complicated structure is required. For example, Figure 12.8-2 shows a three-section DBR laser.

Figure 12.8-2 A three-section, continuously tunable MQW-DBR laser (after T.L.Koch et al., Elect. Lett. 24, 1431-1433 (1988)).
This has a multi-quantum well active section; feedback is provided by a cleaved mirror at the left-hand end, and a grating reflector at the right. However, in this case the Bragg wavelength may be altered by a tuning current, which varies the effective index beneath the grating through minority carrier injection. A central passive phase-shifting section (operating by the same mechanism) then allows the position of one longitudinal mode to be matched to the peak in grating reflectivity. Through an interplay between the three currents, tunable single-line operation is therefore possible.

DFB lasers (Figure 12.8-1b) achieve a similar end result, but operate by an entirely different principle. In this case, the grating and the active region coincide, and is no longer possible to consider the feedback as being provided by reflectors at either end of a cavity; instead, it must be distributed throughout the active region. To show that this may give rise to laser operation, we need a simple model. Below threshold, linear coupled mode theory is suitable. All that is required is the insertion of an effective gain parameter $g$ into the coupled equations previously derived for a reflection grating. If this modification is made to Equations 10.10-14, for example, we obtain:

$dA_F/dz - gA_F + j\kappa A_B = 0$

$dA_B/dz + \left( g - j2\Delta\beta \right)A_B - j\kappa A_F = 0$

12.8-1

We can check that the extra terms are realistic, by examining the solutions for the forward and backward mode amplitudes $A_F$ and $A_B$ when the coupling coefficient is zero. They grow exponentially in the ±z directions, as $\exp(\pm g z)$, which seems reasonable. When $\kappa$ is non-zero, the equations may still be solved analytically. However, the solutions are hard to interpret, as they contain many complex quantities. It is therefore simpler to discuss a typical numerical solution, for the boundary conditions of unity input to one end of the device (i.e., for $A_F = 1$ on $z = 0$ and $A_B = 0$ on $z = L$). For example, Figure 12.8-3 shows logarithmic plots of reflectivity $R = A_B^* A_B$ on $z = 0$ versus detuning $\Delta\beta L$, for a coupling length of $\kappa L = 0.4$ and different values of normalised gain $gL$.

For $gL = 0$, the response is that of a passive grating, reaching a peak when the Bragg condition is satisfied and falling gradually to zero with a standard filter response as $|\Delta\beta L|$ increases. As $gL$ rises, however, the shape of the response alters dramatically. The reflectivity improves overall, which is only to be expected. More interestingly, sharp peaks start to appear in the curves. These are displaced symmetrically on either side of the Bragg condition, and their characteristics are broadly analogous to those of the longitudinal resonances occurring in a Fabry-Perot cavity. They arise when the contributions that are scattered by the grating and then amplified by the gain mechanism combine with a suitable relative phase to ensure high net reflectivity. For $gL = 3$, the first-order resonances are extremely pronounced, reaching a peak of $\approx 13,000$, a staggering figure. In fact, for slightly higher gain, $R$ may be infinite, as may the transmissivity $T$. At this point, the device may be said to be lasing, because an output may appear from both ports for a zero input.

From Figure 12.8-3, it is clear that the largest peaks occur at the first-order resonances, so these will have the lowest threshold. The spectrum of a DFB laser will therefore consist of two lines, each of slightly different wavelength. These are automatically aligned with respect to the Bragg wavelength, so no separate tuning mechanism is required to optimise their thresholds. However, the wavelength of each line will alter with drive current, due to the change in effective index, so a DFB laser is tunable over a moderate range. Although this is a considerable improvement over a Fabry-Perot laser, further advances are possible. All that is required is to generate some form of asymmetry in the grating reflectivity. This may be
done, by inserting a step in the phase of the grating at a suitable point along the cavity, a simple process for gratings fabricated by electron-beam writing. Since the required shift is one quarter of a period, such devices are known as \( \Lambda/4 \)-shifted DFB lasers. The net result is that the threshold gains of the two first-order resonances are displaced from one another, so that one longitudinal mode lases preferentially and the output is single-line.

Figure 12.8-3 Reflectivity as a function of detuning \( \Delta \beta L \), for a reflection grating with gain.

Figure 12.8-4a shows a schematic of a typical long-wavelength ridge waveguide DFB laser. This is very similar in construction to the Fabry-Perot laser shown in Figure 12.5-4, but an extra layer of material has been inserted above the active layer. The grating is fabricated by etching a pattern of electron-beam-defined grooves into this layer, using a preferential wet chemical etch that forms self-stopping V-shaped grooves with a precise depth (giving a well-controlled coupling coefficient). Note that etching of the active layer itself is inadvisable, because most etch processes create surface damage sites that degrade the recombination efficiency, increasing the threshold.

Because it is possible to fabricate complicated grating profiles, the perturbation of the guide need not be sinusoidally varying (although it must be periodic in \( \Lambda \)). In general, the perturbation may be written as a Fourier series, with components of period \( \Lambda \), \( \Lambda/2 \), \( \Lambda/3 \) etc. These have associated grating vectors \( K_1 = 2\pi/\Lambda \), \( K_2 = 2K_1 \), and \( K_3 = 3K_1 \), respectively.

We have already seen that the fundamental component of a grating may be used to provide Bragg reflection at a wavelength \( \lambda_0 = 2n_{\text{eff}}\Lambda \) (say, a near-infrared wavelength). Clearly, the second harmonic component will operate in a similar way at half this wavelength (somewhere in the visible spectrum). This is not particularly useful in itself. However, if the grating period is now doubled, the second harmonic may operate at the original IR wavelength. This is highly advantageous, because it allows the resolution of the process used to define the grating pattern to be relaxed. **Second-order gratings** are therefore common.
Second-order gratings do have one additional characteristic, which is best illustrated using K-vector diagrams. Figure 12.8-5a shows the diagram for Bragg reflection using the second-harmonic grating vector $K_2$, which generates a backward-going diffraction order from a forward-travelling input. This is as expected. However, the fundamental grating vector $K_1$ may also cause diffraction, as shown in Figure 12.8-5b. Since it is half the length of $K_1$, the diffraction orders it generates propagate vertically upwards into the guide cover layer and downwards into the substrate. The fundamental grating therefore acts as a grating coupler, dumping energy out of the cavity. Consequently, it acts as an additional source of loss, increasing the lasing threshold. Although this is normally undesirable, many manufacturers are prepared to pay the price for the reason given above. The same output coupling mechanism is also important in surface-emitting lasers, which we shall describe shortly.
All forms of grating feedback share one further advantage over Fabry-Perot lasers. Because they do not require cleaved mirrors, the substrate is not limited to the length of the laser. It is therefore simple to combine a DFB or DBR laser with other guided wave devices. Particular attention is now being paid to the integration of arrays of lasers with beam-combining components, so that their outputs may be combined in a single channel. For example, an array of DFB lasers (each arranged to emit a slightly different wavelength) might be integrated with a Y-junction tree. The output of the tree will then consist of a number of wavelengths, each of which can be modulated separately. Such a source might form the basis of a wavelength-division multiplexed (WDM) communications system.

12.9 ARRAY LASERS

We shall now consider a different question, namely how to increase the power of a semiconductor laser. Now, the main limitations of stripe-geometry Fabry-Perot lasers are their restricted active volume and the requirement that the optical output must pass through end-facets of small cross-sectional area - indeed, catastrophic mirror failure was a major problem in early devices. One obvious way to improve total output is to use several lasers, placed side-by-side on a chip, so that the active volume and output-coupling window are increased simultaneously. The problem with this arrangement is that each of the lasers will oscillate independently. As a result, it is impossible to combine their outputs into a single, high-power beam, because they must sum together with random phases.

To obtain a useful increase in power, the laser must be forced to emit light of a fixed spatial pattern. Preferably, the outputs of all emitters should be of equal amplitude and phase, and a variety of possible structures have been investigated in pursuit of this goal. For example, the earliest geometry - the coupled waveguide array laser, shown in Figure 12.9-1 - was based on a combination of a directional coupler and a Fabry-Perot cavity. A number of parallel laser stripes were arranged close together on a chip, so that the modal fields in adjacent stripes overlapped. It was anticipated that co-directional coupling would then introduce phase locking. Unfortunately, this was not the case, as we show below; however, the experience allowed better structures to be devised very quickly.

![Figure 12.9-1  A seven-element co-directionally coupled ridge waveguide array laser.](image)

To analyse the coupled waveguide array laser, we need a simple model. Once again, this may be provided by coupled mode theory. Ignoring the feedback from the mirrors for the time being, we may start by considering the related problem of wave propagation in an N-element coupled waveguide array. We shall assume there is equal gain in each stripe (defined by an effective gain g) and nearest-neighbour coupling between stripes (defined by a coupling coefficient κ). In this case, the coupled mode equations governing the change...
in amplitude of the mode in each stripe with distance along the array may be obtained by a
c simple modification of our earlier two-guide equations, as:

\[ \frac{dA_1}{dz} - gA_1 + j\kappa A_2 = 0 \]
\[ \frac{dA_n}{dz} - gA_n + j\kappa \{ A_{n+1} + A_{n-1} \} = 0 \quad \text{for } n = 2, 3 \ldots N - 1 \]
\[ \frac{dA_N}{dz} - gA_N + j\kappa A_{N-1} = 0 \]

Equation 12.9-1 shows that the two outermost guides are coupled to a single neighbour,
while all the other guides are coupled to two. Now, in Chapter 10, we solved two similar
equations for fixed boundary conditions. This time, we shall follow a different course and
look for eigenmode solutions. These describe patterns of mode amplitude that may
propagate along the array with a change in overall amplitude or phase, but not of shape.
Each such pattern may therefore be written in the form:

\[ A_n = a_n \exp(-j\gamma z) \]

Equation 12.9-2 describes a form of travelling wave, moving across
the array and changing in phase by discrete amounts between guides. There is a strong
analogy between this solution and the Bloch waves of Chapter 11; indeed, the behaviour of
light in a periodic array of guides is very similar to that of electrons in an array of potential
wells. Here we have approached the mathematics slightly differently, but we should not be
surprised if similar conclusions are reached. Substituting this solution into Equation 12.9-1
and eliminating all common terms gives the following relation between \( \gamma \) and \( \phi \):

\[ \gamma = jg + 2\kappa \cos(\phi) \]

Ignoring the gain term, Equation 12.9-4 shows that the eigenvalue \( \gamma \) may only take a
restricted range of values (so that \(-2\kappa \leq \gamma \leq 2\kappa\)), varying cosinusoidally with \( \phi \) over the
range. This should be compared with the \( E - k \) relation predicted by the Krönig-Penney
model in Chapter 11.

We now reintroduce the finite lateral extent of the array, by assuming that the upper and
lower equations in 12.9-1 will effectively be satisfied if we assume coupling to two
fictitious guides (at \( n = 0 \) and \( n = N + 1 \)), which support modes of zero amplitude. This
may be done by taking \( a_n = A' \sin(\phi_n) \), i.e. by assuming a standing-wave solution. Applying
the boundary conditions \( a_0 = a_{N+1} = 0 \), we then find that \( \phi \) is restricted to the following set of
discrete values:

\[ \phi_v = \nu\pi / (N+1) \]

where \( \nu = 1, 2 \ldots N \) is the mode number. The complete solution to Equation 12.9-2 is
therefore defined by mode amplitudes \( a_{\nu n} \) and propagation terms \( \gamma_v \), given by:
\[ a_{vn} = A \sin \left[ vn \pi / (N+1) \right] \quad \gamma_v = jg + 2\kappa \cos \left[ \nu \pi / (N+1) \right] \]

12.9-6

Now, the terms \( a_{vn} \) are the amplitudes of the modes in each guide. If each individual mode is defined by a transverse field \( E_i(x, y) \), the total field pattern of the \( \nu \)th eigenmode must therefore be:

\[ E_{\nu}(x, y) = \sum_{i=1}^{N} a_{\nu i} E_i(x, y) \]

12.9-7

Figure 12.9-2a shows a one-dimensional plot of the lowest-order transverse mode (\( \nu = 1 \)) for a seven-guide array. Within each guide, the pattern is like that of the individual guide mode, but there is an overall sinusoidal envelope defined by Equation 12.9-6. Similarly, Figure 12.9-2b shows the second-order mode (\( \nu = 2 \)). This is similar, but the spatial frequency of the envelope is now doubled. The trend continues as \( \nu \) increases, so that the envelope of the highest-order mode has the greatest spatial frequency, with adjacent mode amplitudes of opposite sign.

![Near-field patterns of a) the lowest-order and b) the second-lowest order transverse mode of a seven-element coupled waveguide array.](image)

We now return to the provision of feedback by the end-facets. If their reflectivity is uniform across the array, each eigenmode will be reflected without any change in shape, i.e. without mode conversion. The eigenmodes found above are therefore identical to the transverse modes of the waveguide array laser itself, since they may travel a round-trip through the cavity unaltered. If the laser can be forced to oscillate on just one of these modes (the lowest-order is the most uniform, and hence the most useful), the required phase-locking will have been introduced, because the output from each stripe will then have a fixed phase relationship with its neighbours.

To investigate the mode selectivity of the structure, we must examine the conditions for lasing. The total propagation constant of the \( \nu \)th transverse mode is \( \beta_0 + \gamma_v \), including the average term \( \beta_0 \). Only \( \text{Re}(\gamma_v) \) will affect phase variations, so the longitudinal resonance condition is therefore:

\[ 2\Lambda \{ \beta_0 + \text{Re}(\gamma_v) \} = 2\mu \pi \]

12.9-8

where \( \mu \) is an integer and \( L \) is the cavity length. Since \( \text{Re}(\gamma_v) \) is different for each transverse mode, we may deduce that each will have a different spectrum of longitudinal modes. Similarly, only \( \text{Im}(\gamma_v) \) governs the overall gain of each mode, so the condition for unity round-trip gain is:

\[ R_1 R_2 \exp \{ 2L \text{Im}(\gamma_v) \} = 1 \]

12.9-9
where $R_1$ and $R_2$ are the amplitude reflectivities of the mirrors. Examining Equation 12.9-6, however, we can see that $\text{Im}(\gamma_v) = g$ for all $v$. Consequently, the threshold gain must be the same for each transverse mode, so co-directional coupling has actually introduced no selectivity at all. We would therefore expect the laser to oscillate on a random mixture of all the possible transverse modes.

In practice, early array lasers tended to lase only on the highest-order mode. The explanation lay in the fact that only the active region of each stripe is pumped, so that there is very little gain between the stripes. An accurate calculation shows that the coupling coefficient is then complex, so that $\text{Im}(\gamma_v)$ is actually different for each mode, being greatest for the highest-order one. This is physically realistic; this particular mode has a field that changes in sign from guide-to-guide, falling to zero in between, so it is unaffected by interstripe loss. However, being the least uniform mode, it also is the least usable. Despite some attempts to introduce selectivity by separately pumping each stripe or using non-uniform stripe spacings, directionally coupled arrays have therefore largely been abandoned. Nevertheless, high outputs were achieved; for example, 2.6 W CW was obtained from a 40-stripe array.

A more effective mechanism is required force single-transverse-mode operation. The obvious solution is to introduce a mode-dependent loss, which reduces the effective gain of undesired modes while allowing desired ones to travel up and down the cavity unaffected. Now, in Chapter 9 it was shown that a single-mode Y-junction will radiate any anti-symmetric component that is injected at its forked end. Consequently, Y-junctions can be used to construct a filter that discriminates against transverse field patterns of high spatial frequency (which have fields of opposite sign in adjacent stripes). Figure 12.9-3 shows a device based on such a filter, the Y-junction-coupled array laser. At either end of the device, there is an array of uncoupled single-moded laser stripes. In this example, there are seven stripes in the left-hand array, and eight in the right. The two arrays are linked together in the central section by a network of back-to-back Y-junctions, leaving two unconnected stubs on the left. All sections of guide are equally pumped, and the structure is contained in a Fabry-Perot cavity.

![Y-junction coupling network](image)

**Figure 12.9-3** A Y-junction-coupled semiconductor array laser.

To analyse the structure, we assume for generality that there are $N$ stripes at the left-hand end of the array, and $N+1$ at the right. Once more, we assume that each guide has a total length $L$, an amplitude gain $g$ and a propagation constant $\beta_0$. Similarly, the end-mirror amplitude reflectivities are again $R_1$ and $R_2$. The Y-junctions act as linear beam splitters or combiners, as described in Chapter 9. We proceed by defining the amplitudes of the modes in the individual stripes at the left-hand end of the cavity as $a_1 \ldots a_N$ at the start of a round
trip. It is then simple to show that the amplitudes $b_1 \ldots b_{N+1}$ before reflection at the right-hand facet are:

$$b_n = 1/2 \{a_n + a_{n+1}\} \exp(-j\beta_0 L) \exp(gL)$$

12.9-10

Here once again we have introduced two ficticious amplitudes $a_0$ and $a_{N+1}$ (those of the modes in the stubs), which are both zero. After reflection at the right-hand facet, propagation back to the left-hand end of the cavity, and further reflection, the new mode amplitudes $a_1' \ldots a_N'$ are:

$$a_n' = 1/4 \{a_{n+1} + 2a_n + a_{n-1}\} R_1 R_2 \exp(-j2\beta_0 L) \exp(2gL)$$

12.9-11

Now, if the original amplitudes $a_1 \ldots a_N$ represent an eigenmode, they should be related to the new amplitudes $a_1' \ldots a_N'$ by a constant factor (otherwise the pattern will have changed in a round-trip). We may thus put $a_n' = \delta a_n$, where $\delta$ is a constant. This allows us to solve Equation 12.9-11 using our previous methods. Initially, we ignore the finite extent of the array, and assume a solution as a kind of travelling wave, writing $a_n = A \exp(-j\phi_n)$. We then obtain:

$$\delta = 1/2 \{1 + \cos(\phi)\} R_1 R_2 \exp(-j2\beta_0 L) \exp(2gL)$$

12.9-12

However, since the array is actually finite, we must satisfy the boundary conditions $a_0 = a_{N+1} = 0$. The solution must therefore again be a standing-wave pattern, so that $a_n = A' \sin(\phi_n)$. Since it is clear that the analysis will give very similar results to those obtained for the co-directionally-coupled array, we may immediately deduce that $\phi$ is again restricted to the set of values given in Equation 12.9-5, and that the overall shapes of the modes at the left-hand end of the array are as specified in Equation 12.9-6.

Each mode will lase when its round-trip gain is unity, i.e. when $\delta = 1$. From Equation 12.9-12, we can see that the longitudinal resonance condition for all transverse modes is now the same, namely $2\beta_0 L = \mu \pi$. More interestingly, the threshold gain $g_\nu$ is now different for each, being given by:

$$g_\nu = 1/2L \log_e(1/f_\nu R_1 R_2)$$

12.9-13

Here the factor $f_\nu$ (the transmission of the Y-junction network for the $\nu^{th}$ mode) is:

$$f_\nu = 1/2 \{1 + \cos[\nu \pi / (N+1)]\}$$

12.9-14

From this, we can see that the highest transmission efficiency (slightly less than unity) is obtained for the lowest-order transverse mode, which must therefore have the lowest lasing threshold. In contrast, the transmission of the highest-order mode is close to zero, so its threshold will be very high. The Y-junction network has therefore introduced the desired selectivity in favour of a single transverse mode with a useful spatial distribution. Practical Y-junction-coupled lasers have excellent stability, showing little change in their near- or far-field patterns over the entire range of operation, and their output powers are typically of the order of 400 mW CW at present.

12.10 SURFACE-EMITTING LASERS

In the future, it may be desirable to arrange for the emission from a laser chip to be taken from its surface, rather than an end-facet. This would allow the fabrication of two-
dimensional arrays of lasers, which might have applications in optical processors. At the very least, it would enable an entire wafer of lasers to be tested for failures before individual devices are separated by cleaving. A variety of surface-emitting semiconductor laser geometries are currently being investigated. These are all still based on single-mode waveguides, but there are two major distinctions between types.

Firstly, the laser cavity may be in the conventional orientation, parallel to the surface. Feedback may be obtained without cleaving, by using distributed Bragg reflectors. If these are based on second-order gratings, diffraction from the grating fundamental can be used to obtain output coupling normal to the surface, as described earlier (although unnecessary cladding layers over the grating are generally removed, to minimise absorption). For example, Figure 12.10-1 shows a Y-junction-coupled, surface-emitting DBR array laser. A laser of this type has already been demonstrated with a 10 x 10 array of emitters occupying an area of 60 µm x 5 mm, with a total output power of more than 1 W.

Figure 12.10-1  A Y-junction-coupled, surface-emitting DBR array laser.

Alternatively, the cavity may be normal to the substrate surface. In this case, the device is known as a vertical cavity surface-emitting laser. Figure 12.10-2 shows a typical example, a circular buried heterostructure GaAs/GaAlAs laser. Although the double heterostructure is still formed from layers parallel to the surface, the cavity is a roughly cylindrical pillar approximately 6 µm in diameter and 7 µm long with its axis normal to the surface. The current is injected from a ring electrode into the lower end of the pillar, which is surrounded by lateral current-blocking layers. Feedback is provided by a gold/silica mirror at the bottom of the cavity, and by a silica/titania multilayer stack deposited in a deep well etched right through the substrate at the top. The optical output is taken from the upper mirror.

More recent vertical cavity lasers have used distributed feedback or distributed Bragg reflection, based on epitaxial multilayers, thus eliminating the requirement for external mirrors. Due to the short length of the cavity, the output consists only of a single longitudinal mode, a feature of all such lasers. The particular laser shown had a threshold current of 6 mA, and an output in excess of 1 mW. Although all surface-emitting lasers are at a very early stage of development, they represent a promising new technology, which may have many applications outside communications.
12.11 TRAVELLING-WAVE LASER AMPLIFIERS

Besides acting as a versatile source, the semiconductor laser can provide other functions. For example, if the end-facets are suitably coated so that they are anti-reflecting, laser operation will be suppressed due to the lack of feedback. However, optical gain will still be available in the cavity, provided the electron density is driven above threshold. In this case, the device may act as a travelling-wave laser amplifier (TWLA). Figure 12.11-1 shows the principle, and a typical application.

The TWLA is connected between two fibres, as an in-line amplifier. A relatively weak optical input will grow exponentially as it passes through the TWLA, emerging as a much stronger amplified signal. The maximum possible input is limited, because the gain saturates at high optical power, but this level is quite sufficient for use in typical systems. Recent devices based on double-channel planar buried heterostructure lasers show little polarization sensitivity (1 dB variation in gain between orthogonal polarizations), gains of $\approx 30$ dB.
(although this figure is reduced by coupling losses), and bandwidths of $\approx 75$ GHz. Unfortunately, the amplification is not noiseless, due to the presence of spontaneous emission. This may be minimised by passing the output through a narrowband filter, which blocks the majority of the wide-band spontaneous emission noise, while passing the desired signal. Figure 12.11-2 shows a fully packaged travelling wave laser amplifier with fibre pigtails.

![Figure 12.11-2 A travelling-wave laser amplifier (photograph courtesy Telesis, a publication of Bell Northern Research Ltd.).](image1)

Travelling wave laser amplifiers can have applications other than as simple repeaters. For example, Figure 12.11-3 shows an amplifying switch based on a passive coupler combined with two TWLAs. The coupler acts as a splitter, which merely directs the input to the two amplifiers. Since the output of each TWLA may be switched off by removing the bias current, the whole network acts effectively as a two-way switch. The isolation achieved in the off-state is good - typically 30 dB for a 200 $\mu$m long device - while the gain provided in the on-state is easily sufficient to make up for coupling losses. Fully integrated switches based on similar principles appear an attractive possibility for the future.

![Figure 12.11-3 An amplifying switch based on travelling wave laser amplifiers (after M.J.O'Mahony, IEEE J. Lightwave Tech. LT-6, 531-544 (1988); © IEEE 1988).](image2)
PROBLEMS

12.1 A silicon p-n junction diode has doping levels of $N_a = 10^{24} \text{ m}^{-3}$ and $N_d = 10^{23} \text{ m}^{-3}$. Estimate the electric field across the depletion layer, under 5 V reverse bias. Assuming that the mobilities of electrons and holes in Si are $\mu_e = 0.135 \text{ m}^2/\text{Vs}$ and $\mu_h = 0.048 \text{ m}^2/\text{Vs}$, respectively, estimate their transit times across the depletion layer. Discuss the implications for photodiodes.

[20.4 \times 10^6 \text{ V/m}; 0.1 \text{ psec}; 0.3 \text{ psec}]

12.2 A silicon p-n$^+$ diode has a cross-sectional area of 1 mm$^2$, and doping levels of $N_a = 10^{21} \text{ m}^{-3}$ and $N_d = 10^{24} \text{ m}^{-3}$. Assuming that the distance $L_p$ from the depletion layer to the p-side contact is 0.1 $\mu$m, and that the corresponding distance $L_n$ is $\gg L_p$, calculate the current through the diode under a forward voltage of 0.5 V. The hole and electron diffusion coefficients may be found using the Einstein relation $D_{e,h} = \mu_{e,h} kT/e$. If the minority carrier lifetimes are $\approx 1 \mu$s, are your assumptions valid?

[284 \text{ mA}]

12.3 Compare the performance of InP/InGaAsP phase modulators based on a) the electro-optic effect and b) minority carrier injection. For the former, you may assume that $r_{41} = -1.4 \times 10^{-12} \text{ m/V}$ is the relevant electro-optic coefficient, that the overlap factor $\Gamma$ is 0.5, and that the breakdown field is $3 \times 10^7 \text{ V/m}$. For the latter you may assume that $m_e^*/m = 0.077$, $m_h^*/m = 0.64$, and that an upper limit on the injected carrier density is $\Delta N = 10^{24} \text{ m}^{-3}$. The refractive index of InP is 3.2, and the operating wavelength is $\lambda_0 = 1.5 \mu$m.

12.4 A $G_{d0.47}In_{0.53}As$ PIN photodiode, with a cross-sectional area of 0.1 mm$^2$, produces a photocurrent of 0.996 $\mu$A when illuminated by a beam of 1.3 $\mu$m wavelength and 1 mW/cm$^2$ intensity. Calculate a) the quantum efficiency, and b) the photocurrent at 1.5 $\mu$m and 1.7 $\mu$m wavelength.

[0.95; 1.15 $\mu$A; 0]

12.5 A surface-emitting GaAs LED, emitting at $\approx 820 \text{ nm}$, launches 50 µW of optical power into a step-index fibre when driven by a current of 50 mA. Assuming that the fibre has a NA of 0.3, estimate the internal efficiency of the LED. If the minority carrier recombination time is 10 nsec, what is the radiative recombination time? Estimate the 3 dB modulation bandwidth of the LED.

[54%; 18.4 nsec; 27.6 MHz]

12.6. Show that the effective power gain experienced by a guided mode with normalised transverse electric field $E$ due to the presence of a non-uniform power gain distribution $g_p$ may be written as $g_{peff} = \langle g_p, |E|^2 \rangle$. Discuss the way in which the double heterostructure maximises $g_{peff}$.

12.7. Show that, if a non-unity confinement factor and non-zero losses are included in the calculation, the power obtained from a semiconductor laser in the lasing regime alters to $P = \eta h c (1 - I/I_l)/(e\lambda_e)$, where:

$$\eta = \{1/L \log_e(1/R_1 R_2)\} / \{[\Gamma\alpha_{pcore} + (1-\Gamma)\alpha_{pclad} + 1/L \log_e(1/R_1 R_2)]\}.$$

The table below shows measurements of the power out of one facet of a GaAs/ GaAlAs laser at different currents. The active volume is 250 x 0.1 x 3 $\mu$m, and...
emission is at 820 nm. Estimate the threshold current density, and the average power absorption coefficient in the cavity.

<table>
<thead>
<tr>
<th>I</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 mA</td>
<td>15 µW</td>
</tr>
<tr>
<td>30 mA</td>
<td>500 µW</td>
</tr>
<tr>
<td>35 mA</td>
<td>2 mW</td>
</tr>
<tr>
<td>37.5 mA</td>
<td>2.75 mW</td>
</tr>
</tbody>
</table>

12.8 Assuming that the drive current of a semiconductor laser is in the form of a DC bias current $I'$ plus a small time-varying modulation $I''(t)$, and that the optical output power is the sum of a constant term $P'$ and a similar time-varying term $P''(t)$, show that the differential equation governing $P''(t)$ is:

$$\frac{d^2 P''}{dt^2} + 2 \zeta \omega_r \frac{dP''}{dt} + \omega_r^2 P'' = \left\{ \frac{hc \omega_r}{\lambda_g} \right\} I''$$

where the resonant frequency and damping factor are:

$$\omega_r = \sqrt{G \lambda_g P' / hv} \quad \text{and} \quad \zeta = \frac{1/\tau_e + \tau_p \omega_r^2}{2 \omega_r}.$$

12.9 Solve the differential equation in Question 12.8 for the case of a step change in the modulating current from $I'' = 0$ to $I'' = i''$ at $t = 0$, assuming $\zeta$ is small. Sketch the response for $\zeta = 0.05$.

12.10 A Y-junction-coupled GaAs/GaAlAs array laser has nine lasing stripes at one end of the cavity, and ten at the other. Calculate the lasing threshold for all the transverse modes, assuming a cavity length of 500 µm. Sketch the near-field patterns of the lowest- and highest-order transverse modes at either end of the cavity. How would you calculate the far-field patterns?
SUGGESTIONS FOR FURTHER READING


Casey H.C., Somekh S., Ilegems M. "Room-temperature operation of low-threshold separate-confinement heterostructure injection laser with distributed feedback" Appl. Phys. Lett. 27, 142-144 (1975)


Wada O., Yamakoshi S., Hamaguchi H., Sanada T., Nishitani Y., Sakurai T. "Performance and reliability of high-radiance InGaAsP/InP DH LEDs operating in the 1.15 - 1.5 μm wavelength region" IEEE J. Quant. Elect. QE-18, 368-373 (1982)
CHAPTER
THIRTEEN

OPTICAL DEVICE FABRICATION

13.1 OVERVIEW

We will now consider some of the methods used to make the guided wave devices discussed in the preceding Chapters. Because the range of these devices is itself so great, it is hardly surprising that many different techniques are needed. However, these can be broken down reasonably conveniently into two groups, which we will describe separately.

Firstly, we note that planar and channel guide integrated optics are both analogous to very large scale integrated (VLSI) microelectronics, in that the aim in each case is to fabricate a number of small, interconnected devices on a single substrate. LEDs, semiconductor lasers and detectors are also mass-produced on common substrates, which are subsequently separated into individual elements by dicing. The planar processing strategy of VLSI can therefore be used for optoelectronics, and many of the individual steps involved can be carried out using similar equipment.

In contrast, the aim of optical fibre manufacture is to produce a waveguiding medium in extremely long lengths. Since this lacks any planar characteristics or localized features, the fabrication procedure is entirely different. In fact, the whole operation is much more akin to wire-drawing - optical fibre is also manufactured by a drawing process, but in a unique construction known as a fibre-pulling tower. While it can be pulled directly from melts, a preform is generally used in modern processes. In effect, this is a large size 'copy' of the fibre, which is drawn down to the final cross-section. Such preforms are often built up using deposition methods analogous to those developed for planar substrates.

13.2 PLANAR PROCESSING

The basic steps involved in planar processing are as follows. The starting point is always a substrate, which must be specially prepared so that it is correctly orientated, polished flat, and cleaned. Thin layers of material can then be added to the substrate surface. If the additional material is amorphous, it may be deposited directly onto any type of substrate; crystalline layers, on the other hand, may only be added to crystalline substrates, and then only by strain-free epitaxial growth. However, both of these techniques may be carried out in a number of different ways, each offering a specific advantage. Alternatively, the properties of the substrate can be modified near its surface, for example, by diffusion or implantation of different material species. Finally, material can be removed from the surface by etching. All of these techniques can be applied to the entire area of the substrate, but it is often necessary to delineate small areas for local processing. This is carried out using a pattern transfer process called lithography, which itself involves deposition and removal of material.

Because the dimensions involved in most optoelectronic devices are very small (of the order of 1 µm - waveguide tracks are usually several microns wide, while grating structures normally have sub-micron periodicity), nearly all the processing stages are carried out in a special-purpose room known as a clean room. This is fed with filtered air to remove any dust particles that might otherwise introduce flaws during pattern transfer, and maintained at constant temperature and humidity to ensure process repeatability. Similarly, the services to the room (water, and process gases) are also highly purified, and clean-room technicians wear special clothing, designed to minimise the contamination they introduce. These precautions greatly increase the yield of working devices, and improve their reliability and
lifetime (extremely important in devices destined for undersea use, where replacement of failed components is difficult and costly).

Finally, after the majority of the process operations are completed, the device must be packaged and sealed. At this point, any required electrical connections must be made by wire bonding between electrode contacts on the substrate and terminals on the package. Similarly, any optical connections - e.g., optical fibre pigtails - must be attached by precision mechanical alignment.

13.3 SUBSTRATE GROWTH AND PREPARATION

We begin by considering typical substrate materials. For passive integrated optic devices, the substrate could be a sheet of amorphous material (say, glass) or, alternatively, a wafer of crystalline material such as Si might be required as a substrate beneath a passive silica guide for its mechanical or electronic properties. For electro-optic devices, insulating crystalline substrates (e.g., LiNbO3) are used, while light-emitting devices require semiconducting materials (InP or GaAs). Substrate shapes and dimensions vary widely, but usually the thickness is of the order of 0.5 mm, while the lateral dimensions are typically several centimetres.

Crystalline substrates originate from a boule, which is itself a large, perfect crystal of the material concerned. In general, the process of crystallization can involve a phase change from any of the following phases - vapour, liquid, or amorphous solid - into a crystalline solid phase. However, the liquid-to-crystalline solid transition is used most often. In this case, the crystal is grown by the slow solidification of material at the interface between a liquid (a highly purified melt of the desired material composition) and a solid (which itself originates as a small, perfect seed crystal), so that each successive plane of atoms attaches to the previous one in the correct relative position.

For example, silicon is often grown by the Czochralski process, using a crystal-pulling rig. Here, a large charge - normally, tens of kg - of previously-prepared electronic-grade Si is placed in a crucible, and melted by an R.F. induction heater or a resistive heating coil. Ideally, the crucible should be extremely unreactive with molten Si, to avoid contamination by impurities. Now, the melting point of silicon is 1412°C, which eliminates most potential crucible materials except fused silica (SiO2). Since this is rather fragile, the crucible requires further mechanical support by a graphite susceptor, and is normally destroyed at the end of the run by stress fracture on cooling. To avoid contamination, melting is performed in either an inert gas atmosphere (e.g. argon) or vacuum.

The seed crystal is held by a chuck on the end of a long screw thread, so that the desired growth plane is accurately parallel to the surface of the melt. At the start of the growth, the seed is lowered into the melt, and it is then gradually withdrawn by a slow rotation of the screw. The crucible is generally also rotated, and slowly raised to keep the surface level at a constant height as the melt is consumed. Careful control of a number of parameters - e.g., temperature, crucible rotation rate and pulling rate - is required for defect-free growth, and to maintain a uniform boule diameter. However, commercial processes are now highly automated, and extremely large boules are grown routinely. These may be several metres in length, and yield wafers of 8" diameter after grinding and sawing.

Gallium arsenide (melting point 1238°C) may also be prepared by Czochralski growth, using a variant known as the liquid encapsulated (LEC) process. The basic difference is that since GaAs decomposes when heated, the As vapour pressure above the melt must be
controlled or the stochiometry of the melt will alter as the As evaporates. This is done by covering the melt and the hottest part of the growing crystal in a thick layer of molten boric oxide (B₂O₃, a glassy viscous liquid insoluble to both Ga and As), while pressurising the growth chamber to 300 - 500 psi using an inert gas (e.g. N₂). Other differences are the use of a boron nitride crucible, and the application of a magnetic field to the melt, to damp down thermal convection through the 'magnetic viscosity' effect; this greatly reduces the number of dislocations in the crystal. GaAs growth technology is less highly developed than that used for silicon, but wafers are now available up to four inches in diameter. Indium phosphide (melting point 1062°C) also suffers from the problem of decomposition on heating, and so must be grown by the LEC process. There are two additional problems, however: the dissociation pressure of InP is higher than that of GaAs, and the increased viscosity of boric oxide at 1062°C results in an increase in stacking faults and a smaller yield of perfect crystals.

After recovery from the crystal puller, the boule is reshaped. The ends are first removed by sawing, together with any flawed regions, and the remainder is ground into a cylinder on a lathe. The crystal axes are then located by x-ray diffraction, and orientational flats are ground on the cylinder so that the axes may be identified later on. The boule is then sawn into wafers, using a diamond-impregnated saw blade. To minimise wastage (or kerf loss), this must be as thin as possible. For small boules, a disc-shaped blade may be used, with the cutting edge at its periphery. This type of blade cannot be used for large boules, however, because of the likelihood of buckling as the blade diameter increases. Instead, an annular blade is used; this is supported and tensioned at its outer rim, while the inner rim provides the cutting edge. After separation, the wafers are lapped flat, and their edges are rounded (to minimise edge chipping in later mechanical handling). The wafers are then chemically etched to a depth of about 10 µm (more for GaAs), to remove crystallographic defects induced by the cutting and shaping operations. Finally, one or both surfaces are polished, to a flatness of better than 2 µm.

13.4 DEPOSITION AND GROWTH OF MATERIAL

(1) DEPOSITION

There are several methods of depositing material in the form of dielectric or metallic films, either for insulating layers or for contacts. All are limited to about 1 µm coating thickness. The simplest technique is vacuum evaporation. The principle involved is trivial; the substrate and the coating material are both placed in an evacuated enclosure, some distance apart. The coating material is then heated to its melting point, so that it evaporates. Enough thermal energy is supplied to enable individual atoms to escape from the surface of the molten material, and (since there are no intervening gas molecules to collide with) these travel in a straight line towards the substrate, where they adhere. Several methods may be used to melt the evaporant. For example, Figure 13.4-1a shows resistive heating; here, the evaporant is simply loaded into a coiled wire (typically, a tungsten filament), through which a heavy D.C. current is passed. Alternatively, the heating element may be boat-shaped, and made of sheet metal (e.g. Mo) or graphite.

Many metals (e.g. Ag, Au and Al) may be evaporated very successfully using resistive heating. However, some metal evaporants (e.g. Ti) have high melting points, and are extremely reactive when molten. This may lead to dissolution and failure of the filament, and contamination of the film by impurities. Similarly, many dielectrics have even higher melting points, and also very poor thermal conductivity. Consequently, alternatives to resistive heating are often required.
One suitable alternative is electron-beam heating, shown in Figure 13.4-1b. Here a hot-wire filament is again used, but this time simply as a thermionic electron emitter. A charged aperture above the filament then accelerates the electrons in the vertical direction (through, say 5 keV) while simultaneously limiting them to a roughly collimated beam. A magnetic field is arranged to lie perpendicular to the direction of the electron beam, so that it is bent into an arc and falls on a water-cooled crucible containing the evaporant. This is now heated by the kinetic energy of the electrons, rather than by direct contact with a hot body, eliminating the problems described above.

Figure 13.4-1c shows a complete evaporator. The evacuated enclosure is a glass bell-jar, held at a pressure of $\approx 10^{-6}$ Torr; higher pressures result in oxide formation during the evaporation of metal films. The sample is held face down over the evaporant by a support frame, and evaporation takes place upwards. The thickness of the film is monitored, by allowing the simultaneous coating of a small quartz crystal, arranged as a part of an oscillator circuit. Changes in the resonant frequency of the circuit may then be related to the mass loading of the crystal, and hence to the thickness of the coating. The uniformity of the deposition may be improved by mounting the substrate on a planetary stage, so that it follows a rotating, spinning orbit during coating. It should be noted that, due to the direct line-of-flight of atoms from the evaporant to the substrate, evaporation cannot be used to coat any features of the surface topology that lie 'in shadow'.

An entirely different method of vacuum coating is R.F. sputter-deposition, which can provide conformal coatings of both metals and dielectrics, albeit at the expense of requiring a more complicated apparatus. Here the substrate is placed much closer to the coating material, which is normally arranged in a large sheet known as the target (Figure 13.4-2). This time, the process does not take place under high vacuum; instead, an inert gas (e.g. Ar) is bled into the chamber at a pressure of $\approx 10^{-3} - 10^{-2}$ Torr. The gas is then excited into a plasma - a gaseous mixture of ions and free electrons - by an A.C. radio-frequency electric field (usually at 13.56 MHz, a standard scientific frequency) in the space between target and...
substrate. However, the former is arranged as the cathode, and the latter as the anode, so that positively charged Ar\(^+\) ions strike the target preferentially, and eject atoms from it. This process - known as **sputtering** - is a physical one, which operates by transfer of momentum between the incident, fast-moving ion and the static atom. Although the sputtered atoms are ejected in random directions, many move towards the nearby substrate, where they adhere to the surface.

![Apparatus for planar RF-sputtering.](image)

Because of further collisions with heavy ions inside the plasma, the sputtered atoms strike the substrate from a wide range of angles, so the technique can be used to coat non-planar substrates. However, since argon atoms are chemically inert, no additional reactions take place. Sputtering may also be induced using a D.C. field, but with a dielectric target the arrival of positively charged ions leads to a build-up of surface charge, which very quickly halts the process. Consequently, D.C. sputter-deposition is only appropriate for metals. Naturally, just as the target is bombarded by positive ions, the substrate is continually struck by electrons. Since these are much lighter, they do not eject material, but they can raise the substrate temperature considerably (by several hundred degrees), unless water-cooling is used. Alternatively, a permanent magnetic field may be used to force the electrons to follow cycloidal tracks, which avoid the substrate for the most part. The sputtering rate is highly dependent on the target material, and is generally low (≈0.01 μm/min for glassy materials). The time taken to deposit material to the thickness required for a guiding layer (microns) may therefore be large. Despite this, sputtering was one of the first techniques used successfully for the fabrication of step-index planar waveguides, using SiO\(_2\) and Al\(_2\)O\(_3\) films.

We now consider a further method of deposition, **chemical vapour deposition** (or CVD). This is not a vacuum process. Instead, a hot gas mixture (often at atmospheric pressure) is passed over the substrate in a furnace, so that a film of the correct composition is gradually deposited following chemical reactions, which take place at the substrate surface. The physical layout of the equipment used can take a variety of different forms, depending on the direction of the gas flow (which might be horizontal or vertical) and the form of heating used. For example, Figure 13.4-3 shows a horizontal-tube hot-wall reactor, consisting of a quartz tube with an external resistive heater. The substrate is placed on a quartz support, and loaded into the tube through a removable end-cap. The process gases are injected from one end, and heated prior to arrival at the substrate by contact with the walls. They flow over the substrate as a laminar boundary layer, and are adsorbed on the surface where they react to
form the desired film. By-products of the reaction and unreacted process gas are removed to exhaust at the downstream end of the tube.

Figure 13.4-3 A horizontal-tube hot-wall CVD reactor.

The temperature and pressure are both highly dependent on the process used. One major branch of CVD involves the thermally activated oxidation of hydrogen compounds, such as silane (SiH₄), phosphine (PH₃) and arsine (AsH₃). For example, silica layers may be deposited at atmospheric pressure by a mixture of silane and oxygen, at a temperature of 400 - 450°C, following the reaction:

\[
\text{SiH}_4 + \text{O}_2 \rightarrow \text{SiO}_2 + 2\text{H}_2
\]

Similarly, silicon nitride might be produced from dichlorosilane and ammonia at \( \approx 750°C \), or from silane and ammonia at \( \approx 900°C \), following:

\[
3\text{SiCl}_2\text{H}_4 + 4\text{NH}_3 \rightarrow \text{Si}_3\text{N}_4 + 6\text{HCl} + 6\text{H}_2 \\
\text{or} \quad 3\text{SiH}_4 + 4\text{NH}_3 \rightarrow \text{Si}_3\text{N}_4 + 12\text{H}_2
\]

As a final example of the versatility of the basic CVD process, we mention that it is possible to fabricate doped glassy layers easily. For example, phosphosilicate glass (SiO₂ doped with P₂O₅) may be produced following the two simultaneous reactions:

\[
\text{SiH}_4 + 2\text{O}_2 \rightarrow \text{SiO}_2 + 2\text{H}_2\text{O} \\
2\text{PH}_3 + 4\text{O}_2 \rightarrow \text{P}_2\text{O}_5 + 3\text{H}_2\text{O}
\]

The index changes needed in planar silica-based waveguides (for example) may therefore be obtained merely by alterations in the process gas flows.

A number of important variations on the standard CVD process exist. For example, SiO₂ may also be produced by decomposing the metal organic compound TEOS (tetraethylorthosilicate, \( \text{Si(OC}_2\text{H}_5)_4 \)), at a reduced pressure and a higher temperature (650 - 750°C). A profusion of acronyms is used to distinguish these process variants; low-pressure processes come under the heading of LPCVD, while the initials MOCVD describe processes based on metal organic compounds. The latter are rapidly increasing in importance for the growth of thin crystalline films, as we will describe shortly.

Though CVD is conceptually a simple process, in reality it is significantly complicated by two aspects. The first is the difficulty of ensuring uniform film deposition - in terms of both thickness and stochiometry - since the temperature, concentration, chemical composition and velocity of the gas mixture can vary considerably along the length of the tube and across its diameter. The second is the need to handle gases that are often toxic, explosive or corrosive (or a combination of all three). Some gases are highly dangerous; in particular, we mention
silane, which is both toxic and explodes in contact with air. Double-layer stainless steel pipework is therefore a standard safety precaution.

Using a modification of the CVD process, it is possible to deposit films at lower substrate temperatures (200 - 350°C), by supplying the necessary energy from an electrically excited gas plasma rather than as heat. This variant is known as **plasma-enhanced chemical vapour deposition**, or PECVD. Figure 13.4-4 shows typical apparatus; the substrate is placed on a heated susceptor, which is arranged to act as one of a pair of R.F. electrodes. A plasma is then established between the electrodes, and the process gases are then bled into this region (which may have a very high electron temperature). The plasma might be derived from an inert gas; for example, silicon dioxide films may be deposited by reacting silane and nitrous oxide in an argon plasma. Alternatively, it might itself form one of the reactants. Considerable interest is currently being shown in PECVD processing for the fabrication of waveguides on silicon substrates. These require a thick dielectric spacer (at least 12 µm of SiO₂) to separate the waveguide (doped silica) from the substrate.

![Figure 13.4-4 Basic PECVD apparatus.](image)

(2) **EPITAXIAL GROWTH**

All of the above processes deposit amorphous films, with a random rather than a regular crystalline structure. To add material while maintaining a regular arrangement of atoms (as required in III - V optoelectronic devices), the layer must be grown epitaxially. This can be done by matching the lattice parameters of the film to those of the substrate, so that the latter acts as a template for ordered growth. If the grown material has the same chemical composition as the substrate (e.g., as when a GaAs layer is grown on a GaAs substrate), the process is called **homoepitaxy**. Often, however, the two materials have different compositions. In this case, the process is known as **heteroepitaxy**. For guided wave optical devices, there are currently two important materials systems. Firstly, it is possible to grow the ternary alloy gallium aluminium arsenide (Ga₁₋ₓAlₓAs) on GaAs substrates, over the complete range of alloy composition, since the lattice dimensions of the former are almost entirely independent of the mole fraction x.

Secondly, the quaternary alloy indium gallium arsenide phosphide (In₁₋ₓGaₓAs₁₋ₚₚ) can be grown on InP substrates. In this case, the lattice dimensions are a strong function of alloy composition. To a reasonable approximation, the lattice parameter a(x, y) of the alloy obeys **Vegard's law**, and may be described by the polynomial function:
\[ a(x, y) = xy a_{\text{GaP}} + x(1 - y) a_{\text{GaAs}} + (1 - x)y a_{\text{InP}} + (1 - x)(1 - y) a_{\text{InAs}} \]

where \( a_{\text{GaP}} = 5.4512 \text{ Å} \), \( a_{\text{GaAs}} = 5.6536 \text{ Å} \), \( a_{\text{InP}} = 5.8696 \text{ Å} \), and \( a_{\text{InAs}} = 6.0590 \text{ Å} \) are the lattice parameters of the binary compounds GaP, GaAs, InP and InAs, respectively. This reduces to a linear relation for any possible ternary alloy. Figure 13.4-5 shows contours of constant lattice spacing on the \( x - y \) plane. Also shown is the locus of all compounds that are lattice matched to InP, which may be grown without strain on InP substrates. This is almost exactly a straight line, obeying the relation \( y \approx 1 - 2.14x \). When \( y = 0 \), the locus intercepts the \( x \)-axis at \( x \approx 1/2.14 \), i.e. at the alloy \( \text{In}_{0.53}\text{Ga}_{0.47}\text{As} \).

Figure 13.4-5 Variation of the lattice parameter with composition, for InGaAsP.

Just as in crystal growth, epitaxial processes may be based on vapours or liquids, but epitaxial growth may be performed at temperatures well below the melting point of the substrate. Processes based on liquids are known as liquid-phase epitaxy (LPE); these exploit the principle that mixtures may have lower melting points than those of their stoichiometric counterparts. For example, a mixture of 5% of arsenic with 95% gallium melts at 880°C (considerably lower than the melting point of GaAs, 1238°C). However, when the mix is slowly cooled, some of the arsenic must precipitate out of solution. Most importantly, it does so stochiometrically, with each atom of As combining with one of Ga. The result is GaAs, which grows epitaxially on the substrate surface. Clearly, the ratio of gallium to arsenic in the melt will change in the process, but if the ratio is high to begin with, the effect will be negligible. As a result, continuous growth of GaAs films on GaAs substrates can be carried out, without melting the latter. \( \text{Ga}_{1-x}\text{Al}_x\text{As} \) films may be grown in a similar way, from melts containing Ga, Al and As, and \( \text{In}_{1-y}\text{Ga}_{x}\text{As}_{1-y}\text{P}_y \) can be grown from four-component melts. However, it is difficult to maintain a specified composition throughout the thickness of a layer.

Figure 13.4-6 shows one LPE technique, the horizontal slider method for the fabrication of GaAlAs/GaAs epitaxial films, which has been extremely successful in its time. The key component is a carefully machined graphite boat, fitted with a horizontal graphite slider.
Normally, several substrates are placed in shallow wells machined into the top of the slider; one of these is the wafer destined for overgrowth, while the others are sacrificial dummies. Similarly, charges of the growth material are placed in deeper wells in the boat, which are sealed with further wafers capped with graphite plugs. The whole assembly is placed inside a resistively heated furnace tube, in an inert gas atmosphere, so that the growth material is melted. Using push rods, the relative positions of the boat and slider may be manipulated, so that the wafer is covered in the required melt. This will grow an epitaxial layer on the wafer surface when the temperature is suitably lowered. The dummy wafers are used to seal all unused melts. When it is time to change the composition of the epitaxial film, the slider is simply moved under a new melt. Due to the close fit of slider and boat, the old melt is wiped cleanly off the wafer surface, while the new melt is wiped on. Further growth then takes place over a prescribed program.

![Figure 13.4-6 The horizontal slider method for growth by liquid phase epitaxy.](image)

Although a simple process, LPE suffers some disadvantages. Firstly, some crystal faults are inevitably introduced between the layers as melts are exchanged, reducing the perfection of interfaces. Secondly, because of the high temperature, some interlayer diffusion also occurs; this renders the junctions between layers of different composition less abrupt, which can be undesirable for many device applications. Thirdly, since the process can operate relatively near to equilibrium, slight changes in temperature can cause dissolution of the substrate instead of growth. Finally, it is hard to maintain the composition of quaternary layers with the precision required for accurate lattice matching. As a result, LPE is rapidly being overtaken by methods based on vapour-phase epitaxy (VPE).

Vapour-phase epitaxy simply involves the use of chemical vapour deposition for epitaxial rather than disordered growth. In particular, non-equilibrium techniques based on metal organics (known as MOVPE) appear highly promising, especially for the growth of InP/InGaAsP films, which require considerably better process control than GaAs/GaAlAs structures. As in CVD processes, the Group V elements arsenic and phosphorus are derived from the hydrides arsine (AsH$_3$) and phosphine (PH$_3$), while the Group III elements indium and gallium are obtained from the metal alkyls trimethyl indium ((CH$_3$)$_3$In) and trimethyl gallium ((CH$_3$)$_3$Ga). For example, GaAs may be deposited at 600 - 800°C, following the reaction:

\[
(CH_3)_3Ga + AsH_3 \rightarrow GaAs + 3CH_4
\]

Deposition of InP follows a similar reaction, using trimethyl indium and phosphine, while InGaAsP requires all four constituents. Hydrides of sulphur or selenium (H$_2$S and H$_2$Se) are used as sources of n-type dopants for both InP and InGaAsP; p-type dopant sources include dimethyl zinc and dimethyl cadmium. Figure 13.4-7 shows a schematic of an MOVPE process for the growth of InGaAsP on InP. Here the hydrides and dopants are added directly to the vapour stream, while the metal organics (contained in temperature controlled baths) are transported by a carrier gas such as H$_2$ or He.
The main defect of earlier CVD processes - lack of uniformity - appears to be well under control in modern MOVPE rigs, which offer a number of advantages. Firstly, the stochiometry of the gas mixture may be set very accurately using mass-flow controllers. Secondly, any transitions between layers of different composition may be made merely by switching from one gas mixture to another, and gas switching may be extremely fast with electrically-controlled solenoid valves. Thirdly, the process temperature is lowered to ≈ 650°C, which reduces interlayer diffusion. As a result, the extent of any interlayer lattice imperfections is greatly reduced, and heterostructure interfaces may be extremely abrupt (a few atomic layers in thickness). This allows the fabrication of very thin, pure films. More importantly, multilayer films may be built up with very sharp boundaries between the layers. These may be used as multiple quantum well (MQW) structures.

A final growth method is molecular beam epitaxy, or MBE. This is a low-temperature process, based on evaporation under ultrahigh vacuum (10⁻⁸ - 10⁻¹⁰ Torr). Essentially, several crucibles containing the constituents of the growth material are heated (normally, by electron-beam heating) so that their contents evaporate simultaneously. The departing atoms combine in flight to form a molecular beam, which is then adsorbed on the substrate. Precise control of the evaporation rates of the separate materials is required to establish the correct stochiometry of the growing layer. However, the advantage is that growth occurs at such a low temperature that hardly any diffusion takes place between the layer and the substrate, so that MBE may also be used to grow MQW structures - in fact, it was the original MQW process. Its main disadvantage as a production tool is that the growth rate is extremely small compared with other epitaxial methods, typically 0.01 - 0.03 µm/hr.

13.5 MATERIAL MODIFICATION

There are several methods of modifying the properties of a substrate near its surface. The simplest is diffusion. Here the substrate is placed in intimate contact with the dopant material - which may be a solid, a liquid or a gas - and both are heated to 800 - 1000°C. Due to their greatly increased thermal energy, the dopant and substrate atoms are all considerably...
more mobile than at room temperature. However, the latter are often anchored reasonably well to fixed lattice sites, about which they vibrate. As they do so, some sites may be vacated, into which migrant dopant atoms may then move (as shown in Figure 13.5-1a).

Alternatively, dopant atoms may diffuse into the substrate without the necessity for such vacancies, by moving through the spaces between the lattice sites, known as interstitials (Figure 13.5-1b). The overall process is known as in-diffusion. It is often possible to avoid any large-scale transport of material in the opposite direction; however, this can and does occur, and is known as out-diffusion; it may involve either host atoms or impurities.

Mathematically, diffusion is governed by two rate equations, whose derivation is based on the assumption that the behaviour of any system of mobile particles - for example, a gas - is a random process driven by concentration gradients. Since we have already derived the necessary mathematics in Chapter 11 for the analogous phenomenon of carrier diffusion, we will not repeat it here. Suffice it to say that a one-dimensional gradient in the dopant concentration $C$ leads to a flux of material of:

$$J = -D \frac{\partial C}{\partial x} \quad 13.5-1$$

where $D$ is the diffusion coefficient. Equation 13.5-1 is known as Fick's first law; it shows that a non-uniform concentration will result in a flux of material in the opposite direction to any concentration gradient. Now, at the start of an indiffusion process, there is a large difference in the value of $C$ inside and outside of the substrate. Consequently, dopant atoms will tend to diffuse inwards, altering the local concentration in the process. To analyse this effect, we require a continuity equation. Once again, a suitable equation has already been derived for carrier diffusion in Chapter 11. Restated in terms of the dopant concentration, this becomes:

$$\frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x} \quad 13.5-2$$

Equation 13.5-2 shows that a flux gradient will result in a decrease in the local concentration, which tends to equalise the concentration everywhere. Assuming the diffusion coefficient is independent of concentration, we may combine Equations 13.5-1 and 13.5-2 to obtain a single equation in $C$ alone:

$$\frac{\partial C}{\partial t} = D \ \frac{\partial^2 C}{\partial x^2} \quad 13.5-3$$
Equation 13.5-3 - known as Fick's second law, or the diffusion equation - may be solved to find the concentration as a function of position and time, subject to given boundary conditions.

In-diffusion is often used for the doping of semiconductors (e.g. to form p-n junctions). It is also used for waveguide fabrication, for example in the Ti:LiNbO$_3$ process. This is illustrated in Figure 13.5-2, which shows the result of diffusing a thin surface layer of Ti metal into a LiNbO$_3$ substrate. In this case, it can be shown that the solution to Equation 13.5-3 for times larger than that required simply to diffuse all the metal into the crystal is:

$$C(x, t) = \frac{S}{\sqrt{\pi Dt}} \exp[-x^2/(4Dt)]$$

where $x$ is the distance from the surface, and $S$ is the amount of dopant per unit area at $t = 0$. This solution is plotted in Figure 13.5-2, for a three successive times $t$. As can be seen, the concentration profile is Gaussian; as the diffusion time is increased, the surface concentration is lowered, but the Ti atoms penetrate further into the crystal.

The quantity $L_D = 2\sqrt{Dt}$ is known as the diffusion length. It depends on both the diffusion coefficient and the time; for a deep diffusion, either the diffusion coefficient must be large or a lengthy diffusion must be used. Now, the dependence of $D$ on temperature is normally of the form:

$$D = D_0 \exp(-E_a/kT)$$

Where $E_a$ is an activation energy. To reduce the process time, it is therefore normally carried out at an elevated temperature, at which the dopant molecules are more energetic (giving a large value of $D$).

Ti indiffusion causes an increase in refractive index, that is proportional to the local concentration of Ti metal (at least, for one polarization). Consequently, a planar guide with the index variation:

$$n(x) = n_0 + \Delta n \exp\{-(x/L_D)^2\}$$

may be obtained very simply, by coating a substrate with a layer of metal and placing it in a furnace for a number of hours (e.g. for 8 hours, at 1050°C). More generally, diffusion is a three-dimensional process. In this case, we may replace the one-dimensional flux $J$ by a
vector $\mathbf{J}$ indicating the flow direction as well as its strength, and Equations 13.5-1 and 13.5-2 modify to:

$$\mathbf{J} = -D \nabla C$$ and $$\frac{\partial C}{\partial t} = -\nabla \cdot \mathbf{J}$$

Similarly, the diffusion equation becomes:

$$\frac{\partial C}{\partial t} = D \nabla^2 C$$

Equation 13.5-8 may be used to model the effect of, say, the diffusion of a narrow strip of Ti metal into a LiNbO$_3$ crystal. In this case, the concentration profile varies in two dimensions, so it is not hard to see that a two-dimensional index change (and hence a buried channel guide) will be formed.

An alternative to diffusion is **ion exchange**. This time, the substrate is placed in a liquid melt, usually at a much lower temperature (200 - 400°C). Normally, there is a higher concentration of one type of mobile ion inside the substrate than in the melt, and a higher concentration of another inside the melt than in the substrate. As a result, these two species counter-diffuse, so that material is exchanged between substrate and melt. Ion exchange may also be used for waveguide fabrication, since differences in the polarizability of the two species cause a change in refractive index. For example, Figure 13.5-3 shows a process for the fabrication of channel guides in soda-lime glass (a mixture of SiO$_2$, Na$_2$O and other metal oxides). Here, the glass substrate has been coated in a layer of mask material, in which narrow stripe openings have been made, and placed in a bath of molten AgNO$_3$. The mask acts as an effective barrier to the ions, so that ionic motion only occurs through the openings. In these regions, Na$^+$ ions diffuse out of the substrate and into the melt, to be replaced by Ag$^+$ ions. The exchange of relatively light Na$^+$ ions with heavier Ag$^+$ ions causes an increase in refractive index in the region of the stripe opening, and hence a channel guide. The motion of the Ag$^+$ ions into the substrate may be enhanced by the application of an electric field, which drastically reduces the process time and allows deep, uniform diffusion profiles.

![Figure 13.5-3 The silver-sodium ion-exchange process for channel guide fabrication in glass (after C.D.W.Wilkinson et al., Elect. Lett. **14**, 599-600 (1978)).](image)

Waveguides may be fabricated in LiNbO$_3$ using a similar process known as **proton exchange**, which involves the replacement of Li$^+$ ions by H$^+$ ions from a melt of benzoic acid. However, owing to the crystallinity of the substrate, the refractive index change that
results is strongly dependent on the crystal orientation and the optical polarization. For example, Figure 13.5-4 shows typical refractive index profiles for X-cut, MgO-doped LiNbO₃, exchanged for 3 hours in pure benzoic acid. The change in ne near the surface is large, positive, and almost constant, so a near-step-index guide will be formed. However, the change in no is actually negative, the opposite sign to that required. A waveguide is therefore formed for one polarization only.

![Refractive index change](image)

Figure 13.5-4 Change in refractive index with depth, for a typical proton-exchange process (data taken from M. Digonnet et al., Opt. Lett. 10, 235-237 (1985)).

As a final example of a technique for material modification, we mention ion implantation, which is mainly used for doping semiconductors. Here, the change also follows from the insertion of ions into the substrate, but in an entirely different manner. The process is carried out under high vacuum, and the equipment required is both sophisticated and expensive. The source of ions is a crucible of molten material, from which a beam of ions is extracted electrostatically. This may contain several different types of ion - singly- and doubly-charged ions of the desired element, impurity ions, and so on - so it is passed through a filter sensitive to charge and atomic weight (known as a mass filter, or a Wien filter) to select a single ion species.

The emerging beam is then accelerated by a high voltage (typically 10 - 1000 kV) so that ions strike the substrate at speed. Instead of sputtering atoms from the surface, they now have enough energy to penetrate some distance before stopping. As they slow, the ions lose energy by undergoing several successive collisions with the atoms in the substrate, as in Figure 13.5-5a. Since the substrate atoms may be displaced considerably from their normal positions as a result, a thermal annealing step is almost always required after implantation.

One attractive feature is that the distribution of implanted ions is very well defined, and controllable by varying the acceleration voltage. To a reasonable approximation, the variation of the concentration of implanted ions with depth follows the Gaussian distribution:

\[
N(x) = \frac{\phi}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - r_p)^2}{2\sigma^2}\right\}
\]

13.5-9
Here $\phi$ is the density of implanted ions per unit surface area, defined by $\phi = \int_0^\infty N(x) \, dx$, $r_p$ is the projected range of the ions, and $\sigma$ is the standard deviation of the projected range. The peak concentration occurs at $x = r_p$, and is given by $N_p = N(r_p) = \phi/\sqrt{(2\pi)\sigma} \approx 0.4\phi/\sigma$. $r_p$ itself depends on several parameters: the mass and charge of the ionic species, the substrate material, and the acceleration voltage. However, $r_p$ is generally limited by the maximum practical voltage ($\approx 1$ MV) to about $1 \mu$m.

When plotted on a logarithmic scale, the concentration distribution becomes a parabola; Figure 13.5-5b shows typical profiles obtained for three different acceleration voltages, such that $V_1 < V_2 < V_3$. The larger the voltage, the deeper the implantation, but the broader and weaker the distribution. Other, more complicated profiles may be synthesised from the cumulative effect of a succession of implants, each with a different acceleration voltage and implantation dose, or from a combination of an initial implantation with a drive-in diffusion step.

Figure 13.5-5 Ion implantation: a) schematic, and b) distribution of implanted ions, for different acceleration voltages.

Ion implantation is potentially attractive for the direct fabrication of waveguides, since it suffers less from the lateral spreading inherent in comparable processes like diffusion. Channel waveguide formation has been investigated in a number of crystalline materials (e.g. LiNbO$_3$, by implantation of He$^+$ ions). However, the lattice damage causes high optical propagation loss and a significant reduction in the electro-optic coefficient, which is only partially restored by annealing.

### 13.6 ETCHING

Material removal or etching methods can generally be subdivided into classes, which are based either entirely on physical or chemical principles, or on some combination of the two. Additional divisions may also be made between those techniques which do not require the use of vacuum and those which do, and between masked and maskless processes.
Wet chemical etching is commonly required to pattern thin metal and dielectric films on the substrate, either to fabricate specific features (e.g. electrodes), or to open windows in surface mask layers for subsequent local processing. These materials normally etch isotropically, so that the fidelity of the pattern transfer is limited by undercut of the etch mask. However, etches are also often used to alter the surface topology of crystalline substrates. In this case, the etching may follow a particular direction. Anisotropic etching can be performed in crystals with diamond and zincblende lattices (e.g. Si, GaAs and InP), because the <111> crystal plane is more closely packed than the <100> plane, and so is normally etched more slowly.

For example, using the water: ethylene diamine pyrocatechol (EDP) etch system, the etching rates for <100>, <110> and <111> planes in Si are 50, 30 and 3 µm/hr at 100°C. The <111> etch rate is clearly considerably slower. By choosing the substrate orientation correctly, it is therefore possible to fabricate vertical mesa shapes as shown in Figure 13.6-1a, or inclined grooves as in Figure 13.6-1b. The former topology is appropriate for ridge guides in GaAs/GaAlAs and InP/InGaAsP, while the latter is most often used to provide V-groove alignment features for optical fibres in Si.

Suitable etch mixtures for a number of important amorphous and crystalline materials are summarised in Table 13.6-1.

<table>
<thead>
<tr>
<th>Material</th>
<th>Etch System</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al</td>
<td>HNO₃ / CH₃COOH / H₃PO₄ / H₂O</td>
<td>Used for electrode patterning</td>
</tr>
<tr>
<td>Au</td>
<td>KI / I₂ / H₂O</td>
<td>Used for electrode patterning</td>
</tr>
<tr>
<td>Ti</td>
<td>HF / HNO₃ / H₂O</td>
<td>Used to pattern Ti prior to diffusion into LiNbO₃</td>
</tr>
<tr>
<td>SiO₂</td>
<td>HF / H₂O / NH₄F</td>
<td>Used to open windows in SiO₂ surface mask</td>
</tr>
<tr>
<td>GaAs</td>
<td>H₂SO₄ / H₂O₂ / H₂O</td>
<td>Used to etch rib guides on &lt;100&gt; substrates</td>
</tr>
<tr>
<td>InP</td>
<td>HCl / H₃PO₄</td>
<td>Has negligible effect on InGaAsP</td>
</tr>
<tr>
<td>InGaAsP</td>
<td>H₂SO₄ / H₂O₂ / H₂O</td>
<td>Has negligible effect on InP</td>
</tr>
<tr>
<td>Si</td>
<td>KOH / H₂O / Isopropyl alcohol</td>
<td>Used to etch V-grooves in &lt;100&gt; substrates</td>
</tr>
</tbody>
</table>

Table 13.6-1 Wet etch mixtures for a variety of materials used in optoelectronic device fabrication.

Figure 13.6-1 Effect of anisotropic etching of a) <110> and b) <100> substrates.
DRY ETCHING

Although wet etching is still used to a considerable extent, many modern microfabrication procedures involve one of a battery of rather different vacuum etching techniques, all loosely known as dry etches. Although these require complicated equipment, they have the advantage of greatly improved process control and a high degree of selectivity and anisotropy in the etching.

In dry etching, either purely physical or combined physical/chemical methods are used. The simplest physical method is sputtering, in which atoms are ejected from a surface by ion bombardment. The most obvious implementation is R.F. sputter etching, a variant of R.F. sputter deposition (previously shown in Figure 13.4-2). The electrode connections are simply reversed, so the substrate is bombarded by positive ions, rather than the target (which may be a simple electrode). This technique is often used to clean a substrate prior to deposition, but it may also be used as a method of etching. However, since the ions strike the substrate from a range of angles, etching is roughly isotropic.

Another physical method is ion beam milling, shown in Figure 13.6-2. This is similar to D.C. sputter etching (performed by reversing the electrode connections in a D.C. sputter coater); however, the plasma is generated rather differently. Here an inert gas (e.g. argon) is bled into the far end of an evacuated chamber. A thermionic emitter is then used to generate electrons, which ionise the argon. Generally, the interaction between the electrons and the plasma is increased by using a magnetic field, which makes the electrons follow a long spiral trajectory. The Ar ions are extracted from the plasma and accelerated towards the target by a biased grid, and atoms are then sputtered off the substrate by this beam in the normal way. To prevent the accumulation of positive charge at the substrate, a neutralizer filament is used. This is a further thermionic emitter, which injects electrons into the beam. In contrast to R.F. sputter etching, the ions are incident from a defined direction, so that the etching is anisotropic. However, this advantage is limited by ion 'bounce', which causes ions to recoil in random directions, eroding the walls of any deep features. The substrate can be rotated to ensure even etching, and enhanced etch rates in most materials result from machining at a tilt.

![Figure 13-6-2 An argon ion beam mill.](image-url)
Most often, it is necessary to apply particular processes locally; for example, we might wish to define a small area of a substrate for sputter etching or ion milling. This can be done by depositing a layer of a mask material over the substrate, with openings where the etching is to be performed. Clearly, the property required of the mask material is that it should have a much lower sputter yield than the substrate, so that it erodes more slowly. Unfortunately, most purely physical etching techniques offer little differentiation between materials. As a result, the majority of the mask is etched at roughly the same rate as the substrate, and its edges (which define the periphery of the etched area) are often attacked even faster, since they present tilted faces to the beam. This limits the depth of etched features, and prevents the fabrication of anything other than sloping sidewalls. Consequently, the fabrication of deep etched features, with walls that are accurately orthogonal to the substrate (as might be required in a waveguide mirror), is extremely difficult.

**DESIGN EXAMPLE**

We may estimate the maximum groove depth achievable in GaAs by argon ion milling through a mask layer, as follows. Etch rates for a variety of materials milled with Ar⁺ ions at normal incidence, and 1 keV beam energy, 1 mA/cm² beam current are given in Table 13.6-2. We might select photoresist as the masking material, spin-coated to a thickness of 0.3 µm. With the etch rate given, the mask layer protecting areas of GaAs that are not to be milled will erode completely in 0.3 / 600 x 10⁻⁴ = 5 mins. In this time, exposed GaAs areas will be etched to a depth of 5 x 2,600 x 10⁻⁴ = 1.3 µm.

<table>
<thead>
<tr>
<th>Material</th>
<th>Etch rate (Å/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>360</td>
</tr>
<tr>
<td>SiO₂</td>
<td>420</td>
</tr>
<tr>
<td>GaAs</td>
<td>2,600</td>
</tr>
<tr>
<td>Cr</td>
<td>200</td>
</tr>
<tr>
<td>Ti</td>
<td>100</td>
</tr>
<tr>
<td>Al</td>
<td>440</td>
</tr>
<tr>
<td>Shipley AZ1350 photoresist</td>
<td>600</td>
</tr>
</tbody>
</table>

Table 13.6-2  Etch rates for different materials by Ar⁺ ions.

**MASKLESS ETCHING**

One possible solution to the problem of etch selectivity is to avoid the use of a mask altogether. This is the rationale behind **focussed ion beam** (FIB) micromachining, which achieves local sputtering simply by confining the ion beam to a small area. Figure 13.6-3 shows the required apparatus, which is effectively a modified scanning ion microscope. Here, the source of the ions is not a plasma, but a high-brightness, near-point emitter known as a **liquid metal ion source** (LMIS). The ion source is a capillary tube, filled with a metal of low melting point (e.g. In or Ga), surrounded by a resistive heater. A small aperture, raised to a negative potential, is placed just beyond the tip of the capillary. On heating, the metal is liquified, and the molten surface at the very end of the capillary is then able to respond to external forces. Due to the large electric field in the neighbourhood of the aperture, this surface is distorted electrostatically into a shape known as a **Taylor cone**. At the apex of the cone, the field is strong enough to overcome surface forces and extract large
numbers of ions, so that the tip of the cone acts as a near-point ion source. After emission, the ions pass through the aperture as an expanding beam, which may be refocussed by electrostatic lenses to a point further down the column. Any portion of the substrate placed here will be subjected to an intense, highly directed flux of ions, which will eject atoms from the surface by sputtering. The charging of insulating substrates may be prevented, by machining through a thin metal layer, grounded to leak any charge away.

By confining the area of the substrate that is exposed to the ions to a small spot, the sputtering is localised. The beam may also be deflected and blanked under computer control. Consequently, it is possible to program the area that is machined, so that a pattern is etched into the substrate. This may be aligned to other features on the substrate prior to etching, by using the equipment in an alternative ion microscope mode. In this case, a smaller ion flux is used, and the beam is raster scanned across the entire substrate.

Variations in the numbers of secondary electrons (also generated by the sputtering process) emitted during the scan are then used to build up a picture of the surface, so that the substrate can be moved to a desired co-ordinate origin by an x - y stage. The substrate is also eroded in this alignment stage, albeit to a very small extent. Normally, the machining is performed in a number of passes, using a raster scan of the addressed area. Due to the absence of an erodable mask, the etched features may be extremely deep (tens of microns), with near-vertical walls.

Focused ion beam micromachining may be used to cut very deep features, and has been used a number of niche applications in optoelectronics, such as mirror fabrication. However, it suffers from two major defects. Firstly, it is a serial process, which limits throughput. Consequently, it is inappropriate for the definition of, say, ridge guides, where most of the
substrate area must be etched. Secondly, it shares the disadvantage of other physical processes, that there is no distinction between the substrate and the sputtered material (which may be redeposited elsewhere on the surface).

**REACTIVE ETCHING**

Both of these problems are avoided by **reactive ion etching** (RIE), a combined physical/chemical process requiring a mask layer for local feature definition. It is similar to sputter etching, but the inert gas is now replaced by a molecular gas, whose decomposition by the plasma yields by-products that react with the substrate material to form volatile compounds at low temperatures. The gases that have been used to etch a number of important materials are detailed in Table 13.6-3.

<table>
<thead>
<tr>
<th>Material</th>
<th>Etch system</th>
</tr>
</thead>
<tbody>
<tr>
<td>GaAs</td>
<td>CCl$_2$F$_2$</td>
</tr>
<tr>
<td>InP</td>
<td>CCl$_2$F$_2$ / Ar / O$_2$</td>
</tr>
<tr>
<td>LiNbO$_3$</td>
<td>CCl$_2$F$_2$ or CF$_4$</td>
</tr>
</tbody>
</table>

Table 13.6-3 Dry etch mixtures for a number of crystalline materials used in optoelectronics.

An analogous process based on ion milling with a reactive gas also exists; this is known as **reactive ion beam etching** (RIBE). In each case, the use of a mask restores the parallelism of the process, making the etching of large areas feasible, but selectivity between the mask and the substrate is introduced by tailoring the etchant to attack mainly the latter. Etching is also normally highly anisotropic, especially with crystalline substrates; this virtually eliminates mask erosion and undercut, allowing deep features to be etched with ease. A final advantage is that the etched material may be pumped away, avoiding redeposition, although the pumping system must be specially protected against the etch gases.

**13.7 LITHOGRAPHY**

The process of pattern transfer is known as **lithography**. This is normally a two-step process; in the first, a **mask** - normally a glass plate, containing an opaque pattern in chromium metal - is made. In the second, the features on this master are copied onto the substrate by a form of printing.

**MASK FABRICATION**

We begin by considering mask fabrication using **electron beam lithography**. Initially, the design dimensions of the device are specified, and its salient features are assigned to one of a number of mask levels. For example, Figure 13.7-1 shows the plan view of a directional coupler. To fabricate the device using the Ti:LiNbO$_3$ process, two lithographic steps will be needed; patterning of a layer of Ti metal to define the guides prior to diffusion, and patterning of a layer of Al metal to outline the electrodes. All waveguide features - generally in the form of thin, curving tracks - are therefore assigned to Level 1 (unshaded), while the rectangular electrode blocks form Level 2 (shaded). A separate mask will be made for each level, and often each will contain additional alignment features so that the two patterns may be registered together with high accuracy.
Computer-aided design (CAD) tools are then used to specify the pattern still further. As a result of direct interaction with the designer, a set of special purpose code is generated. This is written in a high-level computer language, and each line typically contains a description of the vertices of an elementary shape, corresponding to the whole or part of one feature in the design. The code can be interpreted by plotting packages, allowing the pattern to be verified on a computer screen. When the design is finalised, a compiler is used to generate a new set of low-level code capable of driving the mask-making machine directly.

Figure 13.7-1 Mask levels for a directional coupler switch in the Ti:LiNbO₃ process.

One problem for integrated optics is that while the features contained in waveguide devices are normally simple, some of these (curved guide sections) are inappropriate for conventional CAD systems, which have historically been developed to support integrated electronics, and are therefore based around rectangular shapes, tracks or polygons. Consequently, using high-level code, a typical curved guide is normally approximated by subdivision into a set of polygonal shapes. It is then further approximated by the low-level code as a pixellated pattern, each element of which is an elementary square corresponding to the minimum feature size of the mask-making system.

Figure 13.7-2 illustrates this gradual degradation. Figure 13.7-2a shows the desired pattern - a section of curved guide. In Figure 13.7-2b, this is approximated as a track, formed by concatenating a number of basic parallelogram shapes. Finally, Figure 13.7-2c shows the result of breaking the shapes into individual pixels.

Figure 13.7-2 a) Desired pattern, b) representation using polygons, and c) representation using pixels.
Clearly, some distortion of the original pattern has been introduced, together with considerable edge roughness. The latter can be major problem, since waveguide bends are typically shallow - the schematic of Figure 13.7-1 is artificially compressed in the horizontal direction - and the result can be high scattering loss, unless the pixel size is made small.

The low-level code is then used to write the masks directly, by computed-controlled deflection of an electron beam over a sensitised plate under high vacuum. The equipment required is basically a modified scanning electron microscope, fitted with additional beam-deflection and beam-blanking electronics and a precision motorized positioning table.

Figure 13.7-3 shows a typical electron-beam mask-making machine. The source of electrons - known as an electron gun - is a thermionic emitter located at the upper end of the evacuated column on the left. Its output is focussed by a set of fixed electrostatic lenses to a spot (with a diameter of 0.5 - 0.01 µm) in the writing chamber below the column. The sensitized plate is placed on an x - y positioning table at the focal plane, and writing is carried out by a combination of beam deflection and table motion.

Figure 13.7-3  An electron-beam mask-making machine (photograph courtesy R.A.Lawes, SERC Rutherford Appleton Laboratory).

By deflecting the beam electrostatically, a field of about 0.1 mm² may be addressed, without moving the table. The field is broken up into a number of pixels on an x - y grid; typically, each is of the order of 0.1 µm x 0.1µm and corresponds to one pixel generated by the low-level code. Shapes that pass outside the field boundaries are then composed of a number of fields 'stitched' together, with the entire plate being moved by the motorized table to each new field position. Interferometric position sensors are used to ensure that the stitching is accurate. Now, a modern electron-beam mask-making machine writes at ≈ 10^7 pixels per second. Given that each is about 10⁻¹⁴ m² in area, about 10⁻⁷ m² can be written per second. Since a moderate-sized mask is about 10 cm x 10 cm, the maximum required writing time is of the order of 10⁵ seconds, or a few hours. However, in practice, this figure is greatly
reduced, since only those areas that contain features (a small percentage of the total in waveguide devices) need be exposed to the electron beam.

Two different methods are used in the actual writing of the pattern. Both are based on organic polymers known as electron resists. In negative resist, cross-linking is induced by irradiation with electrons, making the irradiated areas insoluble. Un-irradiated areas are then dissolved away during a later development step. In positive resist, on the other hand, cross-links are broken by electron irradiation, making the exposed areas soluble. Each type therefore forms a relief image after exposure and processing. The most common negative resist is poly (glycidylmethacrylate-co-ethyl acrylate) or COP, while its positive counterpart is polymethyl-methacrylate, or PMMA, developed using methyl isobutyl ketone, or MIBK. PMMA is considerably less sensitive than COP, but offers higher resolution. However, one of the main factors limiting resolution in electron beam lithography is back-scatter of electrons that have passed through the resist into the substrate.

Further processing may then be carried out to transfer the written features to a chromium metal layer. Additive or subtractive methods may be used; for example, Figure 13.7-4a shows a subtractive process for chrome-on-glass masks, using negative electron-beam resist.

![Figure 13.7-4](image)

Figure 13.7-4  Mask fabrication using negative resist and (a) subtractive, (b) additive processing.

The mask plate - a parallel-sided, highly polished glass substrate - is first coated with Cr metal (e.g. by evaporation) to a thickness of $\approx 1000 \, \text{Å}$ (1). A layer of negative resist is then added; typically, this is spin-coated in liquid form, and then baked to drive off any solvent (2). Next, it is exposed by the scanned electron beam (3), and developed to remove resist from the unexposed areas (4). The remaining resist is then used as a mask in an etching step, in which any unprotected areas of chrome are dissolved away (5). Finally, the resist is also removed, leaving a metal pattern on the glass (6).

Figure 13.7-5 shows a typical set of curved S-bend guide patterns in chromium metal, on a mask fabricated using the processes described above. The guide widths are 7 µm.
Figure 13.7-5 S-bend pattern on a chromium metal mask.

Figure 13.7-4b shows the corresponding additive process, which requires one less step. To begin with, the plate is coated only in resist (1). It is then exposed to the electron beam, but this time in different areas (2). After development, the areas of glass that are to be metallized are therefore unprotected (3). Next, the whole plate is coated with chromium (4). The remaining resist is then dissolved away, simultaneously removing the unwanted metal, which simply breaks up in the resist solvent. This is known as **lift-off**; it offers some advantages in the fabrication of fine features, but special processing is needed to give the edge of the resist an undercut profile, so there is a clean break of the metal layer during lift-off.

**PATTERN TRANSFER**

In the second stage, the pattern is transferred from the mask to the substrate, using an optical printing step known as **photolithography**. It is based on the properties of light-sensitive organic materials or **photoresists**, whose behaviour is qualitatively similar to the electron resists described above. Negative and positive optical resists both exist; in the former, exposure to ultraviolet light polymerises the material, and in the latter, cross-links are broken by exposure. Figure 13.7-6 shows a typical process for subtractive pattern transfer using positive resist, with the mask of Figure 13.7-4.

![Photolithography process diagram](image)

Figure 13.7-6 Optical photolithography: a typical subtractive pattern transfer process.
Here, it is assumed that the substrate has already been coated with the film layer to be patterned, together with a layer of photoresist. Exposure is carried out in a machine known as a mask aligner; this contains accurate translation stages, a high-power microscope and a U.V. exposure source. First, the relative positions of mask and substrate are carefully adjusted, so that the mask pattern is aligned with existing features on the substrate. The two are then clamped together by evacuating the space between them, and the resist is exposed by a flood beam (1). After removal from the mask aligner, the resist is developed; negative resists generally require organic solvents, while alkaline aqueous solutions are used for positive resist (2). The result is a pattern of resist on the film layer; this may be transferred to the film itself by etching (3), followed by removal of the remaining photoresist (4).

GRATING FABRICATION

The two-step procedure described above suffices for most patterns in integrated optics. However, the resolution of the copy process is limited by diffraction to approximately the wavelength of the incoherent ultraviolet exposure source. This is inadequate for making reflection gratings, especially in semiconductor materials. For these, single-step processes are generally employed. There are two common techniques. In the first, the grating pattern is written directly by e-beam lithography onto electron resist, which is coated on the substrate itself. After development of the resist, the pattern may be transferred to the substrate by (for example) preferential wet etching. Figure 13.7-7 shows a first-order grating for a DFB laser (period $\approx 0.23 \, \mu m$) etched into InGaAsP in this way.

![Figure 13.7-7](image)

Figure 13.7-7  A first-order DFB laser grating on InGaAsP (photograph courtesy C.J.Armistead, S.T.C.Technology Ltd.).

In the second, an ultraviolet laser is used as a maskless exposure source for photoresist. Now, we have previously seen in Chapter 5 that the local distribution of irradiance resulting from two coherent plane waves that intersect at an angle is a periodic fringe pattern. This may be used to record a grating pattern directly in photoresist, in a technique known as holographic exposure. Figure 13.7-8 shows a particularly simple set-up, which requires only a single recording beam. This uses a prism, and the geometry is arranged so that the second beam necessary for the exposure is generated by total internal reflection of the input wave at the end-face of the prism. Both beams then pass directly into the photoresist layer...
without reflection, since the prism is contacted onto the resist-coated substrate using index-matching liquid.

![Figure 13.7-8 Simple experimental configuration for grating fabrication by holographic exposure.]

**DESIGN EXAMPLE**

We can estimate the parameters of the holographic set-up needed to record a second-order grating for a GaAs/GaAlAs DFB laser, as follows. Assuming that the laser wavelength is 850 nm, and that \( n \approx 3.6 \) for GaAs, the propagation constant of the laser mode will be

\[
\beta = \frac{2\pi}{3.6} / 0.85 \times 10^{-6} = 26.61 \times 10^6 \text{ m}^{-1}.
\]

For Bragg reflection from a first-order grating, a grating vector of magnitude \( 2\beta \) is needed. However, with a second-order grating, the spatial frequency is halved, so here we require \( K = \beta \). This corresponds to a grating period of \( \Lambda = 2\pi/K = 0.236 \text{ µm} \). We shall assume that a He-Cd laser (\( \lambda_0 = 0.442 \text{ µm} \)) will be used for recording, and that the two recording beams are generated inside a glass prism (for which \( n \approx 1.5 \)) and intersect symmetrically with an interbeam angle of \( 2\theta \). In this case, the propagation constant inside the glass is

\[
\beta_r = \frac{2\pi}{1.5} / 0.442 \times 10^{-6} = 21.3 \text{ m}^{-1}.
\]

and the spatial frequency generated by the recording is \( K = 2\beta_r \sin(\theta) \). To obtain the value of \( K \) above, the recording angle should be \( \theta = \sin^{-1}(2\beta_r / 2 \times 21.3) = 38.65^\circ \). This is sufficiently close to 45° for the set-up shown in Figure 13.7-8 to be used.

**13.8 OPTICAL FIBRE FABRICATION**

We now turn our attention to the methods of optical fibre fabrication. Broadly speaking, these may be divided into two groups: those which do not require a preform, and those which do. However, processes based on preforms allow so much more control over the characteristics of the fibre that they are now entirely dominant. We will therefore limit our discussion of preformless methods to a brief description of one historically interesting example, the double crucible method.

As its name suggests, this involves the use of two concentric platinum crucibles, placed one inside and slightly above the other and heated by external windings (Figure 13.8-1). The inner crucible is loaded with the core glass (supplied in the form of rods), the outer one with the cladding glass. When the assembly is raised to a sufficient temperature, both glasses melt. The core material is then allowed to flow through a nozzle at the base of the inner crucible to mingle with the cladding material, which can flow through a larger nozzle at the bottom of the outer crucible. Because of the concentricity of the arrangement, little mixing
of the two glasses occurs; the flow stream of the core material is cylindrical, while that of the cladding is annular. The result is that the core glass emerges inside the cladding, in the geometry of a step-index fibre. After leaving the heated region, the glass quickly solidifies, and can be coated in a protective thermosetting plastic jacket and wound on a drum. Correct choice of nozzle dimensions and pulling speed then yields a fibre of the desired core and cladding diameters.

Figure 13.8-1 the double crucible method for fibre pulling directly from melts.

Although simple, the double crucible method suffers from the disadvantage that the refractive index profile obtained is entirely dictated by the process. If little interdiffusion of the two glasses occurs, the profile is almost exactly a step-index one, but the diffusion of a mobile, highly polarizable dopant (e.g. thallium) out of the core may be used as a mechanism for obtaining graded-index fibres with near-parabolic profiles. Fibre bundles (consisting of a number of parallel fibre cores) may even be drawn, using a multiple nozzle arrangement. However, it is not possible to exercise the degree of control required to fabricate low-dispersion or polarization-preserving fibres, both of which require complicated and well-specified index profiles. This can only be achieved by using a preform.

PREFORM FABRICATION

A preform is, in effect, a large-scale model of the fibre cross-section, which is subsequently reduced to the final diameter by drawing. As an example of an early method of making step-index preforms, we mention the rod-in-tube method. Here the preform was simply fabricated from a rod of core glass, fused inside an annular cylinder of cladding glass. Although direct, and adaptable to more complicated geometries (e.g. non-circular cross-sections), the resulting fibre suffered severely from scattering loss, caused by scratches and flaws at the interfaces between the glasses.

Figure 13.8-2 shows the alternative built-in casting method, used to fabricate preforms in multicomponent fluoride glasses (e.g. a ZrF$_4$-BaF$_2$-GdF$_3$ mixture). In Figure 13.8-2a, molten cladding glass is poured into a cylindrical brass mould, heated to around the glass transition temperature. The mould is then upset (Figure 13.8-2b), so that the central portion of the casting, which is still liquid, runs out, and the core glass is then poured into the resulting cylindrical hollow (Figure 13.8-2c). After cooling, the casting is removed from the brass mould, and its ends are trimmed to yield a cylindrical preform of step-index profile (Figure 13.8-2d).
Neither of these techniques allows the construction of graded-index preforms of arbitrary refractive index profile. To overcome this problem, a range of entirely different fabrication methods, based on chemical vapour deposition were developed. There are three main variants. The first is the **outside vapour phase oxidation** (OVPO) or **soot** process. This was developed by Corning Glass works in 1970, and used to fabricate the first fibre with losses as low as 20 dB/km. In this case, the preform is built up by depositing layers of silica glass on a ceramic mandrel or 'bait'. The silica is produced by hydrolysing a mixture of SiCl₄ and O₂ in a glassburner, following the reaction:

$$\text{SiCl}_4 + 2\text{H}_2\text{O} \rightarrow \text{SiO}_2 + 4\text{HCl}$$  \hspace{2cm} \text{(13.8-1)}$$

The result is a stream of small silica particles, or 'soot', which are blown onto the bait by the oxy-hydrogen flame, where they stick. During deposition, the bait is rotated in a glassblowing lathe, and the soot stream is traversed up and down the axis to ensure uniform coverage. In early experiments, titania was used as a dopant to raise the refractive index of the inner glass layers, which form the fibre core. However, this caused high propagation loss (due to the presence of Ti³⁺ ions), necessitating an additional annealing step. Considerably better results were obtained when germania was used instead, deposited following the similar reaction:

$$\text{GeCl}_4 + 2\text{H}_2\text{O} \rightarrow \text{GeO}_2 + 4\text{HCl}$$  \hspace{2cm} \text{(13.8-2)}$$

Because the composition of each layer may be varied merely by adjusting the flow of the relevant gas, graded-index preforms of arbitrary radial profile may be built up by OVPO. After the deposition of sufficient material, the bait is removed, and the rather porous soot is compacted or 'sintered' in a furnace at \(\approx 1500^\circ\text{C}\) to produce glass of optical quality. Chlorine gas is used as a drying agent during sintering; the effect is to reduce the concentration of hydroxyl (OH⁻) ions, which would otherwise greatly increase propagation loss in the final fibre.

The OVPO process suffers from two main disadvantages. The first is that the removal of the bait from the preform is difficult, and may cause cracking. It also leaves a large hole in centre of the preform; this is closed up during the fibre drawing process, but a central depression in the refractive index profile normally remains. The second is that the process is a batch one, which limits its applicability to the production of fibre in very long lengths.

The second process variant is the **modified chemical vapour deposition** (MCVD) process, developed by Bell Telephone Laboratories and others in 1974. This avoids one of the difficulties above, the requirement for a central bait. In effect, the process is one of inside
vapour phase oxidation (IVPO), with deposition now taking place inside a silica tube (Figure 13.8-3). One major difference from OVPO is that the chemical reaction does not take place in a flame. Instead, unreacted gases are mixed together and injected into one end of the tube. The tube is heated externally to 1500 - 1650°C by a traversing oxyhydrogen burner, and the reaction is induced at the growing surface by contact with the heated region. Since this process takes place in an enclosure, an improvement in purity of the deposited material might be expected over that obtained by OVPO.

![Figure 13.8-3 Improved CVD system with control of the internal pressure of the supporting tube (after M.Okada et al., Elect. Lett. 14, 89-90 (1978))](image)

Once again, the preform is built up slowly in layers, with changes in refractive index obtained by fine control of the gas composition. This time, however, it grows inwards, and the soot is simultaneously sintered by the oxyhydrogen flame. In early processes, there was a tendency for the silica tube to shrink at high temperatures as the deposition progressed, due to surface tension. This resulted in an increase in the thickness of the tube, which made it difficult to maintain the correct internal temperature. The effect was suppressed using slight positive pressure inside the tube, under feedback control from a laser-based measurement of the tube diameter.

Originally, the deposition reaction was based on the oxidation of hydrides (e.g. SiH$_4$) in an oxygen gas stream, much as in the planar CVD techniques described earlier. However, slow deposition rates, (coupled with relatively high contamination by hydroxyl ions) induced a switch to alternative chloride-based processes. For example, the two materials most commonly required, silica and germania, may be deposited from SiCl$_4$ and GeCl$_4$, respectively, following:

\[
\begin{align*}
\text{SiCl}_4 + \text{O}_2 & \rightarrow \text{SiO}_2 + 2\text{Cl}_2 \\
\text{GeCl}_4 + \text{O}_2 & \rightarrow \text{GeO}_2 + 2\text{Cl}_2
\end{align*}
\]

13.8-3

This particular MCVD process has allowed the fabrication of graded-index fibres with extremely low loss (0.2 dB/km at 1.55 µm wavelength). A further process variant exists, in which the heating is performed by a microwave plasma inside the tube instead of an external
flame; this is known as **plasma chemical vapour deposition** (PCVD) and yields similar results.

One advantage of MCVD is the ease with which more complicated preforms may be made. For example, Figure 13.8-4 shows the process for making **bow-tie** high-birefringence fibres (previously described in Chapter 8), developed at Southampton University.

![Fabrication of bow-tie fibre](image)

Figure 13.8-4 Fabrication of bow-tie fibre (after R.D. Birch et al., Elect. Lett. 18, 1036-1038 (1982))

These require the insertion of stress-producing sectors into the preform, so the fibre is not circularly symmetric. Figure 13.8-4a shows the first step, which involves the deposition of a uniform layer of fluorophosphorus silica glass inside a silica tube, followed by a similar layer of borosilicate glass (which will supply the stress-inducing sectors). In Figure 13.8-4b, the rotation of the lathe is stopped, and two traversing burners are used to provide localised hot zones on either side of the tube. The borosilicate layer is then etched away from two diametrically opposed strips by passing a fluorine-liberating gas (SF₆ mixed with N₂) down the tube. In Figure 13.8-4c, two more uniform layers (of fluorophosphorus silica glass, which acts as a buffer, and germania-doped silica, which provides the core) are then deposited. Finally, in Figure 13.8-4d, the tube is collapsed, to yield the characteristic bow-tie geometry.

Despite this flexibility, MCVD still suffers from being inherently a batch process. So far, the only method suitable for the continuous preform fabrication is the **vapour-phase axial deposition** (VAD) process, developed by NTT Ibaraki Laboratories. Figure 13.8-5 shows a schematic of a VAD preform plant.

Once again, the process involves the production of soot from chlorides, which are hydrolysed in a flame exactly as in the OVPO process. The reactions followed in the deposition of silica and germania are therefore as in Equations 13.8-1 and 13.8-2, respectively. The main difference is that the preform is grown axially, starting from one end. A silica rod held in a rotating chuck is used as a bait, and soot is blown onto the end of this rod by a number of fixed torches. As the deposition proceeds, the preform is slowly raised, to keep the growing surface at a fixed height.
The porous soot preform is consolidated into a transparent glassy mass by passing it directly from the growth chamber into a sintering furnace (at 1500 - 1600°C), where it is dried in a chlorine or thionyl chloride (SOCl₂) atmosphere to reduce hydroxyl ion contamination. In this stage, isolated OH⁻ ions are replaced by Cl⁻ ions, generating Si-Cl bonds; however, these do not affect the fibre propagation loss, since their fundamental absorption peak lies at around $\lambda_0 = 25 \mu m$, far beyond the operating wavelength. A typical dehydration reaction follows:

$$Si-OH + SOCl_2 \rightarrow SO-Cl + SO_2 + HCl$$

Figure 13.8-5 shows a multiple-torch plant, where a set of three fixed torches is used to fabricate a germanium-doped silica core surrounded by a thick silica cladding layer in a single deposition run. By careful design of the torch nozzles, the spatial distribution of the individual gas flows may be controlled. This allows the local dopant concentration, deposition rate and surface temperature to be specified accurately, so that graded-index guides with the required profile may be fabricated.

Table 13.8-1 is a comparison of the methods described above - the OVPO, MCVD, PCVD and VAD processes. As can be seen, all can be used to make very low-loss optical fibre in long lengths, but vapour-phase axial deposition probably allows the most rapid construction of the largest preforms.
<table>
<thead>
<tr>
<th></th>
<th>OVPO</th>
<th>MCVD</th>
<th>PCVD</th>
<th>VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical reaction</td>
<td>Hydrolysis</td>
<td>Oxidation</td>
<td>Oxidation</td>
<td>Hydrolysis</td>
</tr>
<tr>
<td>Dehydration</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Heat source</td>
<td>Oxyhydrogen flame</td>
<td>Oxyhydrogen flame</td>
<td>Microwave plasma</td>
<td>Oxyhydrogen flame</td>
</tr>
<tr>
<td>Deposition rate (g/min)</td>
<td>9.0</td>
<td>0.5-1.0</td>
<td>0.5</td>
<td>5-10</td>
</tr>
<tr>
<td>Production maximum</td>
<td>&gt; 10</td>
<td>2</td>
<td>2.5</td>
<td>&gt;10</td>
</tr>
<tr>
<td>Profile control</td>
<td>Easy</td>
<td>Easy</td>
<td>Very easy</td>
<td>Easy for SM</td>
</tr>
<tr>
<td>Typical fibre length (km)</td>
<td>90</td>
<td>15-20</td>
<td>5-10</td>
<td>100</td>
</tr>
<tr>
<td>Loss@1.3/1.55μm(dB/km)</td>
<td>0.35 / 0.2</td>
<td>0.35 / 0.2</td>
<td>0.35 / 0.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 13.8-1 Comparison of fibre preform fabrication processes (after H. Murata "The development of optical fibres in Japan", Gordon and Breach Science Publishers (1989)).

FIBRE PULLING

After consolidation and drying, the preform is inserted into a pulling tower for the final drawing stage, shown in Figure 13.8-6. The preform is fed at constant speed into the electrically heated carbon furnace at the top of the tower, which is maintained at a precisely-controlled temperature (≈ 2000°C) under a clean atmosphere.

![Figure 13.8-6 Fabrication of optical fibre in a fibre pulling tower.](attachment:Figure13_8_6.png)
As the end of the preform melts, it is drawn out of the furnace as a thin liquid stream, which solidifies rapidly on cooling. For single-mode fibre, the outer diameter is typically 125 µm. At this point, the fibre is extremely fragile, and contact with any surface can induce flaws and scratches, which will subsequently cause mechanical failure. It is therefore passed directly into a coating die; this applies a protective plastic jacket, typically in the form of a 100 µm-thick layer of silicone resin, which is hardened in a curing oven further downstream. In order to achieve high coating speeds, the coater is pressurised by a high-pressure gas source; without this, the fibre has been found to slip through the resin without being coated. The fibre is pulled by a capstan, whose rotational speed is controlled in a feedback loop by a non-contact outer-diameter monitor situated just outside the furnace, and then wound onto a take-up spool.
PROBLEMS

13.1 Why are binary compounds used as substrate materials for III - V optoelectronic devices, rather than ternary or quaternary compounds? Distinguish between homoepitaxy and heteroepitaxy. In what optoelectronic applications might the latter growth technique be required? What are the advantages of MOVPE over LPE for the overgrowth of In$_{1-x}$Ga$_x$As$_{1-y}$P$_y$?

13.2 Assuming that the lattice parameter $a(x, y)$ of the quaternary alloy In$_{1-x}$Ga$_x$As$_{1-y}$P$_y$ may be described by the function $a(x, y) = xy a_{GaP} + x(1 - y) a_{GaAs} + (1 - x)y a_{InP} + (1 - x)(1 - y) a_{InAs}$ (where $a_{GaP} = 5.4512 \text{ Å}$, $a_{GaAs} = 5.6536 \text{ Å}$, $a_{InP} = 5.8696 \text{ Å}$, and $a_{InAs} = 6.0590 \text{ Å}$ are the lattice parameters of the binary compounds GaP, GaAs, InP and InAs, respectively), show that the locus of compounds lattice matched to InP is almost exactly a straight line, obeying the relation $y = 1 - 2.14x$. What is the stochiometry of lattice-matched indium gallium arsenide?

13.3 Show by direct substitution that the concentration variation $C(x, t) = \{S/\sqrt{\pi Dt}\} \exp(-x^2/4Dt)$ is a solution of the one-dimensional diffusion equation $\partial C/\partial t = D \partial^2 C/\partial x^2$ for $t > 0$. Show also that the constant $S$ is given by $S = \int C(x, t) \, dx$. What does $S$ represent?

13.4 A particular doping process requires the diffusion of an impurity into silicon for 8 hours. Assuming that the process activation energy is 3 eV, and that the diffusion coefficient tends to $10^{-1} \text{ cm}^2 \text{ s}^{-1}$ as $T \to \infty$, calculate the diffusion length for a) $T = 900 \degree \text{C}$ and b) $T = 1000 \degree \text{C}$. [0.39 $\mu$m; 1.25 $\mu$m]

13.5 i) Show that the concentration variation $C(x, y, t) = \{S/4\pi Dt\} \exp(-(x^2 + y^2)/4Dt)$ represents a solution of the two-dimensional diffusion equation $\partial C/\partial t = D \nabla^2 C$, for $t > 0$. What boundary conditions would this solution correspond to?

ii) A channel waveguide is to be formed by indiffusion of a strip of Ti metal into a LiNbO$_3$ crystal. Assuming that the strip occupies the range $-w/2 \leq y \leq w/2$ on the crystal surface prior to diffusion, show that the resulting refractive index distribution is given by:

$$n(x, y) = n_s + \Delta n \exp(-x^2/L_{D}^2) \left[ \text{erf}\{(w + 2y)/2L_{D}\} + \text{erf}\{(w - 2y)/2L_{D}\} \right]$$

where $n_s$ is the substrate index prior to diffusion, $x$ is the depth below the crystal surface, $L_{D} = 2\sqrt{Dt}$ and $\Delta n$ is a constant. Sketch the variation in refractive index along lines of a) constant $x$ and b) constant $y$.

13.6 A single-mode fibre is to be clamped between two silicon wafers, which carry V-shaped alignment grooves. <100> Si is to be used, etched anisotropically down the <111> crystal planes, which are oriented at 54.74° to the wafer surface. Assuming that the fibre has a diameter of 125 $\mu$m, that 5 $\mu$m clearance is required between the wafers, and that the etch selectivity is perfect, what should be the groove width? What will be the groove depth? [149.55 $\mu$m; 105.76 $\mu$m]

13.7 The following lines of code are a short section of a CAD design for the mask set used to fabricate an integrated optical device, written in the now-obsolete GAELIC design
language. The first line sets the design grid to 0.1 µm intervals on each axis, while the last line terminates execution. Each of the lines \( \text{RECT}(mask)x,y:\Delta x,\Delta y \); generates a filled rectangular shape on level \( mask \) of the mask-set. The bottom left-hand corner of the rectangle is located at the point \( (x, y) \), while the sides of the rectangle are of length \( \Delta x \) and \( \Delta y \) in the \( x \)- and \( y \)-directions. Sketch the device, indicating the main design parameters. Assuming that the device is to be fabricated by the Ti:LiNbO\(_3\) process, what crystal orientation should be used?

\[
\text{UNITS=MICRONS,GRID=0.1;} \quad \text{RECT(1)-25000,35:50000,70;} \\
\text{RECT(1)-25000,-105:50000,70;} \quad \text{RECT(2)-25000,35:50000,500;} \\
\text{RECT(2)-25000,-535:50000,500;} \quad \text{FINISH;} \\
\]

### 13.8

i) The device design in Question 13.7 is to be completed by the addition of symmetric, circular-arc S-bend sections at either end, so that the centre-to-centre separation of the guides at the input and output is increased to 300 µm, for connection to optical fibres. Assuming that the minimum bend radius that gives acceptable loss in the Ti:LiNbO\(_3\) process used is 30 mm, estimate the minimum length of each transition.

ii) The GAELIC command \( \text{POLY}(mask)L,x,y:\Delta x_1,\Delta y_1,\Delta x_2,\Delta y_2, \ldots, \Delta x_n,\Delta y_n \); generates an \( n \)-sided polygonal shape on level \( mask \) of the mask-set. The point \( (x, y) \) specifies the first vertex of the polygon, while the increments \( (\Delta x_i, \Delta y_i) \) define the position of the \( i+1 \)th vertex relative to the \( i \)th vertex. The vertex positions are listed in anticlockwise order along the command line. By breaking the curved waveguide sections into a number of elementary polygonal shapes, use the POLY command to add S-bends to the design of Question 13.7.

[4.14 mm]

### 13.9

The GAELIC code of Question 13.7 is to be used as input to an electron beam mask-making machine. Assuming that negative electron resist and subtractive processing are used to fabricate the resulting chrome-on-glass mask set, outline the process steps by which the complete Ti:LiNbO\(_3\) device might be fabricated. You should assume that the waveguide patterns are to be defined by wet etching, and the electrodes by lift-off.

### 13.10

i) A fibre preform is to be fabricated by MCVD inside a silica tube of length 1 m and inside diameter 20 mm. Assuming that the material is mainly pure silica (density 2.6 g/cm\(^3\)) deposited at a rate of 2 g/min, estimate the mass of the preform and the time taken to form it.

ii) The preform is to be pulled down to fibre in a single step. Assuming that the preform feed and fibre pulling rates are 3.5 mm/min and 1.5 m/s respectively, calculate the length and diameter of the resulting fibre.

[0.817 kg; 6.81 hours; 25.6 km; 125 µm]
SUGGESTIONS FOR FURTHER READING

Beales K.J., et al. "Low loss graded index fiber by the double crucible technique" Proc. 5th European Conf. on Optical Fiber Communication, Amsterdam, Paper 3.2 (1979)
Chartier G.H., Jaussaud P., de Olivera A.D., Parriaux O. "Optical waveguides fabricated by electric-field controlled ion exchange in glass" Elect. Lett. 14, 132-134 (1978)
Flavin P.G. "Fabrication of curved structures by electron-beam lithography" Elect. Lett. 18, 865-867 (1982)
Nahory R.E., Pollack M.A., Johnston W.D. "Band gap versus composition and demonstration of Vegard's law for InₓAₓGaₙ₋ₓAsₙ₋ₙₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚₚ}_{47}
CHAPTER FOURTEEN

SYSTEMS AND APPLICATIONS

14.1 INTRODUCTION

By now, we have encountered a wide range of guided wave components, and it is time to show how a number of these can be combined together in a complete optical circuit. We have chosen to illustrate the process of integration with concrete examples from the literature. Some of these are old applications, which have unfortunately not stood the test of time. Either the performance of the system turned out to be too poor to be useable in the end, or it was simply apparent that there were better ways of solving the problem that did not involve optics at all. Others are so new that they have yet to reveal their limitations, or indeed their potential. Only a small subset can be described as 'successful'. We make no apology for this: new technology obeys the law of the jungle, just like everything else, and there will naturally be few eventual winners. Our criteria for selecting the examples were that they illustrate the process of integration itself, rather than the judgement of history.

14.2 THE PLANAR INTEGRATED OPTIC RF SPECTRUM ANALYSER

We begin with applications that use one single waveguiding system on its own. The first such example is the planar integrated optic radio-frequency spectrum analyser. The function of this particular circuit is to perform a real-time, parallel, spectral analysis of RF signals. Such a package might form a key component in an aircraft-based electronic countermeasures system, used to monitor enemy radio or radar signals. Its operation requires a Fourier transform of the incoming radio signal to be carried out. Although electronics could be used, optics immediately suggests itself because of the ability of a lens to perform a parallel Fourier transform on a spatial distribution of light, as mentioned in Chapter 4. Having chosen an optical implementation, the advantages of integration are lightweight, and a small, rugged package that will be immune to misalignment through shocks and vibration (e.g., during takeoff and landing).

Figure 14.2-1 shows one possible realisation of the circuit, which consists of a planar integrated optic chip, containing a number of the components described in Chapter 7. This particular example is conceptual, being an amalgam of several real prototypes demonstrated using different technologies. It is based on a waveguide fabricated on a piezoelectric, acousto-optic substrate. In this case, we may assume that LiNbO3 has been used, and that the guide has been formed by Ti indiffusion; however, other systems have used silicon as the substrate material.

The circuit works as follows. First, a prism coupler is used to launch light directly from an external laser (typically, a HeNe laser) into the guide. This results in a relatively narrow beam of guided light, which is then passed to a corrugated Bragg grating; this has a rectangular boundary, with fringes oriented at 45° to the input beam direction. It is used to expand the beam cross-section, and also to deflect it through a right-angle. The expanded beam then interacts with a travelling surface acoustic wave, which is itself excited by the radio-frequency signal through the piezo-electric effect, using an interdigital transducer. As a result, there are now two waves propagating near the substrate surface: the guided optical wave, and the SAW.

Since the substrate is acousto-optic, the surface acoustic wave creates a travelling phase grating, which may diffract the optical beam. Because the SAW has a relatively small period at low frequencies, the grating wavelength is rather large so diffraction takes place in the

R.R.A.Syms and J.R.Cozens  Optical Guided Waves and Devices 1
Raman-Nath regime. This results in the production by the grating of many diffraction orders, which may be a nuisance. However, the angle of deflection of the most important order - the first - depends linearly on the RF frequency to a good approximation. For simplicity, we shall ignore all higher diffraction orders from now on, and simply consider the principle order as ‘the diffracted beam’.

Figure 14.2-1 The planar integrated optic RF spectrum analyser.

An overlay Fresnel lens is used to focus this beam, to a point somewhere in the focal plane. In fact, because the lens performs a Fourier transform of the optical amplitude distribution, the distance of the focus from the optical axis also depends linearly on the angle of deflection. An external, linearly-spaced array of photodiodes is placed at the focal plane, so that each diode receives a signal when the RF drive is tuned to the correct frequency. The output of the array is therefore a parallel representation of the power spectrum of the RF input, which can be passed to further processing electronics. Typical performances reported for integrated optic spectrum analysers are 200 - 500 MHz bandwidth, 4 - 8 MHz resolution, 2 µs response time and dynamic ranges greater than 20 dB.

As the RF frequency rises, the period of the acoustic grating decreases. It will then become more selective, and the diffraction regime will gradually change from Raman-Nath to Bragg-type. This will result in a drop in diffraction efficiency, since the optical wave cannot now strike the acoustic grating at the Bragg angle for all possible RF frequencies. Furthermore, the interdigital transducer will only work over a limited frequency range. There several ways to overcome these problems, so that the overall RF bandwidth is increased. The most common is to use several transducers, each with a different ‘finger’ period and angle of orientation, driven in parallel from the same RF source. At any given frequency, only the transducer with approximately the correct period will launch a surface acoustic wave, which will then travel at the correct angle. Alternatively, the acoustic wave may be continuously ‘steered’ with frequency, using a single transducer arranged as a phased array.

Clearly, the circuit shown above could have been implemented in a variety of different ways - a telescope could have been used as the beam expander, the Fresnel lens could be replaced by a geodesic lens, and so on. The important constraint is that all the components must be made using a common technology, and should therefore be compatible. Consequently, it
does not necessarily follow that the component with the best performance in isolation should be used.

DESIGN EXAMPLE

We can estimate the parameters of a typical spectrum analyser using results derived in Chapters 4 and 7. Figure 14.2-2 shows the geometry; the input optical beam is assumed to be roughly collimated, with a Gaussian amplitude distribution of characteristic width $2w_{10}$.

![Figure 14.2-2 Geometry for calculation of RF spectrum analyser design.](image)

Even though the results of Chapter 4 were for three-dimensional waves, we can guess that the effect of focusing the beam by a two-dimensional lens of focal length $f$ will be to produce a Gaussian focal spot of characteristic width $2w_{20}$. This width is found by substituting $\beta$ for $k_0$ in Equation 4.7-4 to get:

$$w_{20} = 2f/\beta w_{10} = \frac{f\lambda_0}{\pi n_{\text{eff}} w_{10}}$$  \hspace{1cm} 14.2-1

For a beam of width $2w_{10} = 2 \text{ mm}$ and wavelength $0.633 \mu\text{m}$, propagating in a guide of effective index 2.2 (typical for Ti:LiNbO$_3$) and focused by a lens of 2 cm focal length, the width of the focal spot is $2w_{20} = 4 \times 2 \times 10^{-2} \times 0.633 / (\pi \times 2.2 \times 2 \times 10^{-3}) = 3.7 \mu\text{m}$. Since it will be impossible to resolve finer details using such a lens, the width of each detector should be of this order.

Now, in Chapter 7 we showed that the directions of the diffraction orders are defined by the grating equation, $\sin(\theta_1) = \sin(\theta_0) + L\lambda_0/n\Lambda$. For acousto-optic diffraction, the grating period $\Lambda$ is $F/v$, where $v$ is the acoustic velocity and $F$ is the frequency. Assuming a small angle of deflection, and normal incidence, the direction of the first order is then approximately:

$$\theta_1 \approx \frac{\lambda_0 F}{n_{\text{eff}} v}$$  \hspace{1cm} 14.2-2

Assuming that $v = 6.57 \times 10^3 \text{ m/s}$ in LiNbO$_3$, and that the transducer centre frequency is $F \approx 500 \text{ MHz}$, we get $\theta_1 \approx 0.022 \text{ radians}$ (or $1.25^\circ$). Neglecting the distance between the acoustic beam and the lens, the centre of the detector array should then be located at a distance $y = f\theta$ away from the optical axis. From the data above, we find $y \approx 2 \times 10^{-2} \times 0.022 \text{ m}$, or $440 \mu\text{m}$.

Differentiating Equation 14.2-2, we can then find the displacement $\Delta y$ of the focal spot that accompanies a change in frequency $\Delta F$, as:

$$\Delta y = \frac{f\lambda_0}{n_{\text{eff}} v} \Delta F$$  \hspace{1cm} 14.2-3
For a transducer of (say) 200 MHz bandwidth, we obtain \( \Delta y \approx 2 \times 10^{-2} \times 0.633 \times 2 \times 10^8 / (2.2 \times 6.57 \times 10^3) \) \( \mu \text{m} = 175.2 \mu \text{m} \). The detector array should therefore be of this length, to cover the full RF bandwidth. The number of output channels may be then be found as the number of resolvable spots in this distance, i.e. as \( N = \Delta y / 2w_20 \). Using the data above, we obtain \( N \approx 175.2 / 3.7 \approx 48 \). The channels will then be located at intervals of \( \approx 200 / 48 = 4.2 \) MHz in frequency-space.

Since the number of channels is so low, we might enquire as to the limitations on \( N \). Combining Equations 14.2-3 and 14.2-1, we obtain the surprisingly simple result:

\[
N = \left( \frac{\pi}{4} \right) \Delta F \tau
\]

14.2-4

where \( \tau \) is the time-of-flight of the acoustic wave across the optical beam, defined as \( \tau = 2w_10 / v \). To increase \( N \), while maintaining the same acoustic bandwidth, we must use either a wider optical beam or a material with a smaller acoustic velocity.

14.3 THE PLANAR INTEGRATED OPTIC DISC READ HEAD

The second application is more recent, a read head for an optical disc player. Here the aim is to use a focused laser beam to read the information contained in tiny pits, a few \( \mu \text{m} \) in diameter, which are etched in the surface of a rotating optical disc. Such a device might be used in a compact disc player, or in an optical data storage system (known as a CD ROM). The advantage to be gained from using optics in either case is the high potential storage capacity.

We may estimate this by calculating the number of pits of diameter \( d \) which may be stored on a disc of diameter \( D \). The value of \( d \) depends mainly on the size of focused spot we can create. This is limited by diffraction, and a typical (conservative) value might be \( d = 2 \mu \text{m} \). Taking \( D \approx 10 \text{ cm} \), we obtain a figure of \( \pi(D/2)^2 / \pi(d/2)^2 = 2.5 \text{ Gbits} \), a capacity much larger than that of current magnetic discs. Now, optical data storage must clearly involve some form of optical reading or writing, however it is configured. This time, the benefits of integration are the cost-reductions made possible by using a mass-production technology. This particular application shows that guided-wave optics may even penetrate the consumer goods market, a considerable achievement for such an embryonic technology.

Figure 14.3-1 shows a prototype read head, which is again based on a planar integrated optic chip, with an external laser diode and processing electronics. This time, the waveguide is fabricated on a silicon substrate, using sputtered Corning 7059 glass as the guiding layer, with a Si-N cladding layer and a SiO2 buffer layer. The circuit works as follows. The laser diode is butt-coupled to the chip edge, and so produces a diverging beam in the guide. This beam then encounters a Bragg grating device, a twin-grating focusing beam splitter (or TGFBS), which is effectively a split-grating lens. The grating is fabricated by writing a curved, periodic pattern in a resist layer (using an electron beam deflected under computer control), and then transferring the pattern to the Si-N layer by reactive ion etching. Most of the beam passes through the TGFBS without diffraction, to another grating component known as a focusing grating coupler (or FGC). This is used to couple light out of the guide and create a focal spot in a plane above the chip, where the optical disc is rotating.

Light reflected from the pit pattern on the disc is then coupled back into the guide, also by the FGC. The return signal is then split into two converging beams by the TGFBS, and each focussed beam falls on a split detector. These are made from p-n junction photodiodes.
fabricated directly in the Si substrate, and the SiO$_2$ buffer layer is tapered from the guiding region to the detecting area so that the light is gradually led onto the diodes. The four detectors give different signals, depending on the precise positions of the two return foci. Using combinations of these signals, external electronics can provide not only the desired readout signal, but also tracking and focusing error signals, which are used to ensure the correct position of the read head relative to the disc, through a feedback control loop. The tracking error signal is particularly important, because the track width is approximately one pit diameter (a few µm). Because likely errors in the concentricity of the central hole locating the disc are vastly greater than this figure, it would be impossible to follow the track correctly without accurate feedback control.

Figure 14.3-1  Integrated optic read head for an optical disc player: a) schematic (after Ura S.et al., IEEE J. Lightwave Tech. LT-4, 913-918 (1986); © IEEE 1986), and b) the actual device (photograph courtesy T.Suhara, Osaka University).

14.4 GUIDED-WAVE OPTICAL CHIP-TO-CHIP INTERCONNECTS

We shall now present some different examples of optical integration, using the channel guide components of Chapters 9 and 10. The first is an application designed for use with VLSI electronics, which is based on the following premise. It is becoming accepted that electrical connections amongst integrated circuits and between circuit boards will encounter serious problems, if either the speed of the circuitry or its complexity rises much further. This is mainly due to two factors. The first is the limitation on signal transmission...
bandwidth caused by RC time constants in the circuit. Contrary to expectations, these do not scale down with decreasing circuit dimensions, so that propagation delays eventually become more significant than gate delays. The second is interference caused by stray capacitances, which also become more significant as components are packed closer together.

Optical interconnects have therefore been proposed as an alternative to electronic links. These could involve either free-space or guided-wave optics, and both techniques are currently the subject of research (although it is likely that guided wave devices will prove more successful in the long run, because they are smaller, more rugged and optically more efficient). Figure 14.4-1 shows an early prototype circuit based on channel guide technology, which is designed to connect four LSI chips together on a common substrate. This particular circuit is a hybrid, owing to the difficulty of fully integrating the different technologies involved. The optical part is composed of semiconductor laser diodes, a multimode ridge waveguide circuit, and photodiodes. The waveguide circuit has four guides for transmitting the optical signals, and a reflective mixing device to broadcast the resulting signal uniformly back to receivers at each LSI chip.

![Guided wave optical chip-to-chip interconnection circuit](image)

Each guide is used for both transmitting and receiving. The laser diodes (LD), which are butt-coupled to the end of the guide, act as the transmitters; these are InGaAsP diodes, operating at $\lambda_0 = 1.3 \mu m$, which are bonded onto the substrate upside-down. Angled reflection mirrors tap a fraction of the return signal from the mixer, and pass it to the photodetectors (PD), which act as the receivers. These are PIN photodiodes, which are also bonded to the substrate in a hybrid fashion. The waveguides are fabricated in a high index material, an SiO$_2$-TiO$_2$ mixture, and a thick SiO$_2$ buffer layer is used to separate the guides from the high index Si substrate. In this example, the measured signal transmission capacity was 1 Gbit/s. Though an early result, this indicates considerable potential for the future.
14.5 THE CHANNEL WAVEGUIDE INTEGRATED OPTIC A-TO-D CONVERTER

We now consider an alternative channel guide device, with an application in signal processing. It turns out that analogue-to-digital conversion is one of the more difficult tasks to perform quickly using electronics, because many of the most common algorithms (e.g. successive approximation) are rather slow. What is really required is a device that compares a signal with a number of different reference levels at once, and then outputs a suitable binary code. Such devices do exist in electronics, and are known as parallel or 'flash' A-to-D convertors. However, they are expensive, and it was thought at one time that optical A-to-D conversion offered an advantage at the high-speed end of the market.

The operation of the optical A-to-D converter is based on the periodic response of the electro-optic Mach-Zehnder interferometer. We have already encountered this device in Chapter 9, in the guise of a modulator. It consists of two back-to-back Y-junctions, linked by straight waveguides. The first Y-junction splits the incoming beam into two components, which pass along the straight guides and are recombined at the second Y-junction. Either or both of the interferometer arms may have surface phase-shifting electrodes, to allow the relative phase of the recombining components to be altered. If the two are in-phase, the guided output is high, and if they are out of phase, it is low. More generally, the output is periodic in the phase imbalance, as shown in Figure 14.5-1a. If an additional static phase shifter is used to bias the interferometer, the sinusoidal response is shifted by a quarter of a period. If the optical output is now detected, and then thresholded by an electronic comparator at a level of half the maximum possible detector output, the comparator output will vary as a square-wave with voltage (Figure 14.5-1b). This corresponds to a one-bit binary representation of the applied voltage.

Because the electro-optic effect is linear, the output of a Mach-Zehnder interferometer must depend on both the drive voltage and the electrode length. For example, a short device will be less sensitive to voltage than a long device, and so on. Figure 14.5-2 shows the output characteristics that might be expected from three similar interferometers, with electrode lengths of L, L/2 and L/4. Clearly, all the responses are still periodic with voltage, but the longest device has twice the sensitivity of the intermediate one, and four times that of the shortest. If the additional processing described above - static phase-shifting, detection and thresholding - is applied to each interferometer, the comparator outputs will all vary as square waves, but with periods differing by factors of two. Taken together, they yield a three-bit binary representation of the drive voltage.
Figure 14.5-2 Power/voltage characteristics for Mach-Zehnder interferometers with different electrode lengths.

Figure 14.5-3 shows a prototype high-speed, four-bit parallel analogue-to-digital converter for radio-frequency signals, operating on the principles described above. The optical circuit consists of four Ti:LiNbO$_3$ Mach-Zehnder interferometers, which are arranged in parallel on a common substrate and fed by a common optical input through a Y-junction 'tree'. Each interferometer has the same RF input, and is equipped with travelling-wave electrodes for high-speed response, but the electrode lengths differ by factors of two, so that the longest has electrodes of length $L$ and the shortest of $L/8$.

Figure 14.5-3  A 4-bit parallel optical A-to-D converter.

The circuit therefore performs a four-bit analogue-to-digital conversion of the input voltage. In this example, the optical input is not continuous-wave; instead, it consists of a train of short pulses, so that the RF signal is sampled optically at a very high rate. Suitable sources include pulsed semiconductor lasers and mode-locked YAG lasers. Conversion speeds of almost 1 Gbit/s have been reached with prototype Ti:LiNbO$_3$ devices, but difficulties with the comparators and other readout electronics have prevented the anticipated market success. Despite this, the electro-optic A-to-D converter remains an ingenious and novel application of guided wave technology.
14.6 OPTICAL FIBRE SENSORS

We now move on to consider systems based around optical fibre and the all-fibre components of Chapters 8 and 10, beginning with fibre sensors. This is not an obvious application, because fibre is normally designed to provide a low-loss transmission medium for optical waves, which are confined near the fibre core and are therefore isolated from their environment by a substantial thickness of cladding. However, it is also possible to arrange for a guided wave to be affected quite strongly by external perturbations. For example, the core might be relocated much nearer the edge of the cladding, so that the evanescent field extends outside it. The guided wave is then directly affected by any external changes, and the fibre can act as a sensor.

Once this is accepted, there are many convincing arguments to support the use of fibre as a sensing medium. Because it is possible to arrange a considerable length of fibre in a coil, high sensitivity to a particular parameter can be achieved within a small sample volume. Furthermore, it has already been shown that fibre can be sensitized to a wide variety of measurands, including temperature, humidity, strain, electric and magnetic fields, pressure, acoustic waves, and the presence of different inorganic and organic chemicals. Optical sensors are non-electrical, which is important in many hazardous environments (e.g. petroleum plants), and can be read out remotely. It is even possible to build multiplexed sensor highways, which contain a number of sensors in series, that may be interrogated individually. Optical fibre sensors are therefore a promising future technology.

It is obviously impossible to do justice to a topic of this importance in a small section. All we can really do is discuss the major issues involved. Broadly speaking, fibre sensors are divided into two classes, operating on non-interferometric and interferometric principles, respectively. In the former, an external parameter is arranged directly to affect the transmission through the fibre by altering the propagation loss. One simple example is the microbend sensor, which works by inducing a periodic bend in the fibre by compression between two grooved plates (Figure 14.6-1). The amount of bending, and the consequent radiation loss, are proportional to the depth of compression or the pressure applied. This allows very cheap, simple sensors to be constructed, in which a change in the sensed parameter is directly reflected as a variation in optical transmission. However, all such sensors suffer from the following major disadvantage: it is hard to differentiate this particular loss mechanism from any other - for example, from the loss which would arise if the fibre links to and from the sensor section were accidently bent. This type of sensor is therefore 'lead sensitive'.

In the latter, the external parameter is arranged to affect the phase of the light instead. Unwanted perturbations generally have little affect on phase, so this minimises the lead sensitivity problem. However, it is still possible to recover the sensor output in the form of intensity variations, by using interferometry. Figure 14.6-2 shows an example of a generic architecture for interferometric fibre sensors, based once again on the Mach-Zehnder interferometer.
Figure 14.6-2 An interferometric optical sensor.

The optical circuit is slightly different this time, since it contains two 3 dB fibre couplers (which act as 50 : 50 beam splitters) instead of Y-junctions. The couplers are linked together by a short reference arm and a much longer sensing arm, which is arranged as a coil. As before, the first splitter divides the optical input into two components. The upper part passes through the sensor coil, and so is affected by changes in the measurand. For example, if the device is to be used as a hydrophone, it is arranged for pressure waves passing through the coil to dilate it, thus altering the phase of the sensing beam. The beam passing through the reference arm is not exposed to the measurand, so its phase is unaffected by such changes. The second coupler is used to recombine the beams. These will add together coherently, provided that (i) a laser source is used, with a coherence length greater than the path imbalance of the two arms, and (ii) the two components have the same polarization. The latter condition generally requires the entire circuit to be constructed from polarization-preserving fibre, with the input adjusted so that only a single polarization component is launched.

Unlike the Y-junction Mach-Zehnder interferometer, the coupler-based MZI has two possible optical outputs. We can calculate these quite simply, as follows. First, we recall that the transmitted and cross-coupled amplitudes $A_1$ and $A_2$ of a synchronous coupler with coupling coefficient $\kappa$ and length $L$ can be written in the form:

$$A_1 = \cos(\kappa L) \quad \text{and} \quad A_2 = -j \sin(\kappa L)$$

In a 3 dB coupler (for which $\kappa L = \pi/4$) these reduce to:

$$A_1 = 1/\sqrt{2} \quad \text{and} \quad A_2 = -j/\sqrt{2}$$

An optical input, of amplitude $A_0$, will therefore be split by the first coupler into two components, of amplitude $A_0/\sqrt{2}$ in the reference arm and $-jA_0/\sqrt{2}$ in the sensing arm. We now assume that the sensing beam will incur an additional phase delay $\Phi$ relative to the reference arm, due to the intrinsic difference in their lengths and to the effect of the measurand. Ignoring any common phase delay, the amplitudes of the two outputs are:

$$B_1 = A_0 \left(1/\sqrt{2}\right)^2 + A_0 \left(-j/\sqrt{2}\right)^2 \exp(-j\Phi) = (A_0/2) \{1 + \exp(-j\Phi)\}$$

and:

$$B_2 = A_0 \left(1/\sqrt{2}\right) \times (-j/\sqrt{2}) + A_0 \left(-j/\sqrt{2}\right) \times (1/\sqrt{2}) \exp(-j\Phi) = -j(A_0/2) \{1 + \exp(-j\Phi)\}$$

The output powers are then proportional to:

$$P_i = B_1 B_1^* = (A_0^2/2)\{1 - \cos(\Phi)\} = P_0 \sin^2(\Phi/2)$$
and:

\[ P_2 = B_2 B_2^* = \left( A_0^2 / 2 \right) \left( 1 + \cos(\Phi) \right) = P_0 \cos^2(\Phi/2) \]

14.6-6

The two outputs are therefore complementary, in that their sum is equal to the input power \( P_0 \). They are shown plotted as a function of \( \Phi \) in Figure 14.6-3. Normally, the phase imbalance is biased so that the deviations imposed by the measurand occur only in a small range about the typical bias point \( \Phi_0 \). This can be accomplished using additional static phase-shifters. Near the bias point, small changes in \( P_1 \) and \( P_2 \) are both linearly dependent on small changes in \( \Phi \), and hence on variations in the measurand, which makes assessment of the state of the sensor by processing electronics particularly simple. Larger excursions of the measurand (e.g. of more than one period) may be tracked by a variety of techniques, including 'fringe-counting'.

![Figure 14.6-3 Variation of the powers \( P_1 \) and \( P_2 \) with the phase imbalance \( \Phi \).](image)

**14.7 THE INTEGRATED OPTIC FIBRE GYROSCOPE**

We now consider mixed systems, employing both fibre and integrated optics. We begin with another type of interferometer, the Sagnac interferometer (which we have already encountered as the fibre loop mirror in Chapter 10). This has also been investigated extensively for sensor applications. It shows little sensitivity to any parameter except rotation, and is therefore of great potential use, since it can act as a gyroscope. In fact, the fibre gyro has considerable advantages over its mechanical equivalent, since it is lightweight, requires no 'run-up' time, and needs little maintenance. Three such gyros (sensing rotation about three orthogonal axes), combined with three linear accelerometers, could form the basis of future solid-state inertial navigation systems, and they are actively being developed for applications as diverse as aircraft navigation, missile guidance and oil-well logging.

Since we have already described the static behaviour of a Sagnac interferometer, we will concentrate here on the situation when the interferometer is rotating. In this case, it can be shown that there is a phase shift between the two counterpropagating beams in the loop, a phenomenon known as the Sagnac effect. Although an accurate description actually requires the invocation of general relativity, we shall base the discussion on approximate theory, which provides the correct result to first order.

Figure 14.7-1 shows an experimental gyroscope based on a Sagnac interferometer. It is constructed from a Ti:LiNbO\(_3\) channel waveguide integrated optic chip (which contains a number of signal routing and processing components), an optical fibre sensing coil, an
external laser diode source, and detector (which will typically be followed by further processing electronics). The optical circuit works as follows. The N-turn fibre coil, which is of radius $R$, is arranged as the sensing loop. Light from the laser is butt-coupled into the chip, split into two by Y-junction #1, and launched in opposite directions round the fibre coil. After passing round the coil, which has a length $2\pi NR$, the two components are recombined by Y-junction #1. A second Y-junction, #2, is then used to pick off a proportion of the power, which is passed to a detector. According to the usual principle of interferometry, the detected signal is high if the two recombining components are in-phase, and low otherwise.

![Diagram of an integrated optic fibre gyro](image)

Figure 14.7-1 An integrated optic fibre gyro (after L.M. Johnson, Proc. SPIE 566, 96-98 (1985)).

We can calculate the phase change accompanying rotation using the following simple argument. Normally, the time-of-flight round the loop is $\tau$, given by:

$$\tau = \frac{2\pi NR}{c}$$

14.7-1

We now assume that the coil rotates at an angular speed $\Omega$, so the recombination point will continually move away from one beam and towards the other. In a time $\tau$, it moves an angular distance $\Omega \tau$, so one beam must travel a distance $2\pi NR + \Omega \tau R$ to reach it, the other a distance $2\pi NR - \Omega \tau R$. The beams therefore take different times to exit the loop. The time-of-flight difference is:

$$\Delta \tau = \frac{(2\pi N + \Omega \tau)R}{c} - \frac{(2\pi N - \Omega \tau)R}{c}$$

14.7-2

This corresponds to a phase shift (known as the Sagnac phase shift) of:

$$\Delta \Phi = \omega \Delta \tau = \frac{4\pi NR^2 \omega \Omega}{c^2}$$

14.7-3

where $\omega$ is the optical frequency. Consequently, the relative phase of the recombining beams is linearly dependent on the angular rotation rate. The sensitivity is controlled by the area enclosed by the coil and by the number of turns. This illustrates the main advantage of fibre optics, which has turned the Sagnac interferometer into a practical proposition - it is easy to make $N$ large without incurring high loss. The phase shift can also be written in terms of the
length $L$ of the fibre, the coil diameter $D$, and the wavelength of the light used. If this is done, we obtain:

$$\Delta \Phi = \frac{2\pi LD\Omega}{\lambda c}$$  \hspace{1cm} \text{(14.7-4)}$$

The phase shift is therefore linearly dependent on the length of fibre used and the coil diameter.

**DESIGN EXAMPLE**

We can calculate the typical sensitivity of a gyroscope as follows. Using a 1 km length of fibre arranged in a coil of 10 cm diameter, and a wavelength of $\lambda \approx 1 \mu m$, we obtain from Equation 14.7-4:

$$\Delta \Phi \approx 2\Omega$$  \hspace{1cm} \text{(14.7-5)}$$

Now, the earth's rotation rate is $\Omega_e = \frac{2\pi}{(24 \times 3600)} \text{rad/s} = 7.3 \times 10^{-5} \text{rad/s}$. Even to register this motion, we must therefore be able to measure a phase-shift of $\Delta \Phi \approx 1.5 \times 10^{-4}$ rad with the design above. For an inertial navigation system, we need to do considerably better than this - say, at least a thousand times better. A typical phase shift will thus be about $10^{-7}$ rad, an incredibly small figure. This presents a considerable measurement problem, which is only partially solved at the moment.

**THE MINIMUM CONFIGURATION**

To achieve this high sensitivity, the optical paths travelled by the two beams must be identical when the gyro is not rotating, so the system must be reciprocal. Figure 14.7-2 is an example of the so-called 'minimum configuration' needed to achieve reciprocity in a gyro. The output is taken from the loop at exactly the point where the input is injected, and a polarizer is used to ensure that only a single polarization mode is launched and detected. Generally, a number of further optical components must also be integrated on the substrate to improve performance. Here, phase- and frequency-shifters have been added. These may be used to increase the signal-to-noise ratio of the detected signal through various signal-processing tricks, or allow closed-loop tracking of the Sagnac phase shift.

**14.8 HIGH SPEED, GUIDED WAVE OPTICAL COMMUNICATIONS**

A major application of guided wave optics is in high-speed optical communications. In fact, point-to-point links have now reached such an advanced stage that data can be transmitted at very high bit rates over staggering distances. Such progress represents a handsome confirmation of the original promise of fibre optics, and shows just how well the considerable investment required to develop low-loss, low-dispersion fibre and the other vital components of a communications system - lasers of high spectral purity, high-speed modulators, and fast, low-noise detectors - has paid off.

**A HIGH-SPEED POINT-TO-POINT LINK**

To illustrate this, Figure 14.8-1 shows a recent experiment, in which data were transmitted at 4 Gbit/s and received after travelling the phenomenal distance of 117 km without regeneration. The importance of the elimination of repeaters cannot be over stressed - not only does this lessen the need for fleets of maintenance engineers for land-based systems,
but it drastically improves the reliability of undersea links, which can only be repaired at quite considerable expense (or not at all!). Here, the source was a high-power InGaAsP/InP buried heterostructure laser, operating at 1.53 $\mu$m wavelength and externally modulated by a travelling-wave Ti:LiNbO$_3$ directional coupler. The detector was an InGaAs avalanche photodiode with a GaAs FET preamplifier, coupled to the fibre using a GRIN-rod lens. This type of system might well act as the basic trunk route of a future long-distance telecommunications network.

![Figure 14.8-1](after S.K. Korotky et al., IEEE J. Lightwave Tech. LT-3, 1027-1030 (1985); © IEEE 1985).

Bidirectional links are clearly feasible using similar technology, and a variety of demonstration land lines have now been installed for assessment and characterisation. However, no single user is likely to require access to the entire bandwidth of an optical fibre, which is vast. Instead, this will be shared between a number of users, or multiplexed. The two major strategies used are wavelength division multiplexing (WDM), whereby each user is assigned a specific wavelength interval with which to communicate; and time-division multiplexing (TDM), where each user is instead offered a set fraction of the available time. Once again, we can do little in the space available here to describe the advantages and disadvantages of each method in depth - the reader is best advised to consult a specialised optical communications text for details. Instead, we offer examples of the implementation of each method, using some of the hardware we have encountered so far.

**OPTICAL WAVELENGTH-DIVISION MULTIPLEXING DEVICES**

We begin with wavelength-division multiplexing. Figure 14.8-2 shows the main features of an optical WDM system, which operates as follows. A set of N users are first assigned wavelengths $\lambda_1$, $\lambda_2$, ..., $\lambda_N$ with which to communicate; these might be derived from a set of lasers, emitting light of different wavelengths. Each user may modulate his source independently, so that wavelength $\lambda_n$ acts as a carrier of information for Channel #n. Data from all N channels is combined together by the multiplexer (MUX), and then passed to a common channel (a single optical fibre) for transmission. At the far end, the signals are separated by the demultiplexer (DEMUX), and forwarded to their final destinations.
The MUX component may consist of a set of discrete optical filters arranged in cascade, as shown in Figure 14.8-3a. For example, a set of asymmetric directional coupler filters might be used as the basic combiner element; alternatives include mode-conversion filters and Bragg gratings. Each filter is arranged to couple maximum power to the common channel at a different wavelength, so that Filter \#n has a response that peaks at wavelength \( \lambda_n \) (as in Figure 14.8-3b). In this way, all N channels may be combined into a single output without loss of power. The DEMUX component may be a similar network, operating backwards.

Alternatively, the multiplexer may be a single dispersive component, which effectively provides all N filters in parallel. This is normally based on a grating of some kind. For example, Figure 14.8-4 shows a prototype four-channel optical WDM chip fabricated by the LETI organization, France. The chip is formed from a silicon substrate, carrying three silica overlayers fabricated by plasma-enhanced chemical vapour deposition.

The lowest layer acts as a low-index buffer layer, separating the guiding layer from the high-index substrate. The middle (guiding) layer is phosphorus-doped, to increase its refractive index by \( \Delta n = 7 \times 10^{-3} \) over that of undoped \( \text{SiO}_2 \). This layer is etched into ridges near the chip edges before the deposition of the upper layer, which acts as a cladding. Consequently, the input and output channels consist of buried ridge guides, which interface to a planar guiding region near the chip centre and to optical fibres at the chip edge. The fibres are automatically aligned in the correct position by grooves anisotropically etched directly into the silicon substrate.

After leaving the input channel guides, the input beams propagate as diverging cylindrical waves, which are combined into a single cylindrical beam converging on the output channel guide by a dispersive imaging device. This is a Fresnel mirror (the reflective equivalent of a Fresnel lens), which provides a wavelength-dependent, reflective focusing action. The
mirror is formed by etching a blazed pattern right through the silica overlayers to the silicon substrate, so that total internal reflection is obtained at the silica/air interface. The chip can combine four optical channels into a single output; the centre wavelength is 1.545 µm, and the channel separation is 20 nm. Operating in the reverse direction, the chip can separate an input containing all four wavelengths into individual outputs. The insertion loss is less than 6 dB, while the crosstalk between channels is better than 20 dB.

Figure 14.8-4 A 4-channel wavelength multiplexer on a silicon substrate: a) schematic (after LETI Technical Data Sheet DOPT 9053), and b) the actual device (photograph courtesy S.Valette, LETI).

The advantage of wavelength-division multiplexing is that it is asynchronous, so that each user may have access to the common channel instantaneously. It is an efficient process, in that the total bandwidth required is approximately N times that of the individual channels. The disadvantage is the need to fabricate the MUX and DEMUX components. This may be an extremely demanding task, if the number of channels is large and the centre wavelengths are closely-spaced.

OPTICAL TIME-DIVISION MULTIPLEXING DEVICES

We now consider the alternative of a time-division multiplexed communications system, the main features of which are shown in Figure 14.8-5. This time, the N users all communicate
on the same carrier wavelength. However, they may no longer have access to the shared channel at will; instead, they are allocated specific time-slots. In Figure 14.8-5, Channel #1 is assigned to the slot centred on time $t_1$, Channel #2 to that on time $t_2$, and so on. Data from all N users is combined together by the multiplexer, which interleaves the time-slots of the individual channels so that the overall data rate is uniform, and then passed to the common channel for transmission. At the far end, the signals are separated again by the demultiplexer. This is done by sampling and redirecting the data, at exactly the rate and sequence of the original interleaving process. Consequently, synchronisation information must also be passed between the MUX and DEMUX components.

Figure 14.8-5 The principle of an optical TDM communications system.

Optical time-division multiplexers and demultiplexers may both be constructed from a set of channel waveguide switches, arranged to connect a set of inputs to a single output in turn. For example, Figure 14.8-6 shows a 1 x 4 demultiplexer, fabricated on a LiNbO$_3$ substrate using three electro-optic Mach-Zehnder interferometric (MZI) switches. These are modified versions of the devices described in Chapter 9; at their output ends, a hybrid X-junction has been used instead of a Y-junction. This works like a 3 dB coupler, and allows each interferometer to function as a two-way switch, rather than as a simple modulator. In fact, the precise device used is unimportant - the MZI could equally well be replaced by a directional coupler. The important feature is that the switch can act as a high-speed 'gate' when driven by a sinusoidal signal. This is clearly true for the Mach-Zehnder switch, which will give an output from one arm when the driving voltage is close to zero, and from the other when it is close to that required for $\pm \pi$ radians phase-shift between the arms. It is also the case for a directional coupler, which will give a cross-coupled output only when the voltage is close to zero.

Figure 14.8-6 A 1 x 4 optical time demultiplexer driven by a single sinewave generator (after M. Haga et al., IEEE J. Lightwave Tech. LT-3, 116-120 (1985); © IEEE 1985)
The optical input is shown at the left, and is assumed to consist of a train of pulses at regular intervals. These are to be distributed cyclically to each of the outputs in turn, following a clocking sequence. The switches are arranged as a binary tree, and are driven from a common signal generator synchronised to the input data stream and having a period of T seconds, equivalent to the time frame occupied by four input pulses. This driving signal is shown as the full line in Figure 14.8-6.

There are slight but important differences between the switches. Modulator #1 has no optical bias, and its electrode length is chosen so that the peak driving voltage causes a ±π radians shift. Consequently, an output is obtained from its upper arm near the times t = 0, T/2, T and so on. Similarly, light passes to the lower arm at t = T/4, 3T/4 etc. These outputs are used to feed the second-stage devices in the tree. Modulator #2, on the other hand, is arranged to have a static phase bias of π/2 radians between its arms, and its electrodes are half as long as those of Modulator #1. As a result, its sensitivity to voltage is halved, and the effect of the voltage can only be an additional ±π/2 phase shift. At t = 0 and t = T, the total shift is then π radians, so an output is obtained from the uppermost guide. At t = T/2, though, the net phase shift is zero, so the output emerges from the second uppermost guide. The operation of Modulator #3 is similar, but an electrical delay line is used to shift the phase of the driving signal (shown dashed in Figure 14.8-6) through 90°. Consequently, outputs are obtained from the second-lowest and lowest guides at t = T/4 and t = 3T/4, respectively.

The net effect is to distribute the pulsed input amongst the four outputs in turn. Experimentally, the chip was driven by a 1 GHz signal, and was able to demultiplex an input optical signal from 4 to 1 Gbit/s. Alternatively, the device could have been run backwards, and used to combine four inputs as a single output data stream. This demonstrates the impressive capability of integrated guided wave components for high-speed data handling.

14.9 FIBRE LASERS AND AMPLIFIERS

Our final example is rather different. In Chapter 12, we described the operation of semiconductor lasers and travelling-wave amplifiers. Here, we will discuss more briefly some similar devices based on rare-earth-doped optical fibre, which appear very promising for use in all-fibre systems.

Remember that optical gain is obtained in semiconductors by inverting the electron population between the conduction and valence bands. In electrically pumped devices, this is achieved by minority carrier injection using a p-n junction. However, this particular mechanism is not unique, and optical gain has been demonstrated in countless other materials systems. More generally, it is not necessary for the material to be crystalline (or even solid), or for the pumping to be performed electrically; all that is required is the existence of a set of suitable energy levels and a method for inverting the population between two particular levels. Consequently, electrically excited gas lasers and optically-pumped solid-state lasers (which operate by transitions in guest atoms embedded in crystalline or amorphous hosts) have been the dominant laser types of the previous two decades. Of the former, He-Ne and Ar⁺ lasers are the most common, while ruby, Nd:YAG and Nd:glass lasers typify the latter.

The inversion mechanism normally involves three or four energy levels. For example, in the former, optical pumping might involve the excitation of an electron from the lowest state (Level #1) to the highest (Level #2) through the absorption of a photon of energy $h\nu = E_2 - E_1$, followed by a rapid decay to a slightly lower state (Level #3). If electrons have a long lifetime in Level #3, it is possible to build up the population in this state by continuous
pumping until it exceeds that in Level #1. Gain is then available at the wavelength corresponding to the transition between Levels #3 and #1.

Recently, it has been found to be possible to incorporate lasing species in the core of silica glass fibres, thus forming the fibre equivalent of a solid-state laser medium. For the active material, attention has been focussed on the rare-earth elements erbium, neodymium, prasodymium and thulium. Of these, the most important are prasodymium, which may be used to provide gain at 1.3 \( \mu m \) wavelength, and erbium, which operates at 1.55 \( \mu m \). For erbium, windows for optical pumping exist at 0.8, 0.98 and 1.48 \( \mu m \) wavelength, allowing the use of powerful and potentially inexpensive laser diode pump sources (especially the 0.98 \( \mu m \) window).

Figure 14.9-1 shows a typical travelling-wave **fibre amplifier** experiment involving Er\(^{3+}\)-doped fibre. Two 0.82 \( \mu m \) GaAs/GaAsAs lasers are used as pump sources; these inject light into opposite ends of the Er\(^{3+}\)-doped fibre, so that population inversion is maintained along its entire length. Each pump beam is coupled into the amplifier using a wavelength-selective fibre coupler; this allows the signal beam (derived from a 1.535 \( \mu m \) laser diode) to be combined efficiently with the pump at the input, and the amplified signal to be separated from any residual pump radiation at the output. A similar arrangement can be used with a 1.48 \( \mu m \) pump, although a high-performance WDM coupler is required because of the close proximity of the pump and signal wavelengths. In this example, the gain medium was 90 metres of GeO\(_2\)-SiO\(_2\) fibre doped with 34 ppm of erbium. The pump threshold for transparency was 15 mW, and 22 dB of gain was obtained from a total pump power of 54 mW.

Figure 14.9-2 shows a typical example of a DBR fibre laser based on a polished fibre coupler grating. This time, the active medium is a 6 metre length of silica fibre doped with 330 ppm of neodymium, and the pump laser is a 0.83 \( \mu m \) GaAs/GaAlAs laser diode coupled into the fibre through a wavelength-selective mirror. The threshold for laser
operation was 2 mW, and ≈ 1.5 mW of output power was obtained for a pump power of 10 mW at the fibre laser wavelength of 1.084 μm.

Figure 14.9-2 A DBR fibre laser (after I.M.Jauncey et al., Elect. Lett. 22, 987-988 (1986)).
PROBLEMS

14.1 An integrated optic RF spectrum analyser is to be constructed on a LiNbO$_3$ substrate, with a tilted interdigital transducer. Assuming that the optical wavelength is 0.633 µm, calculate the angle of transducer tilt required for optimum operation in the Bragg diffraction regime at the transducer centre frequency of 500 MHz. If the transducer bandwidth is 200 MHz, how wide may the acoustic beam be before Bragg selectivity limits the diffraction efficiency? The effective index of the Ti:LiNbO$_3$ waveguide is 2.2, and the acoustic velocity is 6,570 m/s.

[0.627°, 7.5 mm]

14.2 The figure below shows a small portion of a conducting path on a VLSI electronic circuit. The conductor has cross-sectional area $A$ and length $L$, and is separated from the ground plane by an insulator of thickness $h$. Show that, if all the linear dimensions of the circuit are scaled by $\alpha$, the RC time constant associated with the path remains unchanged. Why is this a potential problem? Describe several advantages of guided wave optical interconnects over this type of circuit.

14.3 Binary codes are just one possible way of representing a set of states using a two-valued sequence. An alternative is offered by Gray codes. These have the advantage that only one bit of the code changes in moving from one state to the next, which is useful in preventing errors. Gray codes may be constructed using the following rule. Start by assigning a code of all zeros to the lowest state; then assign codes for successively higher states by changing the least significant bit that generates a new code. Using this rule, write down the codes for all possible two-bit numbers. Design a two-bit, Gray-coded integrated optic A-to-D converter.

14.4 A beam of wavelength $\lambda_0$ propagates through a length $L$ of optical fibre. Show that the phase change accompanying a temperature change $\Delta T$ is $\Delta \phi = 2\pi/\lambda_0 \{n_{\text{eff}} \alpha + dn_{\text{eff}}/dT\} L \Delta T$, where $\alpha$ is the thermal expansion coefficient of the fibre and $n_{\text{eff}}$ is its effective index. Assuming that $\alpha = 0.004 \times 10^{-5}$ °C$^{-1}$ and $dn_{\text{eff}}/dT = 1.1 \times 10^{-5}$ °C$^{-1}$ for silica fibre, and that $n = 1.458$ for bulk silica, estimate the phase change per °C for a 10 cm length of fibre at $\lambda_0 = 0.85$ µm. Which effect contributes most to the phase change, thermal expansion or refractive index variation?

[8.56 rad/°C]

14.5 The figure below shows an interferometric sensor constructed from two 3 dB couplers linked by fibres. The additional time delay experienced by a wave propagating along the sensing arm over that experienced along the reference arm is $\tau$. Show that, for a single-frequency optical input, the output is ambiguous in $\tau$, i.e. that the same output will be obtained for a variety of different values of $\tau$. Show further that the sensitivity of the device to variations in $\tau$ will go to zero whenever $\omega \tau = n \pi$, where $\omega$ is the...
angular frequency of the input. When is the sensitivity maximum? Discuss the implications of these features for a practical sensing system.

4.6 One solution to the ambiguity problem identified in Question 14.5 is to increase the bandwidth of the optical input. The figure below shows the input used in the linear FM method. Here, the input is frequency modulated by a sawtooth wave of period T. For $0 \leq t \leq T$, for example, the frequency variation is linear with time, following $f(t) = f_1 + (f_2 - f_1) t/T$. Assuming that $\tau < T$, calculate and sketch the response of the sensor of Question 14.5 this input.

14.7 Discuss the importance of launching and detecting a single polarization mode in an optical fibre gyro. Calculate the time-of-flight difference for two beams propagating in opposite directions around a 20 cm diameter coil containing 1 km of fibre, at a rotation rate of $10^{-5}$ rad/sec. Would this be directly measurable? What is the corresponding phase difference, at $\lambda_0 = 0.85 \mu m$?

$[2.22 \times 10^{-20} \text{ sec}; 4.93 \times 10^{-5} \text{ rads}]$

14.8 i) The transfer function $P_n(\lambda_0) = \text{sinc}^2\left\{\pi(\lambda_0 - \lambda_n)/\Delta\lambda\right\}$ is a good approximation to the response of many real filters. Sketch the variation of $P_n(\lambda_0)$ with $\lambda_n$. What does the term $\Delta\lambda$ represent?

ii) The figure below shows a demultiplexer used in a WDM system. The channels are regularly spaced at wavelengths $\lambda_1, \lambda_2, \ldots, \lambda_N$, such that $\lambda_{n+1} - \lambda_n = \Delta\Lambda$, and the peak in the passband of the $n^{th}$ filter is centred on $\lambda_n$. However, due to the finite extent of the filter response, a fraction of the power at wavelength $\lambda_{n+1}$ will be accepted by the $n^{th}$ filter. Use the response above to estimate the worst-case interchannel crosstalk, assuming that $\Delta\Lambda \geq \Delta\lambda$. 
14.9 A directional coupler (with a coupling length of $\kappa L = \pi/2$) is driven by a 200 V peak-to-peak sinusoidal voltage of frequency 1 GHz. The voltage required to switch the coupler from the cross state to the bar-state is 10 V. Draw a dimensioned sketch of the cross-coupled output with time, for a CW optical input. What application might this have in a TDM system?

14.10 A point-to-point communications link operating at $\lambda_0 = 1.55 \, \mu\text{m}$ consists of 50 km of single-mode fibre with a loss of 0.2 dB/km. The source is a 10 mW laser diode, coupled to the fibre via a GRIN-rod lens. The coupling efficiency is 25%. There are 10 fusion splices in the link, each contributing 0.1 dB loss. Assuming that the receiver requires a signal of 1 mW to operate, calculate the length of Er$^{3+}$-doped fibre required in a travelling-wave amplifier arranged as a pre-detector signal booster. When pumped at 0.82 $\mu$m wavelength, the Er$^{3+}$-doped fibre has a gain of 0.25 dB/m. [28 m]
SUGGESTIONS FOR FURTHER READING

Barnes W.L., Townsend J.E. "Highly tunable and efficient diode pumped operation of Tm³⁺ doped fibre lasers" Elect. Lett. 26, 746-747 (1990)
1. OVERVIEW

Silica-on-silicon is the most promising materials system for integrated optics yet developed. It has the following major advantages over previously-developed technologies (namely ion-exchanged glass, Ti: LiNbO₃, GaAs and InP):

- The fabrication process is cheap, and compatible with Si-based microelectronics.
- Connection to single-mode optical fibres is cheap, and low loss.
- A wide range of low-loss passive components may be produced, on large substrates.
- Slow-speed switching components are possible, based on the thermo-optic effect.
- Hybrid integration may be used to add optical sources.
- Active components such as amplifiers are currently being developed.

The only major disadvantage is the current lack of an operating mechanism for high-speed switches (for example, an electro-optic effect). Silica-on-silicon devices should therefore find an application as low cost components in systems for fibre-to-the-home. Here, the most immediate requirements are for cheap, passive star components. Figure 1 shows the main elements of a silica-on-silicon device, which is constructed as follows:

- The substrate is single-crystal Si, normally a (100) - oriented, 525 µm thick p-type wafer. Substrate crystallinity is exploited in the fabrication of V-groove alignment.
features, used to position single-mode fibres for accurate butt coupling to guides in the integrated optic circuit.

- A buffer layer is used to isolate the waveguide from the high-index substrate. Normally, this layer is a considerable thickness (> 12 µm) of pure SiO₂, and fabrication of such a thick layer represents a considerable technological challenge.

- Waveguides are typically formed as buried channel guides, for compatibility with optical fibres. Guide cores are made from a layer of doped SiO₂, etched into a rectangular cross-section. Further processing may be used to modify the core shape and reduce sidewall scattering.

- A thick cladding (typically 10 - 20 µm pure SiO₂) is used to bury the waveguide cores, symmetrising the modes and isolate them from their surroundings.

Two main variants are currently being investigated, based on large, low-Δn guides, and smaller, medium-Δn guides. Each offers advantages and disadvantages in terms of propagation loss, fibre coupling loss, and the minimum allowable bend radius. These are summarised in Table 1. Even smaller, high-Δn guides have also been formed in a few materials; however, these suffer very high fibre coupling loss (> 3 dB / facet).

<table>
<thead>
<tr>
<th>Core size (µm)</th>
<th>Δn</th>
<th>Propagation loss (dB/cm)</th>
<th>Fibre coupling loss (dB per facet)</th>
<th>Bend radius (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 x 8</td>
<td>0.25</td>
<td>&lt; 0.1</td>
<td>0.05</td>
<td>25</td>
</tr>
<tr>
<td>6 x 6</td>
<td>0.75</td>
<td>0.3</td>
<td>0.5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1. Characteristics of SiO₂: TiO₂ single-mode channel guides (after M. Kawachi, Opt. Quant. Elect. 22, 391-416 (1990)).

2. MATERIALS

The basic optical material of silica-on-silicon components is SiO₂, whose refractive index is 1.458 at 0.633 µm wavelength. The refractive index differences needed to form waveguides are established by controlled doping with impurities.

Dopants causing a decrease in refractive index include B₂O₃ and F. Dopants causing an increase include Al₂O₃, As₂O₃, GeO₂, N₂, P₂O₅, TiO₂, and ZrO₂. Figure 2 summarises the dependence of refractive index on dopant concentration for the materials above.

Generally, the dopants are similar to those used to manufacture optical fibres; however, some have been found to be more suitable than others for the particular application of integrated optics. For example:

- Since TiO₂ offers a large refractive index change, very high Δn guides may be formed in TiO₂: SiO₂, and this material scheme has been investigated extensively. However, it has relatively high scattering loss.

- Since phosphorus is pentavalent rather than tetravalent, it is a network modifier. P₂O₅: SiO₂ therefore has a low reflow temperature, making it useful for the formation of cores whose cross-sectional shape will be modified by controlled melting.

- GeO₂ has a near IR absorption band located relatively far away from the key 1.5 µm wavelength, and hence gives low absorption loss.

- Guides have also been produced using pure oxide and nitride films. In particular, Si₃N₄ (refractive index 2) has been used to produce extremely small, high-Δn guides.
For each material system, the main fabrication difficulty lies in the formation of thick layers of high-quality silica and doped silica on silicon substrates. A large number of different deposition methods have now been developed to deposit the thick glass layers required for silica-on-silicon integrated optics. These include:

- Thermal oxidation and nitridation
- Sputtering
- Chemical vapour deposition (CVD)
- Plasma-enhanced chemical vapour deposition (PECVD)
- Flame hydrolysis deposition (FHD)
- Sol-gel deposition (SGD)

The first three methods are older and somewhat restrictive, and only the last four are being actively developed for fabrication of optical devices.

**THERMAL OXIDATION AND NITRIDATION**

High quality, thick SiO$_2$ films may be grown on Si substrates by thermal oxidation in a tube furnace. Silica may be grown in dry or wet atmospheres, following the reactions:

$$\text{Si} + \text{O}_2 \rightarrow \text{SiO}_2$$

or

$$\text{Si} + 2\text{H}_2\text{O} \rightarrow \text{SiO}_2 + 2\text{H}_2$$

The growth rate is limited by the need for the gas molecules to diffuse through the film to the oxide-silicon interface, where they react. For long oxidation times $t$, the film thickness $t_f$ varies parabolically with time as:

$$t_f = \sqrt{Bt}$$

Here $B$ is the parabolic rate constant at the oxidation temperature $T$. Table 2 details the rate constant for growth in dry and wet atmospheres. The main disadvantage of the process is the small value of $B$, requiring a long immersion in a wet atmosphere at an elevated temperature to obtain a suitable thickness (24 days at 1100°C for $t_f = 16$ µm).
the process is not easily modifiable for the formation of doped oxides; consequently, it is most commonly used to grow thick films of pure oxide for use as buffer layers. An exception is silicon oxynitride guiding layers, which have been formed by thermal nitridation of SiO₂ layers in NH₃ at similar temperatures.

<table>
<thead>
<tr>
<th>Oxidation temperature T (°C)</th>
<th>B (dry atmosphere) (µm²/hr)</th>
<th>B (wet atmosphere) (µm²/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>0.045</td>
<td>0.720</td>
</tr>
<tr>
<td>1100</td>
<td>0.027</td>
<td>0.510</td>
</tr>
<tr>
<td>1000</td>
<td>0.0117</td>
<td>0.287</td>
</tr>
<tr>
<td>920</td>
<td>0.0049</td>
<td>0.203</td>
</tr>
</tbody>
</table>

Table 2 Parabolic rate constant B for oxidation of silicon in dry and wet atmospheres (from Sze S.M. *VLSI Technology*, McGraw-Hill International Book Company (1983)).

**SPUTTERING**

R.F. sputtering can be used to deposit dielectric layers on Si by physical transfer of material from a target film lying parallel to the substrate. Coating is performed in a vacuum chamber, containing an inert gas such as Ar at a pressure of 10⁻³ - 10⁻² Torr. A plasma is then created by an RF field in the space between target and substrate. The target acts as the cathode, and the substrate as the anode, so that positively charged Ar⁺ ions strike the target and eject atoms from it. Although the sputtered atoms travel in random directions, many move towards the substrate. After undergoing further collisions with the plasma, the sputtered atoms strike the substrate from a wide range of angles. The technique can therefore be used to coat non-planar surfaces. Sputtering has been used to deposit many materials for optical waveguide applications (e.g. Corning #7059 glass, pure SiO₂, pure Al₂O₃ etc.). Its main disadvantage for production is the slow deposition rate (around 1 µm/hr).

**CVD AND PECVD**

Chemical vapour deposition (CVD) is a method of depositing glassy films from hot gas mixtures. It originated in VLSI microelectronics, and its further development for optics has been pursued mainly by AT & T Bell Labs. Horizontal and vertical gas flows are both used. For example, a horizontal-tube hot-wall reactor might consist of a quartz tube with an external resistive heater. The substrate is placed on a quartz support, and loaded into the tube through a removable end-cap. Process gases are injected from one end, and heated prior to arrival at the substrate by contact with the walls. They flow over the substrate as a laminar boundary layer, and are adsorbed on the surface where they react to form a film. By-products of the reaction and unreacted gases are exhausted at the end of the tube.

The temperature and pressure are both highly dependent on the process used. One major branch of CVD involves the thermally activated oxidation of hydrogen compounds, such as silane (SiH₄), phosphine (PH₃) and arsine (AsH₃). For example, pure silica layers may be deposited at atmospheric pressure by a mixture of silane and oxygen, at a temperature of 400 - 450°C, following the reaction:

\[ \text{SiH}_4 + \text{O}_2 \rightarrow \text{SiO}_2 + 2\text{H}_2 \]  

(3)
Doped silica layers can also be formed easily. For example, phosphosilicate glass ($P_2O_5$: $SiO_2$) can be produced following the two simultaneous reactions:

$$\begin{align*}
SiH_4 + 2O_2 & \rightarrow SiO_2 + 2H_2O \\
2PH_3 + 4O_2 & \rightarrow P_2O_5 + 3H_2O
\end{align*}$$

A silica-on-silicon process based on phosphosilicate glass may therefore be implemented in a single reactor equipped with controllable sources of SiH$_4$, PH$_3$, and O$_2$.

Using plasma-enhanced CVD, films can be deposited at lower temperatures (200 - 350°C). PECVD has mainly been applied for optics by LETI and Nortel. The substrate is placed on a heated susceptor, which is arranged to act as one of a pair of R.F. electrodes. A plasma is then established between the electrodes, and the process gases are then bled into this region (which may have a very high electron temperature). The plasma might be derived from an inert gas; for example, SiO$_2$ films may be deposited from SiH$_4$ and N$_2$O in an Ar plasma. Alternatively, it might itself form one of the reactants; SiON may be deposited using SiH$_4$, N$_2$O and NH$_3$, with the resulting film composition being determined by the N$_2$O: NH$_3$ ratio.

**FLAME HYDROLYSIS DEPOSITION**

Flame hydrolysis deposition (FHD) is a method of depositing glasses in the form of a powdery 'soot'. It was originally developed for the fabrication of optical fibre preforms in the VAD process, and its further development for integrated optics has been performed mainly by NTT. In FHD, silica is produced by hydrolysing a mixture of SiCl$_4$ and O$_2$ in a glass-burner, following the reaction:

$$SiCl_4 + 2H_2O \rightarrow SiO_2 + 4HCl$$

The result is a stream of small silica particles, or 'soot', which are blown onto the substrate by the oxyhydrogen flame, where they stick. During deposition, large numbers of substrates (up to 50) are rotated on a turntable, and the soot stream is traversed to ensure uniform coverage as shown in Figure 3. A variety of dopants may also be incorporated in the films very simply. For example, TiO$_2$, and GeO$_2$, may be included, by adding TiCl$_4$, or GeCl$_4$, to the gas stream. P$_2$O$_5$, and B$_2$O$_3$, dopants may be obtained from BCl$_3$, and PCl$_3$, sources. Because the composition of each layer may be varied, merely by adjusting the flow of gas, graded-index guides and tapered structures may be built up by flame hydrolysis.

![Figure 3. Arrangement for flame hydrolysis deposition (after Kawachi M., Opt. Quant. Elect. 22, 319-416 (1990)).](image)
Thick films may be built up very quickly by FHD. However, after the deposition of sufficient material, the porous soot must be compacted or 'sintered' in a furnace at ≈ 1300°C to produce glass of optical quality. This results in considerable (90%) film shrinkage, making the final thickness uncertain.

SOL-GEL DEPOSITION

The sol-gel process is an alternative method of depositing glassy films, based on the hydrolysis and polycondensation of metal alkoxides. Three stages are involved: first, the formation of a stable suspension of particles within a liquid (a sol); second, further processing of the sol to form a continuous solid network permeated by liquid (a gel), and third, drying of the gel to form a porous glass. Usually, the sol is applied to a surface to be coated by spin coating, so sol-gel material is also referred to as spin-on-glass.

One commonly-used process involves the fabrication of silica (SiO₂) films from tetraethylorthosilicate (Si(OC₂H₅)₄, or TEOS). However, a wide range of other metal alkoxides can be used, allowing a corresponding range of dielectric films. For example, refractive index variations can be introduced by variations in the sol composition (e.g. by the use of TiO₂, derived from a tetrapropylorthotitanate or TPOT precursor, as a dopant).

For silica, the basic sol-gel process involves the reaction:

$$\text{Si(OR)}_4 + 2\text{H}_2\text{O} \rightarrow \text{SiO}_2 + 4\text{ROH}$$

However, this proceeds in a number of steps. The alkoxide is first hydrolysed, so that several OR groups on each of the TEOS molecules are replaced by OH. For example, mixing TEOS with a quarter the stoichiometric ratio of water and refluxing in the presence of a catalyst (e.g. HCl), will replace on average one group, as shown in Figure 4a; repeating the process will result in the replacement of two groups as shown in Figure 4b. These reactions may be slowed by dilution with solvent so that the sol is stable enough for use.

![Figure 4](image)

Figure 4. The sol-gel process for formation of silica from TEOS: a) hydrolysis of one ligand group, b) hydrolysis of two groups, and c) polymerisation by condensation.
When some groups have been hydrolysed, polymerization can take place, as shown in Figure 4c. Here the reaction by-product is water, so this causes further hydrolysis and polymerization in a chain reaction until a connected network (known as a gel) is built up. Again, this reaction can be slowed by dilution with solvent. However, when the sol is coated onto the substrate (e.g. by spin coating) the solvent is driven off. This accelerates the polymerisation reaction, resulting in rapid gelation on the wafer surface. Gelation and final solvent evaporation are normally completed by a baking step.

Drying involves a reduction in volume of the spun layer, which results in tensile stresses in the film. After a certain deposited thickness (normally a few µm), this generally results in failure of the film through cracking. Recently, a new process has been developed for the deposition of thick (20 µm) sol-gel films on silicon substrates, based on the repetitive use of spin-coating followed by densification by rapid thermal annealing. This has overcome the problem of stress-cracks, and enabled waveguides to be constructed using both the silica-titania and the phosphosilicate systems. Figure 5 shows a typical rapid thermal processor.

![Figure 5. Arrangement for rapid thermal annealing in sol-gel silica-on-silicon process.](image)

**3. CHANNEL WAVEGUIDES**

**TOPOGRAPHIC WAVEGUIDE FABRICATION**

Each of the processes described in the previous Section may be used to deposit suitably thick layers of silica or doped silica on silicon substrates. Buried channel guides may then be formed in these layers by a variety of different methods.

Etched channel guide fabrication begins with the deposition of a bilayer glass structure (normally consisting of a doped SiO₂ guiding layer on an undoped buffer layer) on an Si substrate. The wafer is then coated with a mask material, which is patterned lithographically with waveguide features. The mask is used as an etch stop during reactive ion etching (RIE) or reactive ion beam etching (RIBE), which is used to remove most of the upper layer material, leaving the waveguide as a rectangular core, as shown in Figure 6a. Reactive ion etching can be based on either fluorine or chlorine chemistry. For example, in the former case, a C₂F₆ - C₂H₄ plasma has been used with an amorphous silicon mask; in the latter, a CHF₃ - Ar - O₂ plasma with a Cr mask. Etch rates are typically in the range 500 - 1500 A/min. After RIE, the core walls are normally rough.
Depending on the materials used, the core may be reflowed to lower the surface roughness and hence reduce scattering losses. Reflow may be carried out very simply if the core glass has a lower melting point than the buffer layer. At the core melt temperature, surface tension pulls the core into a cylindrical cross-section as shown in Figure 6b. After reflow, guides are buried beneath a thick cladding. Using FHD, an extremely thick cladding (30 μm) can be deposited, so that the guide has an ideal symmetric shape. Using CVD, deposition of a thick cladding is more time-consuming. Furthermore, material is deposited on the core and on the exposed buffer layer at a uniform rate, so that the shape of the growing surface gradually evolves without substantial planarization.

**PROPAGATION LOSSES**

The propagation losses of topographic waveguides fabricated by most of the methods described are now extremely low. The main contributions to loss arise from:

- Sidewall scattering: this is largely eliminated by reflow of the guide core and cladding.
- Extrinsic absorption caused by the presence of hydrogen, primarily near 1.39 μm (excitation of 2nd harmonic of stretching vibrations of O-H bonds) and 1.52 μm (3rd harmonic of Si-H bond and 2nd harmonic of N-H bond): this can be dramatically reduced by thermal annealing.
- Intrinsic absorption caused by the tails of the mid-infrared absorption bands of S-O and similar bonds. The most important of these lie at 7.3 μm (B-O), 8.0 μm (P-O), 9.0 μm (Si-O) and 11.0 μm (Ge-O). For operation at near-infrared wavelengths, germania is therefore the most favourable dopant.

**BIREFRINGENCE**

All the processes used for one-sided deposition of thick silica films yield bowed Si wafers, implying that the films will be stressed. The presence of a large biaxial film stress causes refractive index changes through the photoelastic effect. The changes in the refractive index caused by an in-plane film stress $\sigma_f$ is:

\[
\Delta n_0 = -n^3/2 \cdot (q_{11} + q_{12}) \sigma_f; \quad \Delta n_e = -n^3/2 \cdot 2q_{12} \sigma_f
\]  

Here $q_{11}$ and $q_{12}$ are components of the stress-optical tensor $[q]$. Since $q_{11} \neq q_{12}$, the material will become birefringent, with:
\[ n_e - n_0 = -(n^3/2) (q_{12} - q_{11}) \sigma_f \]  

(8)

In silica, \( q_{12} - q_{11} = 2 \times 10^{-12} \text{ m}^2/\text{N} \) so for compressive film stresses \( n_e - n_0 \) is positive. Assuming (say) \( \sigma_f = -200 \text{ MN/m}^2 \), we obtain \( n_e - n_0 = 6.2 \times 10^{-4} \). A similar birefringence \( n_{TM} - n_{TE} \) is observed between the effective indices of waveguide modes. Since the magnitude of the birefringence is significant compared with the index changes used to form the guide, steps may be taken to control it in some applications, particularly interferometric filters.

### COUPLING TO FIBRES

Besides the possibility of low-loss channel guides, silica-on-silicon integrated optics offers cheap, low-loss connection to single-mode fibres. The coupling efficiency \( \eta \) between an input transverse field \( E_{in} \) and the \( \mu \)th transverse modal field \( E_{\mu} \) of a waveguide is found (assuming the fields are normalised) from the inner product \( \eta = | \langle E_{in}, E_{\mu} \rangle |^2 \). For high coupling efficiency, the two fields should be matched in size and shape, and also coincide spatially. For connection to standard 1.5 \( \mu \text{m} \) single-mode fibre (which has a mode-field diameter of \( \approx 8 \mu \text{m} \)), silica-on-silicon guides should have large core size and low \( \Delta n \). As shown in Table 1, 8 x 8 \( \mu \text{m} \) cores with \( \Delta n \) of 0.25% gives \( \eta = 98.8\% \), or 0.05 dB loss.

However, with such a large modal field, tight bending in the planar optical circuit is impossible; the minimum tolerable bend radius here was 25 mm. Some trade-off between bend and coupling loss is possible; decreasing the core size and increasing the \( \Delta n \) to 6 x 6 \( \mu \text{m} \) and 0.75%, respectively, reduced the minimum bend radius to 5 mm, but increases the coupling loss to 0.5 dB. However, the bend radii needed in high-density circuits cannot be achieved without much smaller, higher-\( \Delta n \) guides (e.g. 3 x 3 \( \mu \text{m} \) GeO\(_2\)-doped cores with \( \Delta n \) of 2%, which offer usable bends of 2 mm radius). These cause excessive coupling loss.

Attention has therefore been given to adiabatic tapers capable of transforming the guide from large core, low \( \Delta n \) type at the fibre input to small core, high \( \Delta n \) type in the optical circuit. Two methods have been demonstrated. The first is a modification of FHD; here, the gas composition and torch scan pattern are varied to deposit a thinner, higher \( \Delta n \) guiding layer near the centre of the circuit, while the guide width is varied by lithography. In the second, a small-core, high \( \Delta n \) guide is modified by local diffusion of dopants. One end of the substrate is clamped in a cooled holder, and the other is heated as shown in Figure 7. For example, heating at 1300˚C for 5 hours to form a 25 mm taper decreased the coupling loss of a 6.5 x 6.5 \( \mu \text{m} \) GeO\(_2\)-doped guide with a \( \Delta n \) of 0.75% from 0.43 to 0.08 dB.

![Figure 7. Method of forming a waveguide taper by local thermal diffusion.](after Yanagisawa M., Yamada Y., Kobayashi M. Elect. Lett. 28, 1958-1959 (1992))

Even with good mode matching, efficient coupling will not be obtained if the relative positions of fibre and waveguide are inaccurately specified. For low cost, alignment must be passive and achieved through integrated mechanical alignment features rather than by active
optimisation of position. Early attempts were made to use the thick silica layers themselves to fabricate suitable features; however, the Si substrate is now used almost exclusively.

The technology used is anisotropic etching down Si crystal planes, based on earlier work on fibre connectors. Two etch mixtures are commonly used, based on KOH and EDP (a mixture of ethylenediamine, pyrocatechol and water) respectively. KOH requires a Si₃N₄ surface mask, because it attacks SiO₂; EDP can be used directly with silica, and is therefore compatible with silica-on-silicon integrated optical devices. However, it is a carcinogen.

In the most common process, etching of (100) oriented Si through rectangular openings in a mask results in the formation of V-grooves, which can act as kinematic mounts for optical fibre. The desired shape is generated by the large difference between the etch rates R₉₀₀ normal to the surface and R₁₁₁ normal to the (111) planes. Typically, this is of the order of 40 : 1. Anisotropically etched silicon can be used as a precision mount for ribbon optical fibre, locating the individual cores on a precise spacing. A prepared fibre end may then be polished and glued to the cleaved end facet of a waveguide device. However, it is advantageous if the waveguide and alignment features are fabricated on a common substrate as shown in Figure 8; in this case, the sloped ends of the grooves must be removed. Fully pigtailed devices (e.g. 1 x N power splitters) have now been demonstrated.

Figure 8. Fibres located in etched V-grooves and butt-coupled to a planar lightwave circuit (photo courtesy A.S. Holmes, Imperial College)

4. PASSIVE WAVEGUIDE DEVICES

The main device variants of silica-on-silicon integrated optics are similar to those developed for Ti : LiNbO₃. However, due to the larger substrate sizes, far more complex circuits have become possible. Key components are:

- Power splitters based on Y-junctions and directional couplers
- Stars based on tree-structures and radiative distributors
- MUX/DEMUX components based on Mach-Zehnder interferometers and gratings
- MUX/DEMUX components based on arrayed waveguide gratings

Since it is amorphous, silica is not electro-optic. However, a number of slow-speed switching devices operating via the thermo-optic effect have been developed. These include:
- Phase modulators
- Switches and switch arrays

POWER SPLITTERS AND STARS

1 x 2 and 2 x 2 power splitting functions are performed by Y-junctions and directional couplers. The former are waveguide forks with branch angles of 1° or less, and provide broadband operation; they may also be cascaded into binary tree structures to provide higher-order splitting. Directional couplers consist of two parallel waveguides which exchange power by evanescent coupling. According to standard theory, the coupling efficiency of a directional coupler is given by:

$$
\eta = \sin^2\left(\sqrt{v^2 + \xi^2}\right) / (1 + \xi^2/v^2)
$$

(9)

where \( v = \kappa L \) is the coupling length, \( \kappa \) the coupling coefficient, \( L \) the device length, \( \xi = \Delta\beta L/2 \), and \( \Delta\beta \) the difference in propagation constant between the guides. \( \kappa \) is given by:

$$
\kappa = k_0^2 \left< \Delta n^2 E_1, E_2 \right>
$$

(10)

where \( k_0 = 2\pi/\lambda \), \( E_1 \) and \( E_2 \) are normalised transverse fields of the modes in the two guides, and \( \Delta n^2 \) is the perturbation in relative dielectric constant seen by each guide due its neighbour. Symmetric couplers (\( \Delta\beta = 0 \)) having \( \kappa L = \pi/4 \) can act as 50% (3 dB) power splitters. The coupling coefficient depends exponentially on the guide separation; however, in most silica-on-silicon systems, usable lengths lie in the range \( 0.1 \text{ mm} < L < 5 \text{ mm} \).

Using components of this type, highly efficient low-order star couplers for broadcast systems may be constructed. For example, Figure 9a shows a 2 x 8 star based on a single directional coupler and two Y-junction trees. Figure 9b shows a combinatorial 8 x 8 star based entirely on directional couplers, which divides the power from a single input equally among the 8 outputs. In this case, waveguide intersections are required; these have both low loss and low crosstalk provided the angle of intersection is chosen suitably.
When the number of ports is large, the combinatorial approach is impractical. In this case, an alternative method is used, in which the power is coupled between input and output ports by radiation. Figure 9c shows a typical structure; here the input and output guides are arranged in regular fans, and connected via a section of planar guide. Light emerging into this region from the end of an input guide will spread by diffraction, so that the far-field pattern covers the entrance to the output fan, coupling to the individual waveguide modes with good uniformity. Although this process must necessarily involve some loss, theoretical predictions for total efficiency are as high as 35%, independent of the number of ports. High efficiencies have been achieved in practice, and the number of ports has risen very rapidly. For example, the sizes of 1 x N devices have increased from 1 x 16 to 1 x 128, while the sizes of N x N devices have risen correspondingly from 19 x 19 to 144 x 144.

FOURIER TRANSFORM PROPERTIES OF THE RADIATIVE STAR

The radiative star has additional interesting properties. Particularly, it can behave as a kind of lens (a Rotman lens). In common with other lenses, it can perform a Fourier transform operation. In this case, the transform is a discrete Fourier transformation, as we now show.

Figure 10a shows the geometry of the radiation region, which for now ignores the input and output waveguides. Here, two surfaces, each with radius R, are separated by a distance R.

By simple trigonometry, the distance PQ between points P and P is given by:
\[ PQ^2 = R^2 \{ \sin(\theta) - \sin(\theta') \}^2 + \{ 1 - [1 - \cos(\theta)] - [1 - \cos(\theta')] \}^2 \]  

(11)

Assuming small angle approximations, Equation 11 reduces to:

\[ PQ^2 = R^2 \{ (\theta - \theta')^2 + (1 - \theta^2/2 - \theta'^2/2)^2 \} \]

(12)

Multiplying out the brackets, and using the binomial approximation, we may take the square root of both sides, to get:

\[ PQ \approx R(1 - \theta\theta') \]

(13)

In the case where P and Q can both only be chosen from a set of N equally separated and symmetrically located points points, we may write \( \theta' = \alpha p \) and \( \theta = \alpha q \), where \( \alpha \) is the angular separation of adjacent points and q is the port number, as shown in Figure 10b. In this case, the path length PQ is:

\[ PQ \approx R(1 - \alpha^2 pq) \]

(14)

Consider now the case when a single amplitude \( A_p \) is presented at point P on the input surface. The amplitude at Q may be found by making allowances for the change in phase and amplitude incurred in propagating the distance PQ. If this is done, we obtain:

\[ A_q = (A_p / \sqrt{N}) \exp\{-j\beta R(1 - \alpha^2 pq)\} \]

(15)

Where \( \beta \) is the propagation constant, while the factor of \( \sqrt{N} \) accounts for power conservation, since the input power is split roughly N ways. If more than one amplitude \( A_p \) is presented, the amplitude at Q may be found as a summation of the form:

\[ A_q = k \sum_{p=-M}^{M} A_p \exp(j\beta R\alpha^2 pq) \]

(16)

Here \( k = \exp(-j\beta R) / \sqrt{N} \), and M = (N-1)/2. Equation 16 shows that the output amplitude distribution now has the form of a discrete Fourier transform of the input distribution. The discrete transform is similar to the continuous transform, but in addition has certain cyclic properties, so that the set of outputs obtained for a given input distribution are obtained again in a shifted manner of the input distribution is also shifted. We shall see how this result is exploited in an important type of filter, the arrayed waveguide grating multiplexer.

INTERFEROMETRIC MULTIPLEXERS AND DEMULTIPLEXERS

Efficient wavelength-division multiplexers and demultiplexers are required for WDM systems. Versatile filters may be constructed from Mach-Zehnder interferometers (MZIs), as shown in Figure 11. Figure 11a shows the principle. Two 3 dB couplers are combined into an MZI with an imbalance of \( \Delta L \) between the two arms. According to standard theory, the transfer efficiencies for the two outputs for an input to the lower port are:

\[ \eta_{upper} = \cos^2(\beta \Delta L/2) \]

\[ \eta_{lower} = \sin^2(\beta \Delta L/2) \]

(17)

where \( \beta = 2\pi n_{eff}/\lambda \), and \( n_{eff} \) is the effective index. Thus, by appropriate choice of \( \Delta L \), \( \eta_{upper} \) and \( \eta_{lower} \) may be made unity at two different wavelengths \( \lambda_1 \) and \( \lambda_2 \). A multiplexed input may then emerge with the individual components spatially separated. As \( \Delta L \) increases, \( \lambda_1 \)
and $\lambda_2$, approach each other, so devices with large imbalance may be used for frequency-division multiplexing, as shown in Figure 11b.

![Silica-on-silicon interferometric MUX/DEMUX components](image)

Figure 10. Silica-on-silicon interferometric MUX/DEMUX components

More complicated networks containing multiple interferometers can be constructed to improve the filter response or increase the number of channels. Figure 11c shows a 1.5 $\mu$m MUX based on three MZIs with different imbalances. This combined four channels separated by 8 nm, with an average fibre-to-fibre loss of 2.6 dB and -16 dB crosstalk. Similar devices can act as matched filters.

### ARRAYED WAVEGUIDE GRATING MULTIPLEXERS

As the number of channels rises, a more systematic approach to filter design is required. A simple solution is provided by the arrayed waveguide grating multiplexer (AWG MUX), a compact device that can combine a large number of similar filters in a compact layout. The device consists of two radiative stars, arranged back-to-back and linked by a set of waveguides - the arrayed waveguide grating - of linearly varying path length as shown in Figure 12. A high degree of functionality is obtained from the device, as we now show.

We consider the case of an input to guide $p$ of the left-hand star. We have previously shown that the output amplitude from the $q^{th}$ guide of the star is related to the input by:

$$A_q = (A_p/\sqrt{N}) \exp(-j\beta R[1 - pq\alpha^2])$$

After passing through an additional delay of length $L_q = L_0 + q\Delta L$, the amplitude becomes:

$$A_q' = (A_p/\sqrt{N}) \exp(-j\beta[R[1 - pq\alpha^2]+ L_0 + q\Delta L])$$

This amplitude is then the $q^{th}$ input of the right-hand star, and finally emerges from (say) the $r^{th}$ output waveguide.
Taking into account the effect on both amplitude and phase of the second star, we may obtain the output as:

$$A_r = \left( \frac{A_p}{N} \right) \exp\left(-j\beta \left( R[1 - pq\alpha^2] + R[1 - rq\alpha^2] + L_0 + q\Delta L \right) \right)$$  \hspace{1cm} (20)

Combining terms and summing over all possible values of q, the output amplitude is then:

$$A_r = \left( \frac{A_p}{N} \right) \sum_{q=-M}^{M} \exp\left(-j\beta q[\Delta L - (p + r)R\alpha^2] \right)$$  \hspace{1cm} (21)

Finally, the output power is obtained as:

$$P_r = \left( \frac{P_p}{N^2} \right) \left| \sum_{q=-M}^{M} \exp(-j\beta q[\Delta L - (p + r)R\alpha^2]) \right|^2$$  \hspace{1cm} (22)

Where $P_p = A_p^2$ is the input power. Equation 22 provides a complete description of the response of the AWG MUX. However, it is relatively complicated, and its significance is best explained by considering various simple cases.

If (for example) the input is to the central guide ($p = 0$), and the output is also taken from the central guide ($r = 0$), then Equation 22 becomes:

$$P_{r=0} = \left( \frac{P_p}{N^2} \right) \left| \sum_{q=-M}^{M} \exp(-j\beta q\Delta L) \right|^2$$  \hspace{1cm} (23)

The summation in Equation 23 now represents the addition of a set of terms of different phase. When $\beta \Delta L = 2\pi$, these terms all add in-phase, and the value of the summation must be N. In this case, the output power must be $P_r = P_p$, and the transmission $P_r/P_p$ must be unity. Since $\beta = 2\pi n_{eff}/\lambda$, we therefore see that the transmission is high at a discrete set of wavelengths $\lambda_{\nu} = n_{eff}\Delta L/\nu$.

The wavelengths for maximum transmission are spectrally separated by $\Delta \lambda = \lambda_{\nu} - \lambda_{\nu+1} = \lambda_{\nu}^2/n_{eff}\Delta L$, and hence move closer together as $\Delta L$ increases. Note, however, that $\Delta \lambda$ is essentially unaffected by N, the number of guides in the array.
Away from the optimum wavelengths $\lambda_\nu$, the transmission falls. To estimate how it varies, we first substitute $\lambda = \lambda_\nu + \delta\lambda$, where $\delta\lambda$ is small. Using the binomial approximation, $\beta\Delta L$ may then be written as:

$$\beta\Delta L = (2\pi n_{\text{eff}}\Delta L/\lambda_\nu) \left(1 - \delta\lambda/\lambda_\nu\right)$$

Using the expressions for $\lambda_\nu$ and $\Delta\lambda$, Equation 24 may then be expressed as:

$$\beta\Delta L \approx 2\nu\pi - 2\pi\delta\lambda/\Delta\lambda$$

Clearly, the first term in Equation 25 will make no difference to the summation in Equation 23, which then reduces to:

$$P_{r=0} = \left(\frac{P_{p=0}}{N^2}\right) \left|\sum_{q=-M}^{M} \exp(+j2\pi q\delta\lambda/\Delta\lambda)\right|^2$$

The summation in Equation 26 now has the form $\sum_{q=-M}^{M} x^q$, where $x = \exp(+j2\pi\delta\lambda/\Delta\lambda)$. After some manipulation of the series we can show that:

$$\sum_{q=-M}^{M} x^q = \frac{x^{N/2} - x^{-N/2}}{x^{1/2} - x^{-1/2}}$$

And hence that:

$$P_{r=0} = \left(\frac{P_{p=0}}{N^2}\right) \frac{\sin(\pi N\delta\lambda/\Delta\lambda)}{\sin(\pi\delta\lambda/\Delta\lambda)}$$

Figure 13 shows the variation in transmission efficiency obtained from Equation 28 for an 11-port AWG MUX. Here we have assumed that each star coupler has a radius of curvature $R = 10$ mm, that the length increment in the waveguide array is $\Delta L = 10$ $\mu$m, that the angular separation of the ports is $\alpha = 0.15^\circ$, and that the effective index is $n_{\text{eff}} = 1.50$.

![Figure 13. Variation of transmission with wavelength for an 11-port AWG MUX, with an input to and an output from the central port.](image-url)
10 lies at $\lambda = 1.5 \, \mu m$, and this is the one shown. Away from this optimum wavelength, the transmission falls, so that the overall response has a narrow band-pass characteristic near the maximum and a set of side-lobes beyond. Because of the existence of additional transmission maxima, we might expect peaks in response to be obtained at other wavelengths. They will be similar in shape to that shown in Figure 13. However, they will generally be unimportant if they lie outside the spectral range of the signals being used.

The full expression (Equation 22) may be used to compute the output from the entire set of ports, for an arbitrary input. Figure 14 shows the response obtained for the same parameters as Figure 13. As can be seen, the transmission function from each port is essentially a shifted version of the variation previously shown for port 0. Thus, the AWG MUX represents an extremely efficient method of spatially superimposing a set of highly similar filters in a compact architecture. If (for example) the input is a wavelength multiplexed signal containing spectral elements $\lambda_1, \lambda_2, \ldots, \lambda_N$, and each wavelength is arranged to coincide with a peak in transmission of one of the filters, the wavelengths will emerge spatially separated, one from each port.

![Figure 14. Variation of transmission with wavelength for an 11-port AWG MUX, with an input to the central port. The legend above the graph gives the output port number.](image)

The channel separation is largely controlled by the radius of curvature R of the stars. Similarly, the channel width is mainly set by the number of guides in the array. AWG MUX components have been fabricated as 40 channel devices, with channel separations as low as 50 GHz, and with overall insertion losses as low as 6 dB. Further modifications have also been made to flatten the pass-band characteristic, and to remove polarization sensitivity.

SWITCHABLE COMPONENTS

Slow-speed optical phase shifters and modulators may be constructed very simply, based on the thermo-optic effect. For example, Figure 15 shows a phase modulator, which consists of a Ti metal thin-film heater of length L located above a channel waveguide. Under normal conditions, the optical path accumulated in the device length is $\phi = \beta L = 2\pi n_{eff} L/\lambda$. When a current is passed through the heater, a temperature gradient is established across the silica layers that is sufficient to remove the heat generated, by conduction to the substrate. Some heat is also removed by convection, and a steady-state temperature $\Delta T$ above room temperature is reached.
The resulting phase change is:

\[ \Delta \phi = \frac{2\pi}{\lambda} \left( \frac{dL}{dT} n_{\text{eff}} + L \frac{dn_{\text{eff}}}{dT} \right) \Delta T \]

\[ = \frac{2\pi}{\lambda} \left( \alpha n_{\text{eff}} + \frac{dn_{\text{eff}}}{dT} \right) L \Delta T \]  

(29)

where \( \alpha = 1/L \frac{dL}{dT} \) is the thermal expansion coefficient of the guide, and \( \frac{dn_{\text{eff}}}{dT} \) is the temperature coefficient of refractive index. Assuming that \( n_{\text{eff}} \approx n_{\text{silica}} \), \( \frac{dn_{\text{eff}}}{dT} \) is normally much greater than \( \alpha n_{\text{eff}} \), so that \( \Delta \phi \approx \frac{2\pi}{\lambda} \frac{dn_{\text{eff}}}{dT} L \Delta T \).

For silica, \( \frac{dn}{dT} \approx 1.1 \times 10^{-5} /\degree C \). A temperature rise of 6.8\degree C is therefore required to obtain a phase change of \( \pi \) radians in a device of length 10 mm at 1.5 \( \mu \)m wavelength. This can be achieved with an electrical power of \( \approx 0.5 \) W, with a response time of 2 msec.

Thermo-optic devices are polarization-insensitive, and their operation does not depend on absolute temperature. However, they do suffer some disadvantages. In particular, since the heat they generate is not well localised, they cannot be used to control a directional coupler by desynchronising the two guides. The building block most commonly used in Ti: LiNbO\textsubscript{3} integrated optics for switching and tuneable filtering is therefore inappropriate for silica-on-silicon at present.

Thermo-optic phase modulators can be used to provide modulation or switching when they are incorporated into a Mach-Zehnder interferometer as shown in Figure 16a. In this case, the outputs from the two arms are given by:

\[ \eta_{\text{upper}} = \cos^2(\Delta \phi/2) \]

\[ \eta_{\text{lower}} = \sin^2(\Delta \phi/2) \]

(30)

Here \( \Delta \phi \) is as in Equation 29. Two-way switches of this type can be formed into gate matrix switching arrays. Unfortunately, the required architecture is more complex than for comparable coupler-based Ti: LiNbO\textsubscript{3} devices. Figure 16b shows the construction of a typical gate matrix switch array, which has already been demonstrated in 4 x 4 form. This particular switch is constructed from Y-junction trees and thermo-optically-driven interferometric modulators. Effectively, it is composed of four cross-connected 1 x 4 power splitters and sixteen modulators, and therefore functions as a strictly non-blocking
broadcast switch. Total insertion losses and crosstalk reported are 27 dB and -12 dB, respectively.

Figure 16 a) Thermo-optic Mach-Zehnder interferometer; b) gate matrix switch array (after Himeno A., Kobayashi M. IEEE J. Lightwave Tech. LT-3, 230-235 (1985))

WAVELENGTH ROUTING

By combining AWG MUX and switch components, sophisticated signal routing operations may be carried out in wavelength division multiplexed systems. We have already shown that the AWG MUX may demultiplex the channels in a WDM system, as shown in Figure 17a. When operated in reverse, the AWG MUX may combine a number of channels of different wavelengths onto a single multiplexed output as shown in Figure 17b.

Figure 17. System representation of an AWG MUX used as a) a demultiplexer and b) a multiplexer
Using three AWG MUX components and a set of cross-point switches, an ADD-DROP multiplexer may be constructed as shown in Figure 18. Here, the first DEMUX component is used to spatially separate the channels. Each channel is then connected to one of two possible MUX components by a switch. The upper MUX is used to combine any desired channels in a single output port, while the lower MUX is used to combine any undesired channels into a drop port. Any dropped channels may be replaced using inputs from a set of add ports. In this way, the spectral content of a wavelength-multiplexed signal may be locally modified, with different spectral components being routed at will through the system. Entire components of this type have now been fabricated in silica-on-silicon.

5. ACTIVE DEVICES

HYBRID INTEGRATION

In addition to the range of passive components previously described, silica-on-silicon is being used as a low-cost platform for hybrid integration of components such as lasers and detectors that cannot be fabricated in the glass itself. Considerable use is made of micromachining techniques to provide the alignment features needed to obtain low loss connections. For example, buried silicon terraces can be used to locate the components accurately in the vertical plane, and mesas can be etched into the glass itself to provide lateral alignment.

For example, Figure 19 shows a hybrid-integrated 1.3/1.5 µm transceiver chip developed by NEL, which is based on terraced silica-on-silicon platform. The chip contains waveguide circuitry, an embedded dielectric filter for demultiplexing, a photodiode receiver, and a laser diode source with its own monitor photodiode. The laser is mounted directly on an exposed silicon terrace, so that efficient cooling is obtained via the substrate. The waveguide core is located just above the surface of the terrace, so that it may be butt coupled to the diode. The waveguide circuit contains Y-junction components for signal routing, and the dielectric filter is placed in a slot sawn carefully across the waveguide.
Figure 19. 1.3/1.5 µm transceiver chip based on terraced silica-on-silicon platform (source: NTT Electronics Ltd.)

ERBIUM DOPED WAVEGUIDE AMPLIFIERS

Active devices such as optical amplifiers are also now being constructed using silica-on-silicon integrated optics. The principles involved are very similar to those of existing fibre-based amplifiers, but integration may offer a cost advantage, since pump lasers and filters may all be co-integrated with the active waveguide. These cost advantages become significant when entire arrays of amplifiers are required, for example in WDM systems.

Figure 20a shows the energy levels of a three-level optical amplifier, which might be used as a signal booster. Gain is achieved by stimulated emission. In this process, a signal photon of frequency \( \nu_s \) (such that \( h \nu_s = E_2 - E_1 \)) triggers a downward transition of a bound electron in the active material from state 2 (the excited state) to state 1 (the ground state), generating a second photon of the same frequency.

This process can be made to dominate over absorption of the signal (which promotes electrons from state 1 to state 2) only if the population of state 2 is made to exceed that of the ground state. This is achieved by pumping the material with a source of photons of frequency \( \nu_p \) and energy \( E_p = h \nu_p \). Absorption of these photons results in the promotion of electrons from the ground state to a third state with energy \( E_3 \) such that \( E_p = E_3 - E_1 \). The electrons then decay non-radiatively to state 2. If the pump beam is strong enough and the lifetime of state 2 is long enough, population inversion is obtained.

Figure 20. a) Stimulated emission in a 3-level amplifier; b) spontaneous emission
However, spontaneous emission of photons of the signal frequency also occurs, through the decay of electrons from state 2 to state 1 as shown in Figure 20b. This process acts as a source of optical noise, which is then amplified by stimulated emission.

Silica may be transformed into an active material by doping with rare earth ions. For operation at the common system wavelength of 1.530 µm, doping with Er³⁺ ions is used. Optical gain can be provided at this wavelength by pumping at 0.98 µm, which promotes electrons from the I⁰ 15/2 ground state of Er⁺⁺ to the I⁰ 11/2 state. Electrons in this state quickly decay to the I⁰ 13/2 metastable state. Stimulated transition from this state back to the ground state then provides the gain. Two possible sources can currently be used to generate the pump wavelength: for high powers, Ar⁺-pumped Ti: Al₂O₃ lasers are used, and for lower powers, specially developed semiconductor laser diodes.

The main problem with erbium doping is the difficulty of obtaining high concentrations of uniformly dispersed ions in the glass. Specifically, Er is known to form concentration clusters in a pure silica host, which limit the dopant concentration to \( \approx 100 \) ppm in either silica or germanosilicate glass. Already, this effect has been identified as a key difficulty in fibre-based amplifiers, necessitating the use of long fibre lengths (e.g. 90 metres) to achieve the required gain. In an integrated device, the problem is even more acute, and very careful attention must be paid to the host material.

It has already been demonstrated in fibre amplifiers that the use of alumina or phosphorus pentoxide co-doping can drastically reduce the clustering problem in silica. Clustering occurs because the rare-earth oxides (e.g. Er₂O₃) are generally insoluble in SiO₂. However, Er₂O₃ is soluble in Al₂O₃, which in turn is soluble in SiO₂. The Al₂O₃ can then form a solvation shell around the rare earth ion, and the resulting complex is readily incorporated into the silica network. The same principle is assumed to apply in P₂O₅ co-doping.

Waveguide devices can then be constructed using a four-component glass. For example, fibre-based amplifiers have been constructed using Er₂O₃: Al₂O₃: GeO₂: SiO₂. Here, two of the dopants - germania and alumina - are used to control the optical refractive index and the dispersion of the third dopant (the erbium ions) independently.

Similar principles have already been applied to planar devices. For example, NTT have already demonstrated Nd- and Er-doped silica waveguide lasers, and have constructed a 20 cm-long Er-doped amplifier with 15 dB gain at a pump power of 500 mW. In each case, a three-component glass was used. The waveguide material was SiO₂ co-doped with P₂O₅, formed by flame hydrolysis. The active materials were then introduced by soaking the soot in an alcohol solution containing a hydrated rare earth chloride before sintering, and Er concentrations of 5500 - 8000 ppm were achieved.

Other fabrication methods (e.g. PECVD and sol-gel) are now being adapted to allow the incorporation of erbium in deposited glasses. These may allow the incorporation of higher dopant concentrations, since they involve lower process temperatures. For example, Figure 21 shows the variation of transmission with wavelength, for a 5 cm long amplifier formed by erbium doping of aluminophosphosilicate glass formed by the sol-gel process. At low powers, there is significant absorption near \( \lambda = 1.535 \) µm, caused by the Er³⁺ ions. As the pump power increases, the absorption spectrum essentially inverts and becomes a gain spectrum. Near complete inversion is obtained, and the peak internal gain in this case was 1.35 dB/cm. A relatively short length of active guide (say, 10 –20 cm) may then provide enough gain for a low cost amplifier for use in Metro applications.
Figure 21. Spectral variation of gain at different pump powers, for a sol-gel EDWA (after W. Huang et al. IEEE Photon. Tech. Lett. 14, 959-961 (2002))
SUGGESTIONS FOR FURTHER READING


Suzuki S., Shuto K., Takahashi H., Hibino Y. "Large-scale and high-density planar lightwave circuits with high-Δ GeO₂-doped silica waveguides" Elect. Lett. 28, 1863-1864 (1992)


Day S., Bellerby R., Cannell, Grant M. "Silicon based fibre pigtailed 1 x 16 power splitter " Elect. Lett. 28, 920-922 (1992)


“Application of planar lightwave circuit platform to hybrid integrated optical WDM transmitter/ receiver module” Elect. Lett. 31, 1366-1367 (1995)
ANSWERS TO SELECTED PROBLEMS

CHAPTER 2

2.1 Start with Maxwell’s equations:
\[ \nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t \quad ; \quad \nabla \times \mathbf{H} = \mathbf{J} + \partial \mathbf{D}/\partial t \]

Then assume that in a dielectric medium \( \mathbf{J} = 0 \), \( \mathbf{B} = \mu_0 \mathbf{H} \) and \( \mathbf{D} = \varepsilon \mathbf{E} \)

Hence, we can say that:
\[ \nabla \times \mathbf{E} = -\mu_0 \partial \mathbf{H}/\partial t \quad (1) \]
\[ \nabla \times \mathbf{H} = \varepsilon \partial \mathbf{E}/\partial t \quad (2) \]

Now take the curl of Equation (2), to get:
\[ \nabla \times (\nabla \times \mathbf{H}) = \nabla \times (\varepsilon \partial \mathbf{E}/\partial t) \]

Furthermore, if we assume that \( \varepsilon \) is constant, we have:
\[ \nabla \times (\nabla \times \mathbf{H}) = \varepsilon \{ \nabla \times (\partial \mathbf{E}/\partial t) \} \quad (3) \]

Reversing the order of differentiation in Equation (3), we get:
\[ \nabla \times (\nabla \times \mathbf{H}) = \varepsilon \{ \partial (\nabla \times \mathbf{E})/\partial t \} \quad (4) \]

Substituting for \( \nabla \times \mathbf{E} \) from Equation (1) then gives:
\[ \nabla \times (\nabla \times \mathbf{H}) = -\mu_0 \varepsilon \partial^2 \mathbf{H}/\partial t^2 \]

Now, the following standard identity exists:
\[ \nabla \times (\nabla \times \mathbf{H}) = \nabla \cdot \nabla \mathbf{H} - \nabla^2 \mathbf{H} \]

Since \( \nabla \times \mathbf{B} = 0 \), it follows that
\[ \nabla (\mu_0 \mathbf{H}) = 0, \text{ so } \nabla \mathbf{H} = 0 \text{ and } \nabla \times (\nabla \times \mathbf{H}) = -\nabla^2 \mathbf{H} \]

Hence, the required vector wave equation in \( \mathbf{H} \) alone is:
\[ \nabla^2 \mathbf{H} = \mu_0 \varepsilon \partial^2 \mathbf{H}/\partial t^2 . \]

2.2 a) \( \nu_{ph} = c = f \lambda_0 \Rightarrow \lambda_0 = (3 \times 10^8 \times 2\pi) / (2.978 \times 10^{15}) \text{ m} = 0.633 \times 10^{-6} \text{ m}. \)

b) \( \lambda = \lambda_0/n = \lambda_0/\sqrt{\varepsilon} = 0.633 \times 10^{-6}/\sqrt{2.25} \text{ m} = 0.422 \times 10^{-6} \text{ m}. \)

2.3 \( \mathbf{E} = E_{y0} \exp\{-jk_0 (ax + bz)\} \hat{j} \Rightarrow \mathbf{E} = E_{y0} \exp\{-jk_0 n (x \sin\theta + z \cos\theta)\} \hat{j} \)

a) \( E_{y0} \) is the wave amplitude.

b) The E-vector points in the j-direction, so the wave is polarized in the j-direction.

c) There is no y-variation, so the wave is travelling in the x-z plane.

The angle of propagation relative to the x-axis is \( \theta = \tan^{-1}(a/b) = \tan^{-1}(1/\sqrt{3}) = 30^\circ. \)

d) \( n = a/\sin(\theta) = 1/0.5 = 2. \)
2.4 \( k^2 = \omega^2 \mu \varepsilon \Rightarrow k = \omega \sqrt{\mu \varepsilon} \), where \( k \) is the propagation constant. If:
\[ E_y = E_{y0} \exp(\gamma x) \exp(-j\beta z) \]

Then:
\[ \partial E_y / \partial x = \gamma E_{y0} \exp(\gamma x) \exp(-j\beta z) \quad \text{and} \quad \partial^2 E_y / \partial x^2 = \gamma^2 E_{y0} \exp(\gamma x) \exp(-j\beta z) \]

And
\[ \partial E_y / \partial z = -j\beta E_{y0} \exp(\gamma x) \exp(-j\beta z) \quad \text{and} \quad \partial^2 E_y / \partial x^2 = -\beta^2 E_{y0} \exp(\gamma x) \exp(-j\beta z) \]

If:
\[ \nabla^2 E_y + k^2 E_y = 0 \]

Then:
\[ \gamma^2 - \beta^2 + k^2 = 0, \text{ so } \gamma = \sqrt{\beta^2 - k^2}. \]

2.5 The wave is travelling in the z-direction; its amplitude varies exponentially in the x-direction. If \( \gamma \) is to be real, we must have \( \beta > k \). Hence, \( v_{ph} = \omega / \beta < \omega / k \), so the wave travels slower than a plane wave travelling in the z-direction. This type of wave is an evanescent or boundary wave.

2.6 The propagation constant is \( k' = \omega \sqrt{\mu_0 \varepsilon_0 \varepsilon_r} = 2\pi \sqrt{\varepsilon_r} / \lambda_0 = 2\pi \sqrt{2.25} / 0.633 \times 10^{-6} = 14.89 \times 10^6 \text{ m}^{-1} \). The attenuation coefficient is \( k'' = k' \varepsilon_r'' / (2\varepsilon_r') = 14.89 \times 10^6 \times 10^8 / (2 \times 2.25) = 0.0331 \text{ m}^{-1} \).

The wave propagates as \( E_x = E_{x0} \exp(k'z) \exp(-jk'z) \). The power carried by the wave is proportional to \( |E_x|^2 = E_{x0}^2 \exp(-2k'z) \). The power will therefore decay to 1/e of an initial value when \( 2k''z = 1 \), or when \( z = 1 / (2k'') = 1 / (2 \times 0.0331) = 15.11 \text{ m} \).
2.7 The time averaged Poynting vector is
\[ S = \frac{1}{T} \int_0^T (\mathbf{E} \times \mathbf{H}) \, dt \]

Put:
\[ \mathbf{E} = \text{Re}\{\mathbf{E} \exp(j\omega t)\} \text{ and } \mathbf{H} = \text{Re}\{\mathbf{H} \exp(j\omega t)\} \]

Now:
\[ \text{Re}(z) = \frac{1}{2} \{z + z^*\} \]

Hence:
\[ S = \frac{1}{4T} \int_0^T \{ (\mathbf{E} x \mathbf{H}) \exp(2j\omega t) + (\mathbf{E}^* x \mathbf{H}^*) \exp(-2j\omega t) + (\mathbf{E} x \mathbf{H}^*) + (\mathbf{E}^* x \mathbf{H}) \} \, dt \]

Or:
\[ S = \frac{1}{2T} \text{Re} \left[ \int_0^T \{ (\mathbf{E} x \mathbf{H}) \cos(2\omega t) + (\mathbf{E} x \mathbf{H}^*) \} \right] \]

The first term in the integral averages to zero, if \( T \) is long enough, leaving:
\[ S = \frac{1}{2} \text{Re}(\mathbf{E} x \mathbf{H}^*) \]

2.8 Assume that:
\[ \mathbf{E} = E_y \exp(-jk_0 n [x \sin(\theta) + z \cos(\theta)]) \]

Then:
\[ \nabla \times \mathbf{E} = -\partial E_y / \partial z \hat{i} + \partial E_y / \partial x \hat{k} = -j\omega \mu_0 \mathbf{H} \]

Hence:
\[ H_x = -[k_0 n \cos(\theta)/\omega \mu_0] E_y ; \quad H_y = 0 ; \quad H_z = [k_0 n \sin(\theta)/\omega \mu_0] E_y \]

So:
\[ S = \frac{1}{2} \text{Re}(\mathbf{E} x \mathbf{H}^*) = \frac{1}{2} \text{Re}\{(E_yH_z^*) \hat{i} - (E_yH_x^*) \hat{k}\} \]

Substituting relevant components of \( \mathbf{E} \) and \( \mathbf{H} \) then gives:
\[ S = (k_0 n E_{y0}^2/2\omega \mu_0) \{ \sin(\theta) \hat{i} + \cos(\theta) \hat{k} \} \]

Or:
\[ S = 1/2 \sqrt{\epsilon/\mu_0} E_{y0}^2 \{ \sin(\theta) \hat{i} + \cos(\theta) \hat{k} \} \]

Hence:
\[ S_z = 1/2 \sqrt{\epsilon/\mu_0} E_{y0}^2 \cos(\theta) \]

Given that \( \sqrt{\epsilon/\mu_0} = 377 \Omega \), \( n = \sqrt{\epsilon_r} = 2 \) and \( \theta = 30^\circ \), \( S_z = 2.3 \times 10^3 E_{y0}^2 \text{ W/m}^2 \).

The direction of maximum irradiance is parallel to the vector \( \sin(\theta) \hat{i} + \cos(\theta) \hat{k} \); the maximum irradiance is given by \( |S| = 1/2 \sqrt{\epsilon/\mu_0} E_{y0}^2 = 2.65 \times 10^3 E_{y0}^2 \text{ W/m}^2 \).

2.9 \[ |S| = 1/2 \sqrt{\epsilon/\mu_0} E^2 = 1000 \text{ W/m}^2. \] Hence \( E = \sqrt{(2000 \times 377)} \text{ V/m} = 868.3 \text{ V/m} \).

2.10 If the dispersion characteristic is \( \omega = \sqrt{[\omega_p^2 + c^2k^2]} \), then:
\[ v_{ph} = \omega/k = \sqrt{[(\omega_p^2/k^2) + c^2]} \] and \[ v_g = d\omega/dk = c^2/\sqrt{[(\omega_p^2/k^2) + c^2]} \].

When \( \omega \rightarrow \omega_p \), \( k \rightarrow 0 \), so \( v_g \rightarrow 0 \). This implies that the velocity of propagation of information will tend to zero as \( \omega \rightarrow \omega_p \).

Rearranging the dispersion characteristic, we get:
\[ k = (1/c) \sqrt{(\omega^2 - \omega_p^2)} \].

When \( \omega < \omega_p \), \( k \) must be purely imaginary, so we can write: \( k = -jk'' \). Hence, a z-propagating wave will have the form \( E = E_0 \exp[j(\omega t - kz)] = E_0 \exp(j\omega t) \exp(-k''z) \). This solution does not represent a travelling wave, but one that decays exponentially with distance.
CHAPTER 3

3.1 At $\lambda_0 = 1.3 \ \mu m$:

\[ n' = 1 + 0.6961663 \frac{(1.3^2 - 0.0684043^2)}{(1.3^2 - 0.1162414^2) + 0.8974794 \times 1.3^2 / (1.3^2 - 9.8961610^2)} \Rightarrow n' = 1.44969175. \]

At $\lambda_0 = 1.55 \ \mu m$:

\[ n' = 1 + 0.6961663 \frac{(1.55^2 - 0.0684043^2)}{(1.55^2 - 0.1162414^2) + 0.8974794 \times 1.55^2 / (1.55^2 - 9.8961610^2)} \Rightarrow n' = 1.4440236. \]

3.2 For a single resonance, Sellmeier’s equation reduces to:

\[ n'^2 = 1 + \frac{A_1}{1 - \left(\frac{\lambda_0}{\lambda_1}\right)^2}, \]

Assuming that $\lambda_0 \gg \lambda_1$ ($\lambda_0$ a visible wavelength, $\lambda_1$ in the UV), this may be expanded as a power series:

\[ n'^2 = 1 + \left(\frac{\lambda_1}{\lambda_0}\right)^2 + \left(\frac{\lambda_1}{\lambda_0}\right)^4 \ldots \]

For small $A_1$, we may then approximate this as:

\[ n' \approx 1 + \frac{A_1}{2} \left(1 + \left(\frac{\lambda_1}{\lambda_0}\right)^2 + \left(\frac{\lambda_1}{\lambda_0}\right)^4 \ldots \right) \]

This has the same general form as Cauchy’s equation $n' = A + \frac{B}{\lambda_0^2} + \frac{C}{\lambda_0^4}$, with $A = 1 + \frac{A_1}{2}$, $B = A_1 \lambda_1^2 / 2$ and $C = A_1 \lambda_1^4 / 2$. Since Cauchy’s equation contains only even powers of $\lambda_0$, it cannot describe the whole of the dispersion curve, e.g. the region of anomalous dispersion.

3.3 $N = \frac{\omega_p^2 m_e}{\varepsilon_0^2} = (12.2 \times 10^{15})^2 \times 9.1 \times 10^{-31} \times 8.85 \times 10^{-12} / (1.6 \times 10^{-19})^2 = 4.68 \times 10^{28} / m^3$.

The real part of the relative dielectric constant of a metal is given by:

\[ \varepsilon' = 1 - \frac{\omega_p^2}{\omega^2} \]

for $\varepsilon' < 0$, we require $\omega < \omega_0$ or $\omega < \sqrt{(12.2 \times 10^{15})^2 - (0.09 \times 10^{15})^2} = 12.1996 \times 10^{15}$

Hence, $\lambda_0 > 2 \pi / \omega = 1.545 \times 10^{-7} \ m$.

3.4 The behaviour of an electron in the ionised gas is essentially similar to that in a metal, except that collision damping may be neglected (since there is no lattice, and the gas is low-pressure). Hence, we may deduce from previous results that $\varepsilon' = 1 - \frac{\omega_p^2}{\omega^2}$.

Since $k^2 = \omega^2 \mu_0 \varepsilon$, we then get $k^2 = \omega^2 \mu_0 \varepsilon_0 \{1 - \omega_0^2 / \omega^2\}$. Now, $c = 1/\sqrt{\mu_0 \varepsilon_0}$, so this expression may be rearranged to give the required result, $\omega = \sqrt{\omega_0^2 + c^2 k^2}$.

3.5 Start by assuming that if contributions from bound and free electrons are both included, the relative dielectric constant can be written as $\varepsilon = \varepsilon_{\text{ne}} - j \sigma / \varepsilon_0 \omega$, where $\sigma = N \varepsilon_i / (r + j m^* \omega)$. Hence, $\varepsilon_i = \varepsilon_{\text{ne}} - j N e^2 / \{m^* \varepsilon_0 (\omega_i^2 + j \omega^2)\}$, where $\omega_i = r / m^*$.

Separating this result into real and imaginary parts, we get :

\[ \varepsilon_i = \varepsilon_{\text{ne}} - N e^2 / \{ m^* \varepsilon_0 (\omega_i^2 + \omega^2) \} - j N e^2 \omega_i / \{ m^* \varepsilon_0 (\omega_i^2 + \omega^2) \omega \} = \varepsilon'_i - j \varepsilon''_i. \]
The introduction of free carriers therefore results in a non-zero value of ε'';
consequently, if an attempt is made to modify the refractive index of a semiconductor
by free carrier injection, there will also be significant optical loss.

3.6 Relative retardation of the two polarization components by the birefringent wave plate
is \( \phi_x - \phi_y = 2\pi(n_0 - n_e)d/\lambda_0 = 2\pi (2.286 - 2.2) \times 3.684 \times 10^{-3} / 0.633 \times 10^{-6} = 2\pi (m + 0.5) \)
where \( m = 500 \Rightarrow \) birefringent plate is a half-wave plate, whose action is to rotate
the input polarization by 2\( \theta \), where \( \theta \) is the rotation of plate relative to the x-axis. If
transmission is initially maximum, the input polarization and the axis of the polarizer
must be aligned parallel to x-axis. Hence:

(a) If \( \theta = 22.5^o \), the polarization will be rotated by 45\(^o\), so the resulting field will have
equal components parallel and perpendicular to the polarizer axis. The perpendicular
component will be blocked, giving a transmission of 50%.

(b) If \( \theta = 45^o \), the polarization will be rotated by 90\(^o\), and so will be entirely
perpendicular to the axis of the polarizer.

3.7 If the coordinates are represented by a radius vector \( r \), and the B-coefficients are
assembled into a symmetric, second-rank tensor \([B]\), we may write:

\[
\begin{bmatrix}
  x \\
  y \\
  z
\end{bmatrix} =
\begin{bmatrix}
  B_{xx} & B_{xy} & B_{xz} \\
  B_{xy} & B_{yy} & B_{yz} \\
  B_{xz} & B_{yz} & B_{zz}
\end{bmatrix} \begin{bmatrix}
  x \\
  y \\
  z
\end{bmatrix}
\]

\[
\begin{bmatrix}
  x \\
  y \\
  z
\end{bmatrix} = \begin{bmatrix}
  B_{x}x + B_{y}y + B_{z}z \\
  B_{xy}x + B_{yy}y + B_{yz}z \\
  B_{xz}x + B_{yz}y + B_{zz}z
\end{bmatrix}
\]

Hence, if \( B_{xx} x^2 + B_{yy} y^2 + B_{zz} z^2 + 2B_{xy} xy + 2B_{yz} yz + 2B_{zx} zx = 1 \),
then \( r^T [B] r = 1 \).

3.8 The individual eigenvalues and eigenvectors of the second-rank tensor \([B]\) are defined
by three matrix/vector equations of the form \([B] \chi_i = \lambda_i \chi_i \). These may all be combined
into a single matrix equation \([B] [V] = [V] [\Lambda] \), where \([\Lambda]\) is a diagonal matrix whose
elements are \( \lambda_i \), and \([V]\) contains the vectors \( \chi_i \) arranged in columns.

Post-multiplying by \([V]^{-1} \), we then obtain:

\( [B] [V] [V]^{-1} = [B] = [V] [\Lambda] [V]^{-1} \).

However, since the eigenvectors \( \chi_i \) are orthonormal, we note that \([V]^{-1} = [V]^T \), so the
matrix \([B]\) can be written \([B] = [V] [\Lambda] [V]^T \). Consequently, the equation for the index
ellipsoid may be written in the new quadratic form \( r^T [B] r = r^T [V] [\Lambda] [V]^T r = 1 \).
This may be written as \( r^T [\Lambda] r' = 1 \), if a new set of co-ordinates are defined according
to \( r' = [V]^T r \). The equation of the index indicatrix is thus \( \lambda_1 x'^2 + \lambda_2 y'^2 + \lambda_3 z'^2 = 1 \).
This is the equation of an ellipsoid aligned to the \( x', y' \) and \( z' \) axes, and with semi-
major axes of lengths \( 1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2} \) and \( 1/\sqrt{\lambda_3} \).

3.9 KDP is a tetragonal crystal, of point group 42m. Consequently, the only significant
electro-optic coefficients are $r_{41}$ and $r_{63}$. For an electric field applied in the $z$-direction, the changes to the $B$-coefficients are given by:

\[
\begin{bmatrix}
\Delta B_1 \\
\Delta B_2 \\
\Delta B_3 \\
\Delta B_4 \\
\Delta B_5 \\
\Delta B_6 \\
\end{bmatrix} =
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
r_{41} & 0 & 0 & 0 \\
0 & r_{41} & 0 & 0 \\
0 & 0 & r_{63} & 0 \\
\end{bmatrix}
\begin{bmatrix}
E_z \\
\end{bmatrix}
\]

Only the coefficient $B_6$ (which is the same as $B_{xy}$) changes, by an amount $r_{63}E_z$.

Assuming that the equation of the unperturbed index ellipsoid was $B_{xx}x^2 + B_{yy}y^2 + B_{zz}z^2 = 1$, the new equation is $B_{xx}x^2 + B_{yy}y^2 + B_{zz}z^2 + 2r_{63}E_z xy = 1$.

3.10 The equation of the unperturbed index ellipsoid may be written in the quadratic form of Questions 3.7 and 3.8, with the second-rank tensor $[B]$ given by:

\[
[B] =
\begin{bmatrix}
B_{xx} & 0 & 0 \\
0 & B_{yy} & 0 \\
0 & 0 & B_{zz} \\
\end{bmatrix}
\]

After application of the electric field, the new tensor $[B']$ has the form:

\[
[B'] =
\begin{bmatrix}
B_{yy} & 2r_{63}E_z & 0 \\
2r_{63}E_z & B_{yy} & 0 \\
0 & 0 & B_{zz} \\
\end{bmatrix}
\]

Now, the eigenvalues of $[B']$ are defined by the determinantal equation:

\[
\begin{vmatrix}
B_{yy} - \lambda & 2r_{63}E_z & 0 \\
2r_{63}E_z & B_{yy} - \lambda & 0 \\
0 & 0 & B_{zz} - \lambda \\
\end{vmatrix} = 0
\]

which may be expanded to give:

\[
(B_{xx} - \lambda)(B_{yy} - \lambda)(B_{zz} - \lambda) - 4r_{63}^2E_z^2(B_{zz} - \lambda) = 0.
\]

or:

\[
(B_{xx} - \lambda)\{\lambda^2 - (B_{xx} + B_{yy})\lambda + (B_{xx}B_{yy} - 4r_{63}^2E_z^2)\} = 0
\]

To first-order in $E_z$, this has the solutions $\lambda = B_{xx}, B_{yy}, B_{zz}$. Consequently, to a reasonable approximation the eigenvalues are unaltered by the application of the field, so the values $1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}$ and $1/\sqrt{\lambda_3}$ (the lengths of the semi-major axes of the index ellipsoid) are also unchanged.
CHAPTER 4

4.1 Use the lens maker’s formula \(1/f = (n - 1) \left(1/r_1 + 1/r_2\right)\) with \(n = 1.5\).
(a) \(1/f = 0.5(1/200 + 1/200) = 1/200 \Rightarrow f = 200\) mm.
(b) \(1/f = 0.5(1/400 + 1/\infty) = 1/800 \Rightarrow f = 800\) mm.

4.2 Use the imaging formula \(1/u + 1/v = 1/f \Rightarrow 1/v = 1/f - 1/u\).
(a) \(1/v = 1/200 - 1/400 = 1/400 \Rightarrow v = 400\) mm (real image).
(b) \(1/v = 1/800 - 1/400 = -1/800 \Rightarrow v = -800\) mm (virtual image).

4.3 First define two co-ordinate systems \((x, y, z)\) and \((x', y', z')\) as shown below. The former is centred axially, the latter on the off-axis image point.

Now, from previous results, a paraxial spherical wave is defined in \((x', y', z')\) coordinates by \(E(x', y', z') = A(z') \exp\{-jk_0 (x'^2 + y'^2)/2z'\}\). However, \(x' = x\), \(y' = y - \delta\) and \(z' = z\), so we can also write \(E(x, y, z) = A(z) \exp\{-jk_0 [x^2 + (y - \delta)^2]/2z]\). If \(\delta\) is small, we may approximate this by

\[E(x, y, z) \approx A(z) \exp\{-jk_0 (x^2 + y^2 - 2y\delta)/2z\} \approx A(z) \exp\{-jk_0 (x^2 + y^2)/2z\} \exp\{jk_0 y\delta/z\}\]

Hence, the input to lens is \(E(x, y, u) = A(u) \exp\{-jk_0 (x^2 + y^2)/2u\} \exp\{jk_0 y\delta/u\}\). Now, the lens transform is \(\tau_L = \tau_s \exp\{jk_0 (x^2 + y^2)/2f\}\), so the output from the lens can be found as:

\[E(x, y, u) \tau_L = A'(u) \exp\{jk_0 (x^2 + y^2) (1/f - 1/u)/2\} \exp\{jk_0 y\delta/u\}\]

where \(\delta' = \delta/v/u\) and \(A'(u) = A(u)\tau_s\). Comparison between the expressions for the input to and the output from the lens shows that the output is a converging spherical wave, focused on a point \(v\) away from the lens, \(\delta'\) below the axis. Since the result is independent of \(\delta\), the magnification of the set-up can be found for a general image as \(M = \delta'/\delta = v/u\).

4.4 Lens waveguide geometry must be as shown below. Hence \(d_1 = 2u, d_2 = 2v\). Now, the imaging formula requires that \(1/u + 1/v = 1/f \Rightarrow\) the relation that must be satisfied by the lens separations is \(1/d_1 + 1/d_2 = 1/(2f)\).
4.5 In the far-field, the intensity distribution in the \((\xi, \eta)\) plane can be related to the amplitude in the \((x, y)\) plane by:

\[ |A(\xi, \eta)|^2 = \left( \frac{1}{\lambda_0^2 z^2} \right) \int \int A(x, y) \exp\{+j(\omega_x x + \omega_y y)\} \, dx \, dy \]

where \(\omega_x\) and \(\omega_y\) are spatial frequencies given by \(\omega_x = k_0 \xi / z\) and \(\omega_y = k_0 \eta / z\).

Hence, assuming an input wave of amplitude \(A\), the intensity distribution for the aperture is:

\[ |A(\xi, \eta)|^2 = \left( \frac{A^2}{\lambda_0^2 z^2} \right) \int_{-b/2}^{b/2} \int_{-a/2}^{a/2} \exp\{+j(\omega_x x + \omega_y y)\} \, dx \, dy \]

\[ = \left( \frac{a b A^2}{16 \lambda_0^2 z^2} \right) \text{sinc}^2(\omega_x a/2) \times \text{sinc}^2(\omega_y b/2). \]

4.6 Assuming an input wave amplitude \(A_0\) and adopting a transparency function \(\tau = \tau_0 + \Delta \tau \cos(Ky)\) for the screen, we may put:

\[ A(x, y) = A_0 \{\tau_0 + \Delta \tau \cos(Ky)\} = A_0 \{\tau_0 + (\Delta \tau/2) \{\exp(jKy) + \exp(-jKy)\}\} \]

The far-field diffraction integral will then average to zero, unless \(\omega_x = 0\) and \(\omega_y = -K, 0\) or \(+K\). Consequently, the output will consist of a fan of three beams, travelling in directions defined by \(\eta/z = LK/k_0\) with \(L = -1, 0\) or \(+1\), as shown in the construction below. These are known as diffraction orders.


4.8 (a) \(\theta = 2/k_0 w_0 = 2 / \{(2\pi/0.633 \times 10^{-6}) \times (10^{-3}/2)\} = 4.0298 \times 10^4\) rad.
(b) \(z_0 = k_0 w_0^2/2 = (2\pi/0.633 \times 10^{-6}) \times (10^{-3}/2)^2/2 = 1.2407\) m.

4.9 The lasing threshold is reached when the round-trip gain \(R_1 R_2 \exp(2gL)\) equals unity. Hence,

\[ g_{\text{th}} = (1/2L) \log_e\{1/(R_1 R_2)\} = 1/0.5 \log_e(1/0.9) = 0.2107\] m\(^{-1}\).
Assuming that the refractive index of the low-pressure gas plasma forming the gain medium is approximately unity, the mode separation is $\Delta f = c/2L = 3 \times 10^8/0.5 = 600$ MHz.

4.10 The Gaussian mode is just one of the possible transverse modes of this type of cavity; the others are Hermite-Gaussian modes, which generally have most of their energy distributed further from the optical axis. Consequently, if a circular aperture is inserted into the cavity, these modes will be heavily attenuated and so will have a high lasing threshold. The Gaussian mode, on the other hand, has most of its energy concentrated on-axis, and will be affected only slightly.

The radius of curvature of a Gaussian beam at a distance $z$ from the beam waist is given by $r(z) = (z^2 + z_0^2)/z$.

This must match the radius of curvature of the spherical mirror, when $z$ is equal to the cavity length. Hence: $1 + z_0^2 = 2 \Rightarrow z_0 = 1$ m.

The beam waist is then given by $w_0 = \sqrt{2z_0/k_0} \Rightarrow w_0 = \sqrt{2 \times 1 \times 0.633 \times 10^{-6}/(2\pi)} = 4.488 \times 10^{-4}$ m, and the beam diameter $d$ is then $2w_0 = 0.8977$ mm.
CHAPTER 5

5.1  
   a) \( \theta_c = \sin^{-1}(n_2/n_1) = \sin^{-1}(1/1.7) = 36.03^\circ \).
   
   b) \( \theta_B = \tan^{-1}(n_2/n_1) = \tan^{-1}(1/1.7) = 30.46^\circ \).

5.2 \( \theta_c = \sin^{-1}(n_2/n_1) = 30^\circ \Rightarrow n_2/n_1 = 0.5. \)

   At normal incidence, \( T_E = T_M = 2n_1 / (n_1 + n_1) = 2 / (1 + n_1/n_1) = 2/1.5 = 1.333. \)

5.3 \( \theta_1 = 30^\circ \Rightarrow \theta_2 = \sin^{-1}\left(\frac{n_1}{n_2}\sin(\theta_1)\right) = 19.47^\circ. \)

   \[ \Gamma_E = \frac{n_1 \cos(\theta_1) - n_2 \cos(\theta_2)}{n_1 \cos(\theta_1) + n_2 \cos(\theta_2)} \]
   
   \[ = \frac{0.866 - 1.5 \times 0.9428}{0.866 + 1.5 \times 0.9428} = -0.2404 \]

   \[ \Rightarrow |\Gamma_E|^2 = 0.05779. \]

   \[ T_E = 2n_1 \cos(\theta_1) / \left[n_1 \cos(\theta_1) + n_2 \cos(\theta_2)\right] \]

   \[ = 2 \times 0.866 / [0.866 + 1.5 \times 0.9428] = 0.7596 \]

   \[ \Rightarrow |T_E|^2 = 0.57699. \]

5.4 If light rays are reversible, it should be possible to make the following situation arise:

We can construct this by superposition of the sets of fields that make up the following diagrams, which define the transmission and reflection coefficients for incidence from media 1 and 2:

Multiplying all wave amplitudes in diagram (b) by \( T \), and all waves in (c) by \( \Gamma \) then gives two new diagrams:
Now, by superposition, we may add the waves in diagrams (d) and (e) together, to get:

Since diagram (f) now has the same waves incident on the interface from the right as diagram (a), the waves leaving on the left must also be the same. Comparing amplitudes, we then get: $\Gamma^2 + T'T = 1$ (the first required result) and $T\Gamma + \Gamma'T = 0$, or $\Gamma + \Gamma' = 0$ (the second).

5.5 Proof of the results of Question 5.4: for TE incidence, the relevant coefficients are given by:

\[
\Gamma = \frac{[n_1 \cos(\theta_1) - n_2 \cos(\theta_2)\] / [n_1 \cos(\theta_1) + n_2 \cos(\theta_2)]
\]

\[
\Gamma' = \frac{[n_2 \cos(\theta_2) - n_1 \cos(\theta_1)\] / [n_1 \cos(\theta_1) + n_2 \cos(\theta_2)]
\]

$\Rightarrow \Gamma + \Gamma' = 0$ by inspection.

\[
T = \frac{2n_1 \cos(\theta_1)}{[n_1 \cos(\theta_1) + n_2 \cos(\theta_2)]
\]

\[
T' = \frac{2n_2 \cos(\theta_2)}{[n_1 \cos(\theta_1) + n_2 \cos(\theta_2)]
\]

$\Rightarrow \Gamma^2 + T'T = \frac{[n_1 \cos(\theta_1) - n_2 \cos(\theta_2)]^2 + [2n_1 \cos(\theta_1) \cdot 2n_2 \cos(\theta_2)]}{[n_1 \cos(\theta_1) + n_2 \cos(\theta_2)]^2} = 1$

5.6 The amplitude for Path 3 is $T_{13} \Gamma_{32} \Gamma_{31} \Gamma_{32} T_{31} \exp(-j4k_0 n_3 d)$. Hence, the total reflectivity may be written as an infinite series, in the form:

\[
\Gamma = \Gamma_{13} + T_{13} \Gamma_{32} T_{31} \exp(-j2k_0 n_3 d) + T_{13} \Gamma_{32} \Gamma_{31} T_{31} \exp(-j4k_0 n_3 d) + \text{higher order paths}
\]

$\Rightarrow \Gamma = \frac{\Gamma_{13} + \Gamma_{32} [T_{13} T_{31} - \Gamma_{13} \Gamma_{31}] \exp(-j2k_0 n_3 d)}{1 - \Gamma_{31} \Gamma_{32} \exp(-j2k_0 n_3 d)}$

At this point, we make use of the relations derived in Question 5.4, namely that $\Gamma_{13} = -\Gamma_{31}$ and that $T_{13} T_{31} + \Gamma_{13}^2 = 1$. This allows us to simplify the expression above to:

\[
\Gamma = \frac{\Gamma_{13} + \Gamma_{32} \exp(-j2k_0 n_3 d)}{1 - \Gamma_{31} \Gamma_{32} \exp(-j2k_0 n_3 d)}
\]
We now choose to set the thickness \( d \) equal to \( \frac{\lambda_0}{4n_3} \), so that \( \exp(-j2k_0n_3d) = \exp(-j\pi) = -1 \). Hence, the reflectivity becomes: \( \Gamma = \frac{\{\Gamma_{13} - \Gamma_{32}\}}{\{1 + \Gamma_{31}\}\Gamma_{32}} \).

Clearly, the reflectivity can be zero, if \( \Gamma_{13} = \Gamma_{32} \). Using the Fresnel coefficients, this requires:

\[
\frac{n_1 - n_3}{n_1 + n_3} = \frac{n_3 - n_2}{n_3 + n_2}, \text{ or } \frac{n_1 - n_3}{n_3 - n_2} = \frac{n_1 + n_3}{n_1 + n_1}.
\]

Multiplying out the brackets, it is then simple to show that \( n_3^2 = n_1n_2 \), or \( n_3 = \sqrt{n_1n_2} \).

A quarter-wave dielectric layer of refractive index \( n_3 = \sqrt{n_1n_2} \) may therefore be used to eliminate reflections at the interface between two media of refractive indices \( n_1 \) and \( n_2 \), at least for wavelengths close to the design value. This is known as an **antireflection** coating.

### 5.8

The figure below shows the definitions of the relevant angles and refractive indices.

![Diagram showing definitions of relevant angles and refractive indices](image)

For the first interface, there will be no reflection, provided incidence is at the Brewster angle, \( \theta_B = \tan^{-1}(n_2/n_1) \). This angle can be represented geometrically as shown in (a) below.

Now, the wave is refracted as it enters Medium 2. According to Snell's law, \( \theta_B' \) is given by: \( n_1 \sin(\theta_B) = n_2 \sin(\theta_B') \). Hence:

\[
n_1n_2 / \sqrt{n_1^2 + n_2^2} = n_2 \sin(\theta_B') \Rightarrow \sin(\theta_B') = n_1 / \sqrt{n_1^2 + n_2^2}
\]

\( \theta_B' \) can therefore be represented geometrically as in (b) above. Therefore, \( \tan(\theta_B') = n_1/n_2 \), so the wave is incident on the second interface at the Brewster angle for that interface. Consequently, there will be no reflection at either interface.

### 5.9

a) For TE incidence,

\[
\Gamma^2 = \frac{\{n_1^2\cos^2(\theta_1) - 2n_1n_2\cos(\theta_1)\cos(\theta_2) + n_2^2\cos^2(\theta_2)\}}{\{n_1^2\cos^2(\theta_1) + 2n_1n_2\cos(\theta_1)\cos(\theta_2) + n_2^2\cos^2(\theta_2)\}}
\]
\[ T^2 = 4n_1^2 \cos^2(\theta_1) / \{ n_1^2 \cos^2(\theta_1) + 2n_1n_2 \cos(\theta_1)\cos(\theta_2) + n_2^2 \cos^2(\theta_2) \} \]
\[ \Rightarrow T^2 = n_2 \cos(\theta_2) / n_1 \cos(\theta_1) = 4n_1 \cos^2(\theta_1) \]
\[ n_2 \cos^2(\theta_2) / n_1 \cos(\theta_1) + 2n_1n_2 \cos(\theta_1)\cos(\theta_2) + n_2^2 \cos^2(\theta_2) \}
\[ \Rightarrow \Gamma^2 + T^2 \frac{n_2 \cos(\theta_2)}{n_1 \cos(\theta_1)} = 1 \text{ by inspection.} \]

b) Substitute previous data into a):
\[ 0.05779 + 0.57699 \times 1.5 \times \cos(19.47^\circ) / \cos(30^\circ) = 1. \]

5.10 The time-averaged power per unit area carried by a plane wave is \( P = 1/2 \frac{nE^2}{Z_0} \Rightarrow \)

The power carried by the input wave through a length \( S \) of the boundary is:
\[ P_I = 1/2 \frac{n_1 E_i^2}{Z_0} S \cos(\theta_1) / Z_0. \]

Similarly, the reflected power is:
\[ P_R = 1/2 \frac{n_1 E_R^2}{Z_0} S \cos(\theta_1) / Z_0 \]

and the transmitted power is:
\[ P_T = 1/2 \frac{n_2 E_T^2}{Z_0} S \cos(\theta_2) / Z_0 \]

Since there is no loss, we must have: \( P_I = P_R + P_T \). Hence:
\[ E_i^2 = E_R^2 + E_T^2 \frac{n_2 \cos(\theta_2)}{n_1 \cos(\theta_1)} \text{ or } \Gamma^2 + T^2 \frac{n_2 \cos(\theta_2)}{n_1 \cos(\theta_1)} = 1. \]
CHAPTER 6

6.1 From previous results, we know that the propagation constant is given by $\beta = k_0 \sin(\theta)$, where $\theta$ is the ray angle. Similarly, the eigenvalue equation requires that $2k_0 h \cos(\theta) = 2\nu \pi$. The dispersion relation can therefore be found as:

$$\beta^2 = k_0^2 \{1 - (\nu \pi/k_0 h)^2\} = \left(\frac{\omega}{c}\right)^2 \{1 - (\nu \pi/c/\omega)^2\}.$$  

For large $\beta$, $\omega \to c \beta$ for all modes. For small $\beta$, the individual modes cut off when $\omega = \nu \pi/c/h$.

6.2 For dielectric-filled metal-walled guide, the eigenvalue equation modifies to $2k h \cos(\theta) = 2\nu \pi$. Cutoff occurs when $\theta \to 0$, i.e. when $kh = \nu \pi$. Hence, all modes are cut off when $k < \pi/h$ or when $2\pi n f/c < \pi/h$, so $f < c/(2n h) = 2 \times 10^{14}$ Hz.

6.3 The eigenvalue equation for an asymmetric slab dielectric waveguide is given by:

$$\tan(\kappa h) = \kappa (\gamma + \delta) / (\kappa^2 - \gamma \delta).$$

If the guide is symmetric, $\gamma = \delta 

$$\Rightarrow \tan(\kappa h) = 2\kappa \gamma / (\kappa^2 - \gamma^2)$$

$$\Rightarrow \tan(\kappa h) = 2(\gamma/\kappa) / \{1 - (\gamma/\kappa)^2\}$$

Now, $\tan(A + B) = (\tan A + \tan B) / (1 - \tan A \tan B)$

$$\Rightarrow \tan(\kappa h) = 2 \tan(\kappa h/2) / (1 - \tan^2 (\kappa h/2))$$

Comparison with the result above shows that we must then have $\tan(\kappa h/2) = \gamma/\kappa$.

6.5 Assume the geometry of Figure 6.2-1. For the TM case, it is simpler to work with $H$  

$$\Rightarrow$$ solve $\nabla^2 H + n_i k_0^2 H = 0$ in each of the three layers. Assume $H = H_i(x) \exp(-j\beta z) j$  

$$\Rightarrow$$ waveguide equation is $d^2 H_i/dx^2 + \{n_i^2 k_0^2 - \beta^2\}H_i = 0$ in each layer.

Now guess the solutions:

$$H_1 = H \cos(\kappa x - \phi)$$

$$H_2 = H' \exp(\gamma x)$$

$$H_3 = H'' \exp[-\delta(x - h)]$$

where:

$$\kappa = \sqrt{(n_1^2 k_0^2 - \beta^2)}$$

$$\gamma = \sqrt{(\beta^2 - n_2^2 k_0^2)}$$

$$\delta = \sqrt{(\beta^2 - n_3^2 k_0^2)}$$

Boundary matching:

Tangential components of $H$ must match

$$\Rightarrow H_1 = H_2$$ on $x = 0$ and $H_1 = H_3$ on $x = h$  \hspace{1cm} (1)

Tangential components of $E$ must match. Now, $E$ is given by

$$E = (1/j\omega e) \nabla \times H,$$ or

$$E = (1/j\omega) \{-\partial H_j/\partial z \hat{j} + \partial H_j/\partial x \hat{k}\}$$

$$\Rightarrow$$ $E_z$ is only tangential component, and matching yields:

$$(1/\varepsilon_i) dH_j/dx = (1/\varepsilon_z) dH_j/dx$$ on $x = 0$ and
\[ \frac{1}{\varepsilon_1} \frac{dH_1}{dx} = \frac{1}{\varepsilon_3} \frac{dH_3}{dx} \text{ on } x = h \] (2)

From (1) we have \( H \cos(\phi) = H' \) (3)
and \( H \cos(\kappa h - \phi) = H'' \) (4)

From (2) we have \( (H \kappa n_1^2) \sin(\phi) = H' \gamma n_2^2 \) (5) and \( (H \kappa n_1^2) \sin(\kappa h - \phi) = H' \delta n_3^2 \) (6)

From (3) and (5) we then get:
\[ \tan(\phi) = \left( n_1^2/n_2^2 \right) \gamma \kappa = \gamma' / \kappa \] (7)

From (4) and (6) we then get:
\[ \tan(\kappa h - \phi) = \left( n_1^2/n_3^2 \right) \delta / \kappa = \delta' / \kappa \] (8)

Reduction of Equations (7) and (8) to standard form
\[ \tan(\kappa h) = \kappa (\gamma' + \delta') / (\kappa^2 - \gamma^2') \]
then follows identical procedure to that used in TE case.

6.6 From the results of Question 6.5, we know that the transverse magnetic field must match at each boundary, but the slope of the transverse magnetic field does not, since, for a symmetric guide, \( (1/n_1^2) \frac{dH_1}{dx} = (1/n_2^2) \frac{dH_2}{dx} \) on \( x = 0 \) and \( x = h \).

Because \( n_1 > n_2 \), \( dH_1/dx > dH_2/dx \) at each boundary. Both fields then have the general form shown below, with greater penetration into the substrate by the evanescent tails in case (b).

6.7 The eigenvalue equation is \( \tan(\kappa h/2) = \gamma / \kappa \) for symmetric modes
\[ \Rightarrow \text{ when } \beta = n_{eff} k_0 \text{ has the correct value, the following equation is satisfied:} \]
\[ \tan\left( \frac{h}{2} \sqrt{[n_1^2 k_0^2 - \beta^2]} \right) - \sqrt{\frac{[\beta^2 - n_2^2 k_0^2]}{[n_1^2 k_0^2 - \beta^2]}} = 0 \]

For any other value of \( \beta \),
\[ \tan\left( \frac{h}{2} \sqrt{[n_1^2 k_0^2 - \beta^2]} \right) - \sqrt{\frac{[\beta^2 - n_2^2 k_0^2]}{[n_1^2 k_0^2 - \beta^2]}} = \alpha \neq 0 \] (1).

Solve by iteration, assuming that \( n_2 < n_{eff} < n_1 \).

For any other value of \( \beta \),
\[ \tan\left( \frac{h}{2} \sqrt{[n_1^2 k_0^2 - \beta^2]} \right) - \sqrt{\frac{[\beta^2 - n_2^2 k_0^2]}{[n_1^2 k_0^2 - \beta^2]}} = \alpha \neq 0 \] (1).

Solve by iteration, assuming that \( n_2 < n_{eff} < n_1 \).

When: \( n_{eff} = n_2 \), \( \alpha = 0.682 \)
When: \( n_{eff} = (n_2 + n_1)/2 \), \( \alpha = -0.54 \)
When: \( n_{eff} = n_1 \), \( \alpha = -72 \)

The sign changes between the first two values of \( n_{eff} \Rightarrow \) the solution lies between these points; try half way.

When: \( n_{eff} = (3n_2 + n_1)/4 \), \( \alpha = 4.747 \times 10^{-3} \)
This is very close to zero, so terminate the iteration  
\[ n_{\text{eff}} = \frac{(3n_2 + n_1)}{4} = 1.50125. \]

6.8 Assume that the modes concerned are defined by the transverse fields \( E_a(x) \) and \( E_b(x) \), where:

\[
E_a = E \cos(\kappa_a x) \quad \text{for} \quad 0 < |x| < h/2 \\
E \cos(\kappa_{a,h/2}) \exp\{-\gamma_a [x - h/2]\} \quad \text{for} \quad |x| > h/2
\]

\[
E_b = E \cos(\kappa_b x) \quad \text{for} \quad 0 < |x| < h/2 \\
E \cos(\kappa_{b,h/2}) \exp\{-\gamma_b [x - h/2]\} \quad \text{for} \quad |x| > h/2
\]

where \( \tan(\kappa_{a,h/2}) = \gamma_a / \kappa_a \) and \( \tan(\kappa_{b,h/2}) = \gamma_b / \kappa_b \) (N.B. These expressions are obtained from the standard results, merely by shifting the coordinate origin.)

Now, for orthogonality, we require that \( I = \int I_a I_b dx = 0 \). The integral can be divided into two components, \( I_1 \) and \( I_2 \), such that:

\[ I_1 = 2 \int_0^{h/2} E \cos(\kappa_a x) E \cos(\kappa_{b,h/2}) \exp\{-\gamma_a [x - h/2]\} dx \]

\[ I_2 = 2 \int_{h/2}^{\infty} E \cos(\kappa_{a,h/2}) \exp\{-\gamma_a [x - h/2]\} E \cos(\kappa_{b,h/2}) \exp\{-\gamma_b [x - h/2]\} dx \]

For simplicity, we shall define the following new notation:

\( s_{a,b} = \sin(\kappa_{a,b} h/2), \ c_{a,b} = \cos(\kappa_{a,b} h/2) \) and \( t_{a,b} = \tan(\kappa_{a,b} h/2) \). Hence:

\[ I = 2E^2 \left\{ (s_{a} c_{b} + c_{a} s_{b}) / \{2(\kappa_a + \kappa_b)\} + (s_{b} c_{a} - c_{b} s_{a}) / \{2(\kappa_a - \kappa_b)\} \right\} \]

\[ = (2E^2 / f) \left\{ 2(\kappa_a + \kappa_b)(\gamma_a + \gamma_b)(s_{a} c_{b} + c_{a} s_{b}) + 2(\kappa_a + \kappa_b)(\gamma_a + \gamma_b)(s_{b} c_{a} - c_{b} s_{a}) + 4(\kappa_a^2 - \kappa_b^2)c_{a} c_{b} \right\} \]

where \( f = 4(\kappa_a^2 - \kappa_b^2)(\gamma_a + \gamma_b) \).

Recombining terms in the above, we then get:

\[ I = (8E^2 / f c_{a} c_{b}) \left\{ (\kappa_a \gamma_a + \gamma_b)c_{a} c_{b} - \kappa_b(\gamma_a + \gamma_b)c_{a} s_{b} + (\kappa_a^2 - \kappa_b^2)c_{b} c_{b} \right\} \]

Now, the individual eigenvalue equations require that \( t_{a,b} = \gamma_{a,b} / \kappa_{a,b} \) so that:

\[ I = (8E^2 / f c_{a} c_{b}) \left\{ (\kappa_a \gamma_a + \gamma_b)c_{a} c_{b} - \kappa_b(\gamma_a + \gamma_b)c_{a} s_{b} + (\kappa_a^2 - \kappa_b^2)c_{b} c_{b} \right\} \]

Manipulation of the eigenvalue equation then shows that \( c_{a,b} = \kappa_{a,b} / \sqrt{\kappa_{a,b}^2 + \gamma_{a,b}^2} \), so that:

\[ I = (8E^2 / f c_{a} c_{b}) \left\{ (\kappa_a^2 + \gamma_a^2) - (\kappa_b^2 + \gamma_b^2) \right\} \]

Recall the definition of \( \kappa_{a,b} \) and \( \gamma_{a,b} \):

\[ \kappa_{a,b} = \sqrt{(n_i^2 k_0^2 - \beta_{a,b}^2)}, \gamma_{a,b} = \sqrt{\beta_{a,b}^2 - n_i^2 k_0^2}. \]

Hence, \( \kappa_a^2 + \gamma_a^2 = (\kappa_b^2 + \gamma_b^2) \). \( n_i^2 \) and \( k_0^2 \) and I = 0  

\[ \Rightarrow \text{the two modes are orthogonal. Note that this result is not valid if the two modes are identical, since f = 0 in this case; the integral then gives a finite, non-zero value.} \]

6.9 Define the transverse electric fields in Guides 1 and 2 as \( E_y(x) = E_0 \exp(-\alpha x^2) \) and \( E_{y'}(x) = E_0 \exp(-\alpha(x - \delta)^2) \), respectively. Choose \( E_0 \) so that the fields are normalised, i.e. so that:

\[ \int_{-\infty}^{\infty} E_y^2(x) dx = \int_{-\infty}^{\infty} E_{y'}^2(x) dx = 1. \]
\[ <E_y, E_y> = <E_y', E_y'> = 1. \] This implies that we must set
\[ \int_{-\infty}^{\infty} E_0^2 \exp\{-\alpha (\sqrt{2x})^2\} \, dx = 1. \]

The coupling efficiency is then
\[ \eta = \left| \frac{<E_y, E_y'}{<E_y, E_y> <E_y', E_y'>} \right|^2 \]
\[ = \left| <E_y, E_y'> \right|^2 \]

So that:
\[ \eta = \left( \int_{-\infty}^{\infty} E_0 \exp(-\alpha x^2) \cdot E_0 \exp\{-\alpha (x - \delta)^2\} \, dx \right)^2 \]
\[ = \left( \int_{-\infty}^{\infty} E_0^2 \exp\{-\alpha (2x^2 - 2x\delta + \delta^2)\} \, dx \right)^2 \]
\[ = \left( \exp(-\alpha \delta^2/2) \int_{-\infty}^{\infty} E_0^2 \exp\{-\alpha (\sqrt{2x} - \delta/\sqrt{2})^2\} \, dx \right)^2 \]

Now, the integral above must yield unity - it is merely a shifted version of the normalisation integral defined previously. Hence, we obtain \( \eta = \exp(-\alpha \delta^2) \). Clearly, the coupling efficiency is maximum (and 100%) when there is no lateral offset, and the variation with \( \delta \) is also Gaussian.

6.10 a) (i) If the discontinuity is vanishingly small, only the lowest-order mode of the right-hand guide will be excited. (ii) For larger discontinuities, there will also be conversion to other symmetric modes (e.g. higher-order guided modes and radiation modes). Backward modes will also be excited. However, since the whole geometry is symmetric, there can be no conversion to anti-symmetric modes.

(b) (i) as (ii) above; (ii) as (i) above.
CHAPTER 7

7.1 In layer 1, the solution consists of incident and reflected plane waves. Hence we may write:

\[ E_{y1} = E_i \exp\{-jk_0n_1[z \sin(\theta_1) - x \cos(\theta_1)]\} + E_R \exp\{-jk_0n_1[z \sin(\theta_1) + x \cos(\theta_1)]\} \]

Define \( \beta = k_0n_1 \sin(\theta_1) \) and \( \kappa = \sqrt{n_1^2k_0^2 - \beta^2} \) ⇒ solution can be written as:

\[ E_{y1} = \{E_i \exp(+j\kappa x) + E_R \exp(-j\kappa x)\} \exp(-j\beta z) \]

If incidence is at an angle beyond \( \theta_c \), \[ \left| \frac{E_R}{E_i} \right| = 1 \]

Hence, the solution can be written as: \[ E_{y1} = E \cos(\kappa x - \phi) \exp(-j\beta z) \] (1)

In layer 2, solution consists of growing and decaying evanescent waves ⇒

\[ E_{y2} = E_G \exp\{-jk_0n_2[z \sin(\theta_2) - x \cos(\theta_2)]\} + E_D \exp\{-jk_0n_2[z \sin(\theta_2) + x \cos(\theta_2)]\} \]

where \( \theta_2 \) is found from Snell's law, \( n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \).

Define \( j\gamma = k_0n_2 \cos(\theta_2) = k_0n_2 \sqrt{1 - (n_1/n_2)^2 \sin^2(\theta_1)} = j\sqrt{\beta^2 - n_2^2k_0^2} \)

⇒ \[ E_{y2} = \{E_G \exp(-\gamma x) + E_D \exp(+\gamma x)\} \exp(-j\beta z) \] (2)

In layer 3, solution consists of a single transmitted wave:

\[ E_{y3} = E_T \exp\{-jk_0n_1[z \sin(\theta_1) - x \cos(\theta_1)]\} = E_T \exp(+j\kappa x) \exp(-j\beta z) \] (3).

Boundary conditions:

Match tangential electric field at both interfaces ⇒ match \( E \) at \( x = 0 \) and \( x = h \).

Match tangential magnetic field at both interfaces ⇒ match \( \frac{\partial E}{\partial x} \) at \( x = 0 \) and \( x = h \).

7.2 Write solution in each layer in the form \( E_{yi} = E_i(x) \exp(-j\beta z) \), i.e. as in Equations (1), (2) and (3) in the solution to Question 7.1 above. The boundary conditions then modify to the requirement that \( E_i \) and \( \frac{dE_i}{dx} \) match at both interfaces.

Matching \( E_i \):

\[ E \cos(\kappa h - \phi) = E_G \exp(-\gamma h) + E_D \exp(+\gamma h) \] (1)
\[ E_T = E_G + E_D \] (2)

Matching \( \frac{dE_i}{dx} \):

\[ -\kappa E \sin(kh - \phi) = \gamma \{E_D \exp(+\gamma h) - E_G \exp(-\gamma h)\} \] (3)
\[ j\kappa E_T = \gamma \{E_D - E_G\} \] (4)

Assuming that \( E \) is known, Equations (1) - (4) are four simultaneous equations that must be solved to find the four unknowns \( E_G \), \( E_D \), \( E_T \) and \( \phi \). Combining Equations (2) and (4) gives:

\[ E_G = E_T \gamma(\gamma - j\kappa)/2\gamma \] and \[ E_D = E_T \gamma(\gamma + j\kappa)/2\gamma \].

Substitution into Equations (1) and (3) then yields:

\[ E \cos(\kappa h - \phi) = E_T \{(\gamma - j\kappa)/2\gamma \} \exp(-\gamma h) + (\gamma + j\kappa)/2\gamma \exp(+\gamma h) \] (5)
\[ -\kappa E \sin(kh - \phi) = \gamma E_T \{(\gamma + j\kappa)/2\gamma \} \exp(+\gamma h) - (\gamma - j\kappa)/2\gamma \exp(-\gamma h) \] (6)

Finally, dividing Equations (5) and (6) gives:

\[ -\kappa \tan(kh - \phi) = \gamma(\{(\gamma + j\kappa)/2\gamma \} \exp(+\gamma h) - (\gamma - j\kappa)/2\gamma \exp(-\gamma h)) / (\{(\gamma - j\kappa)/2\gamma \} \exp(-\gamma h) + (\gamma + j\kappa)/2\gamma \exp(+\gamma h)) \]
In principle, this equation could now be solved analytically for $\phi$. However, it is simpler to make some approximations. If $\gamma h$ is reasonably large, $\exp(-\gamma h) \approx 0$ in the above, so that $\tan(kh - \phi) = -\gamma/\kappa$. In this case, $\phi$ is a constant, independent of $h$.

Now, from (5) above,
$$E_T = E \cos(kh - \phi) / \left[ \left( \frac{\gamma - j\kappa}{2\gamma} \right) \exp(-\gamma h) + \left( \frac{\gamma + j\kappa}{2\gamma} \right) \exp(+\gamma h) \right]$$
Using the same approximation, $E_T \approx E \cos(kh - \phi) (2\gamma/(\gamma + j\kappa)) \exp(-\gamma h)$.
Consequently, $E_T$ falls off exponentially as the thickness $h$ increases.

7.3 Phase matching occurs when $k \sin(\theta_1) = \beta$, or when $k_0 n_p \sin(\theta_1) = k_0 n_{\text{eff}}$
$\Rightarrow \theta_1 = \sin^{-1}(n_{\text{eff}}/n_p)$.

There will be three phase-matching angles, at
$\theta_A = \sin^{-1}(1.515/1.7) = 63.021^\circ$,
$\theta_B = 64.158^\circ$ and
$\theta_C = 65.343^\circ$.

This will give rise to three dips in the angular variation in reflectivity. Corresponding propagation constants are:
$\beta_A = 2\pi n_{\text{eff}A}/\lambda_0 = 6.346 \times 10^6 /\text{m},$
$\beta_B = 6.408 \times 10^6 /\text{m}$ and
$\beta_C = 6.471 \times 10^6 /\text{m}$.

7.4 Phase-matching diagrams:

a) Reflection and refraction take place. Transmitted beam will be the strongest; both reflected beams will be fairly weak, so little mode conversion occurs.

b) $\theta_1 > \sin^{-1}(1.5025/1.504)$, so total internal reflection occurs. There is then no transmitted beam, and strong mode conversion probably occurs.

7.5. The lens maker's formula implies that $1/f = \{(n_{\text{eff}L} - n_{\text{eff}M}) / n_{\text{eff}M} \} \{1/r_1 + 1/r_2 \}$. $f$ is then smallest when $r_1$ and $r_2$ are equal, and as small as possible, so that $2r = \text{beam}$
width in each case. Hence, the most compact arrangement uses circular overlay lenses in the geometry shown below:

For the input lens, \( f_a = \frac{n_{\text{effL}}}{(n_{\text{effL}} - n_{\text{effM}})} \cdot \frac{r}{2} = (1.6/0.02) \times (0.5/4) = 10 \text{ mm}. \) For x5 beam expansion, the focal length of the output lens is \( f_b = 50 \text{ mm}. \) Hence, the lenses are 60 mm apart.

7.6. \( 2 \Lambda \sin(\theta_0) = \frac{\lambda_0}{n} \Rightarrow \theta_0 = \sin^{-1}\left\{ \frac{1.5}{(2 \times 1 \times 1.6)} \right\} = 27.953^\circ. \)
\[
\Delta \lambda = \frac{\lambda_0}{d} \cdot \cot(\theta_0) = (1.5 \times 10^{-6} \times 1/50) \times \cot(27.953^\circ)
\]
\[
= 5.653 \times 10^{-6} \text{ m} \text{ or } 56.533 \text{ nm}.
\]

For an input angle of \( \theta, y = d \tan(\theta) = 50 \tan(27.953^\circ) \mu\text{m} = 26.533 \mu\text{m}. \) Since \( \Lambda = 1 \mu\text{m}, \) approximately 27 fringe planes are crossed.

7.7 Start with Bragg’s law, \( 2 \Lambda \sin(\theta_0) = \frac{\lambda_0}{n}. \)
Differentiating both sides gives \( 2 \Lambda \cos(\theta_0) \Delta \theta_0 = \Delta \lambda_0 / n. \) Now, the wavelength bandwidth is defined by \( \Delta \lambda_0 = (\Lambda \lambda_0 / d) \cot(\theta_0) \) so substituting into the above gives \( \Delta \theta_0 = \frac{\Lambda}{2n \sin(\theta_0)} \) \( \Delta \lambda_0 / \Lambda = \frac{\Lambda}{d} \) for the angular bandwidth.
\[
\Lambda = \frac{\lambda_0}{[2n \sin(\theta_0)]} = 0.633/[2 \times 1.6 \times \sin(30^\circ)] \mu\text{m} = 0.3956 \mu\text{m}
\]
Hence, \( \Delta \theta_0 = 0.3956/20 \text{ rad} = 0.01978 \text{ rad} \text{ or } 1.133^\circ. \)

7.8 Bragg’s law is \( 2 \Lambda \sin(\theta_0 - \phi) = \frac{\lambda_0}{n}. \)
Hence, \( 2(2\pi n/\lambda_0) \sin(\theta_0 - \phi) = 2 \pi / \Lambda \Rightarrow 2 \sin(\theta_0 - \phi) = \frac{\lambda_0}{2 \Lambda}. \)

This relationship is a reflection of the geometrical construction shown in the k-vector diagram below, where \( k \sin(\theta_0 - \phi) = \frac{|K|}{2}. \)
7.9  

a) Transmission grating: the period is $\Lambda = 5 \, \mu\text{m}$. Hence, the grating vector is $|\mathbf{K}| = \frac{2\pi}{\Lambda} = \frac{2\pi}{5} \times 10^{-6} = 1.256 \times 10^6 \, \text{m}^{-1}$

$k = \frac{2\pi n}{\lambda} = \frac{2\pi}{1.6/0.633} \times 10^{-6} = 15.88 \times 10^6 \, \text{m}^{-1}$; $|\mathbf{K}| \ll k$, and $d \approx \Lambda$.

Hence, diffraction will take place in the Raman-Nath regime, with many diffraction orders being produced. The deflection angle of the first diffraction order is therefore approximately $\theta = \tan^{-1}(1.256/15.88) = 4.522^\circ$.

b) Reflection grating: the period is $\Lambda = 0.5 \, \mu\text{m}$

$|\mathbf{K}| = \frac{2\pi}{\Lambda} = \frac{2\pi}{0.5} \times 10^{-7} = 1.256 \times 10^7 \, \text{m}^{-1}$.

Hence, the deflection angle is given by $(\theta_0 - \phi) = \sin^{-1}\left(\frac{|\mathbf{K}|}{2k}\right) = \sin^{-1}\left\{\frac{12.566}{2 \times 15.88}\right\} = 23.3^\circ$. 

R.R.A. Sym and J.R. Cozens  Optical Guided Waves and Devices  22
CHAPTER 8

8.1 If attenuation is 4 dB/km, loss in 10 km of fibre is 4 x 10 = 40 dB
⇒ 40 = -10 \log_{10}(P_{out}/P_{in}) ⇒ P_{out}/P_{in} = 10^{-4} or 0.01%.

If the loss is caused by Rayleigh scattering, it falls off as \(1/\lambda^{4}\).

Hence, at \(\lambda_0 = 1.3\ \mu m\), the loss will have reduced by \((0.65/1.3)^4 = 1/16\) so the attenuation will be 0.25 dB/km. The fraction of power transmitted is now \(P_{out}/P_{in} = 10^{-0.25} = 0.562\) or 56.2%.

8.2 The definition of numerical aperture is \(NA = \sqrt{(n_1^2 - n_2^2)}\). Similarly, the definition of V-value is \(V = k_0 a \sqrt{(n_1^2 - n_2^2)}\). Hence, \(a = V/(k_0 NA)\).

Now, \(N \approx V^2/2\), so that \(V \approx \sqrt{2N}\). Hence, \(d = 2a = 2\sqrt{(2N)/(k_0 NA)} = 2\sqrt{(2000) \times 0.85 \times 10^{-6}/(2\pi \times 0.2)} = 60.5 \times 10^{-6}\ m\).

8.3 i) Consider the propagation of a group of rays emanating from a point \((x, 0)\) at the left-hand end of the lens. Now, the general solution for ray paths is

\[x = x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2} \sin(z/x_0 + \Psi)\]

Hence, at \(z = 0\), we have \(x = x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2} \sin(\Psi)\)
⇒ \(\Psi = \sin^{-1}[x/x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2}]\).

After propagation through half a pitch length, \(z/x_0 = \pi\). Hence, at this plane, the transverse position of a general ray is

\[x' = x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2} \sin(\pi + \Psi) = -x_0 \sqrt{1 - \beta^2/k_0^2 n_0^2} \sin(\Psi) = -x\].

All rays emanating from \((x, 0)\) therefore intersect at a single point, so a real point image is formed. Since \(x' = -x\), the image is inverted.

ii) Consider a general forward-travelling field \(E_i(x)\), specified at the input plane \(z = 0\). This may be expanded in terms of guided modes as \(E_i(x) = \sum a_v E_v(x)\), where \(a_v\) is the amplitude of the \(v^{th}\) mode, and \(E_v(x)\) is its transverse field. After propagating a distance \(z\), this field will be modified to \(E_{out}(x) = \sum a_v E_v(x) \exp(-j\beta_v z)\), where:

\[\beta_v = k_n n_0 - (2v + 1)/2n_0\].

Inserting these values, we get

\[E_{out}(x) = \exp[-j(k_n n_0 - 1/2n_0)z] \sum a_v E_v(x) \exp(jvz/n_0)\]

If \(z = \pi x_0\) (i.e., if \(z = \lambda/2\)), this reduces to \(E_{out}(x) = \exp(-j\Psi) \sum a_v E_v(x) \exp(jv\pi)\), where the phase term \(\Psi = (k_n n_0 - 1/2n_0)\lambda/2\). At this point, \(E_{out}\) is very similar to \(E_i\), multiplied by an unimportant exponential phase factor. However, even-numbered modes have a phase-shift of \(\pi\) with respect to odd modes. This corresponds to a sign change in all antisymmetric components in the image, so the image must be inverted.

8.4 The TE reflection coefficient for a dielectric interface is:

\[\Gamma_E = \{n_1 \cos(\theta_1) - n_2 \cos(\theta_2)\} / \{n_1 \cos(\theta_1) + n_2 \cos(\theta_2)\}\]
This may be written as:
\[ \Gamma_E = \frac{1 - \alpha}{1 + \alpha} \]

where:
\[ \alpha = \frac{n_2 \cos(\theta_2)}{(n_1 \cos(\theta_1))}, \]

or
\[ \alpha = \sqrt{\frac{(n_2^2 - n_1^2 \sin^2(\theta_1))}{(n_1^2 - n_2^2 \sin^2(\theta_1))}} \]

This function is discontinuous, tending to \( \alpha = 1 \) and \( \alpha = -j \) as \( \theta_1 \to 90^\circ \) and \( n_1 \to n_2 \), for \( n_1 < n_2 \) and \( n_1 > n_2 \) respectively. In the latter situation (which corresponds to that of a turning point), \( \Gamma_E = \frac{1 + j}{1 - j} = \exp(j\pi/2) \).

8.5 \[ a = \sqrt{\frac{2r_0/n_0 k_0}{n_0 k_0}} = \sqrt{\frac{2 x 25 x 10^{-6} x 0.85 x 10^{-6}}{(1.5 x 2\pi x 2.123)}} = 2.123 x 10^{-6} \text{ m.} \]

\[ \theta = \frac{2}{k_0 a} = \frac{2 x 0.85}{(2\pi x 2.123)} = 0.1274 \text{ rad or 7.3}^\circ. \]

8.6 Differentiating the assumed solution for the transverse field, we have:
\[ \frac{\partial E_{\mu,\nu}}{\partial x} = \left\{ \frac{\sqrt{2/a} H'_\mu(\sqrt{2x/a})}{H_\mu(\sqrt{2x/a})} \right\} H_\nu(\sqrt{2y/a}) \exp\{-(x^2 + y^2)/a^2\} \]

\[ \frac{\partial^2 E_{\mu,\nu}}{\partial x^2} = \left\{ \frac{2/a^2 H''_\mu(\sqrt{2x/a})}{H_\mu(\sqrt{2x/a})} - \frac{4\sqrt{2x/a^3} H'_\mu(\sqrt{2x/a})}{H_\mu(\sqrt{2x/a})} \right\} H_\nu(\sqrt{2y/a}) \exp\{-(x^2 + y^2)/a^2\} \]

\[ \frac{\partial E_{\mu,\nu}}{\partial y} \text{ and } \frac{\partial^2 E_{\mu,\nu}}{\partial y^2} \] may be found in a similar way. Substituting into the waveguide equation, eliminating all common terms, and defining \( \xi = \sqrt{2x/a} \) and \( \eta = \sqrt{2y/a} \), we then get:

\[ \left\{ \frac{H''_\mu(\xi)}{H_\mu(\xi)} - 2\xi \frac{H'_\mu(\xi)}{H_\mu(\xi)} \right\} - 1 \] + \left\{ \frac{H''_\nu(\eta)}{H_\nu(\eta)} - 2\eta \frac{H'_\nu(\eta)}{H_\nu(\eta)} \right\} - 1 \] + (a^2/2) {n_0^2 k_0^2 - \beta_{\mu,\nu}^2} = 0

Now, the contents of the first curly bracket in the above are a function of \( \xi \) only. Those of the second are a function only of \( \eta \). The third is a constant. For the equation to be satisfied for all \( \xi \) and \( \eta \), we must therefore have:

\[ \left\{ \frac{H''_\mu(\xi)}{H_\mu(\xi)} - 2\xi \frac{H'_\mu(\xi)}{H_\mu(\xi)} \right\} - 1 \] = constant = A, say, so that:

\[ H''_\mu(\xi) - 2\xi \frac{H'_\mu(\xi)}{H_\mu(\xi)} = 0 \]

Similarly, we may write:

\[ H''_\nu(\eta) - 2\eta \frac{H'_\nu(\eta)}{H_\nu(\eta)} = 0. \]

The propagation constant is then defined by:

\[ (a^2/2) \left\{ n_0^2 k_0^2 - \beta_{\mu,\nu}^2 \right\} + A + B = 0 \]

or \( \beta_{\mu,\nu}^2 = n_0^2 k_0^2 + (2/a^2) (A + B) \).

The two differential equations above are each of the standard form satisfied by Hermite polynomials, provided \( A = -(2\mu + 1) \) and \( B = -(2\nu + 1) \). This leads to the required definition of the propagation constant, namely:

\[ \beta_{\mu,\nu} = n_0 k_0 \sqrt{1 - 2(\mu + \nu + 1) / (n_0 k_0 r_0)}. \]

8.9 Define the new parameter \( A = w_0/a \). The end-fire coupling efficiency is then:

\[ \eta(A) = 4A^2 / \left\{ A^2 + 1 \right\}^2 = 4(1/A^2) / \left\{ 1 + (1/A^2) \right\}^2. \]

Hence, \( \eta(A) = \eta(1/A) \), so the variation will be symmetric in
\[ \alpha = \log(A) \ comma \ \frac{d\eta}{dA} = \frac{8A}{(A^2 + 1)^2} - \frac{16A^3}{(A^2 + 1)^3} \]

\[ \Rightarrow \frac{d\eta}{dA} = 0 \] when \( 8A (A^2 + 1) - 16A^3 = 0 \)

\[ \Rightarrow \eta \text{ is maximum when } A = 1, \text{ or when } a = w_0 \]

The variation of \( \eta \) with \( \alpha \) is then as follows:

8.10 Use two 8 x 8 stars and eight extra fused tapered couplers, as shown below.
CHAPTER 9

9.1 For Ga$_{1-x}$Al$_x$As, \( n \approx 3.57 \) when \( x = 0 \), and \( n \approx 3.35 \) when \( x = 0.35 \).
Hence, the refractive index varies as \( n \approx 3.57 - 0.6286x \).
Hence, for \( x = 0.15 \), \( n = 3.57 - 0.6286 \times 0.15 = 3.476 \).

From Chapter 6, a symmetric slab guide will be single-moded provided
\[
h \leq \frac{\lambda_0}{2\sqrt{(n_1^2 - n_2^2)}}.
\]
Assuming that \( n_1 = 3.57 \), \( n_2 = 3.476 \) we get \( h \leq 0.9 / 2\sqrt{(3.57^2 - 3.476^2)} = 0.553 \mu m \).

9.2 For bends of radius \( r \) formed in a given channel guide system, the optical power \( P_z \) at a distance \( z \) round the bend is related to the input power \( P_0 \) by \( P_z = P_0 \exp(-\alpha z) \),
where \( \alpha = c_1 \exp(-c_2 r) \) and \( c_1 \) and \( c_2 \) are constants independent of \( r \).

For 90° bends made using bend radii of \( r_1 \) and \( r_2 \), define the distance travelled as \( z_1 = \pi r_1 / 2 \) and \( z_2 = \pi r_2 / 2 \) and the transmission as \( P_{z1}/P_0 = T_1 \) and \( P_{z2}/P_0 = T_2 \).

Using the formula above,
\[
T_1 = \exp\{-c_1 \pi r_1 / 2 \exp(-c_2 r_1)\} \quad \text{and} \quad T_2 = \exp\{-c_1 \pi r_2 / 2 \exp(-c_2 r_2)\}
\]

Hence:
\[
\log_e(T_1) / \log_e(T_2) = (r_1/r_2) \exp\{-c_2 (r_1 - r_2)\}
\]
\[
\Rightarrow c_2 = \log_e\{r_2 \log_e(T_1) / r_1 \log_e(T_2)\} / \{r_2 - r_1\}.
\]

For \( r_1 = 1 \) mm, the radiation loss is 10 dB \( \Rightarrow T_1 = 0.1 \)
For \( r_2 = 10 \) mm, the radiation loss is 3 dB \( \Rightarrow T_2 = 0.5 \). Hence,
\[
c_2 = \log_e\{10 \log_e(0.1) / \log_e(0.5)\} / \{10 - 1\} /\text{mm} = 0.389 /\text{mm}.
\]
Similarly,
\[
c_1 = -2 \log_e(T_1) / \{\pi r_1 \exp(-c_2 r_1)\}
\]
\[
= -2 \log_e(0.1) / \{\pi \times \exp(-0.389)\} /\text{mm} = 2.163 /\text{mm}.
\]

Hence, the transmission through a 45° bend of radius \( r_3 = 5 \) mm is:
\[
T_3 = \exp\{-c_1 \pi r_3 / 4 \exp(-c_2 r_3)\} = \exp\{-2.163 \pi x 5 / 4 \exp(-0.389 x 5)\} = 0.2968,
\]
corresponding to a loss of 5.27 dB.

9.3 Transmission through the reflective bend may be analysed by considering propagation through an equivalent image guide, the reflection of the actual output guide in the line M - M' as shown below. Ignoring the diffraction that must occur in the short distances when the light is not fully confined, we may then consider the problem as one of butt coupling between the two sections of guide shown. If the normalised transverse field in guide 1 is \( E \), and that in guide 2 is \( E' \), the transmission efficiency is then \( \eta_r = |<E', E>|^2 \) as usual.

Even if \( \delta = 0 \), the transmission will not reach exactly 100% - since the mirror does not extend over all space, it cannot reflect the entire transverse field. The evanescent tails extending outside the guide will not be reflected, and consequently some power must be lost to radiation.
9.4 The input power is proportional to
\[ P_{\text{in}} = \sum |A_i|^2 = 1 + 4 + 9 + 0 + 1 + 4 + 16 + 1 = 36. \]

The Y-junction tree is an amplitude-summing device, giving the output
\[ A_{\text{out}} = (1/\sqrt{2})^3 \sum A_i = (1/\sqrt{2})^3 \{1 - 2 + 3 - 0 + 1 + 2 + 4 - 1\} = 3/\sqrt{2}. \]

The output power is proportional to \[ P_{\text{out}} = |A_{\text{out}}|^2 = 4.5. \]
The efficiency of the summing process is then \[ \eta = P_{\text{out}}/P_{\text{in}} = 4.5/36 = 0.125, \text{ corresponding to } 12.5\%. \]

If the inputs are all arranged to be in-phase, the output amplitude rises to:
\[ A_{\text{out}} = (1/\sqrt{2})^3 \{1 + 2 + 3 + 0 + 1 + 2 + 4 + 1\} = 7/\sqrt{2}. \]

Hence, \[ |A_{\text{out}}|^2 = 24.5 \text{ and } \eta = P_{\text{out}}/P_{\text{in}} = 24.5 / 36 = 0.681, \text{ or } 68.1\%. \]

9.5 i) Uses the horizontal component of the electric field.
ii) Uses the vertical component of the electric field.
Field lines and waveguide positions are as shown below.

9.6 Cutoff will occur when the electrically-induced index change \( \Delta n_{\text{eff}} \) reduces the effective index of the guide to a value equal to the index of the substrate. Hence, we require:
\[ |\Delta n_{\text{eff}}| = n_e^3 r_{33} |V| \Gamma/2g = n_{\text{eff}} - n_e, \text{ or:} \]
\[ |V| = 2g(n_{\text{eff}} - n_e)/n_e^3 r_{33} \Gamma \]
\[ = 2 \times 7 \times 10^{-6} \times 0.001 / \{2.23 \times 30.8 \times 10^{-12} \times 0.6\} V = 71.1 \text{ V}. \]

As cutoff is approached, the confinement of the optical field will be greatly reduced. Hence, the overlap factor \( \Gamma \) will also reduce, and a larger voltage will actually be required to obtain reasonable extinction of the guided beam.
9.7 For a sinusoidal time-varying phase shift $\Phi(t) = \Phi_{\text{max}} \sin(\omega_m t)$, the output electric field from the modulator is $E_{\text{out}}(x, y, t) = E_{\text{out}}(x, y) \sum_n (-1)^n J_n(\Phi_{\text{max}}) \exp(jn\omega_m t)$, leading to an infinite number of sidebands at angular frequencies $\omega + n\omega_m$, where $\omega$ is the carrier frequency.

For a 10 MHz modulation frequency, the sidebands in the range 20 MHz below the carrier frequency to 20 MHz above it are those with $n = -2, -1, 0, +1$ and $+2$. Now, $(V_L)_o$ for the modulator is 50 V mm; given that $L = 10$ mm and $V_{\text{p-p}} = 10$ V (so that $V_{\text{max}} = 5$ V), it follows that $\Phi_{\text{max}} = \pi$. The fractional power in the range is then $P = J_0^2(\pi) + 2J_1^2(\pi) + 2J_2^2(\pi)$. From standard tables:

$J_0(3.1) = -0.29206$  \quad $\Rightarrow J_0(\pi) \approx -0.30376$

$J_1(3.1) = 0.30092$  \quad $\Rightarrow J_1(\pi) \approx 0.28446$

$J_2(3.1) = 0.48620$  \quad $\Rightarrow J_2(\pi) = 0.48508$

$\Rightarrow P = (-0.30376)^2 + 2 \times (0.28446)^2 + 2 \times (0.48508)^2 = 0.7247$, or 72.5%.

9.8 The modulating signal is a travelling wave, of the form $V(z, t) = V_0 \exp[j(\omega_m t - \beta_m z)]$  

where $V_0$ is the peak voltage across the electrodes and $\beta_m$ is the propagation constant of the stripline. This wave travels at a phase velocity $v_{\text{pm}} = \omega_m / \beta_m$ and gives rise to a moving variation in effective index in the optical waveguide. From our previous analysis of DC phase modulators, we may write the travelling variation in effective index as:

$\Delta n_{\text{eff}}(z, t) = -(n_0 \sqrt{r} \Gamma / 2g) V(z, t) = -\{n_0 r V_0 \sqrt{\Gamma / 2g}\} \exp[j(\omega_m t - \beta_m z)]$

where $r$, $\Gamma$ and $g$ are the electro-optic coefficient, overlap factor and electrode gap, respectively.

Assume that the optical wave is at $z = 0$ at $t = t_0$. Since it travels at a phase velocity $v_{\text{po}}$, it will reach a point $z = v_{\text{po}}(t - t_0)$ at time $t$. In this case, the effective index change it encounters at any point along the modulator may be written as:

$\Delta n_{\text{eff}}(z, \tau) = \Delta n \exp[j\{\omega_m (z/v_{\text{po}} + t_0 - \beta_m z)\}]$, where $\Delta n = -(n_0 r V_0 \sqrt{\Gamma / 2g})$

$= \Delta n \exp[j\{\omega_m t_0 + (v_{\text{po}}/V_0 - 1) \beta_m z]\}$

$= \Delta n \exp(j\omega_m t_0) \exp(j\alpha z)$ where $\alpha = (\omega_m n_m / c) (n_0 / n_m - 1)$

Now, the phase change $\Phi(t_0)$ accumulated in travelling the length $L$ of the modulator may be found by integration, as:

$\Phi(t_0) = \int_{t_0}^{t} 2\pi \Delta n_{\text{eff}}(z) / \lambda_{oz} \ dz = \Delta \beta \exp(j\omega_m t_0) \int_{t_0}^{t} \exp[j\alpha z] \ dz$

where $\Delta \beta = 2\pi \Delta n / \lambda_{oz} = \Delta \beta L \exp(j\alpha L / 2) \exp(j\omega_m t_0)$

This can be written as $\Delta \beta = \Delta \beta L \phi(\alpha) \exp(j\omega_m t_0)$

where $\phi(\alpha) = \exp(j\alpha L / 2) \exp(j\omega_m t_0)$
\( \Phi(t_0) \) therefore varies sinusoidally with \( t_0 \), as might be expected. The term \( \phi(\alpha) \) accounts for the frequency-dependence of the induced phase shift, and falls off as the modulation frequency is increased. The effectiveness of the modulator is determined by \( |\phi| \). In terms of angular frequency, this can be written

\[
|\phi(\omega_m)| = \left| \text{sinc}\left( \frac{\omega_m n_m L}{2c} \left( \frac{n_0}{n_m} - 1 \right) \right) \right|.
\]

\( |\phi(\omega_m)| \) falls to zero for the first time when \( \left( \frac{\omega_m n_m L}{2c} \left( \frac{n_0}{n_m} - 1 \right) \right) = \pi \), or at a frequency \( f_m = \frac{c}{n_m L} / \left( \frac{n_0}{n_m} - 1 \right) \).

Consequently, the effectiveness of a travelling-wave modulator is reduced at high modulation frequencies, if there is any difference between \( n_m \) and \( n_0 \) (and hence if there is any difference between the velocities of the electrical and optical waves). The larger the velocity difference, the lower the frequency at which the induced phase modulation is reduced significantly.

9.9 The normalised output from a dual-modulator Mach-Zehnder interferometer is \( P = \cos^2(\Phi) \), where \( \Phi \) is the phase shift applied by one of the modulators. If \((V L)_{\pi} = 50\ \text{V mm}\), and \( L = 10\ \text{mm} \), then \( V_{\pi} = 5\ \text{V} \). Consequently, for \( V = 1.25\ \text{V} \), \( \Phi = \pi/4 \), and for \( V = 6.25\ \text{V} \), \( \Phi = 5\pi/4 \). The interferometer therefore has a static phase bias of \( \pi/4 \) radians, with a superimposed dynamic variation over a further range of \( \pi \) radians. The output is therefore as shown below.

9.10 The device is a reflective variant of the Mach-Zehnder interferometer known as a Michelson interferometer. The normalised reflected output is \( P = \cos^2(\Phi) \) where \( \Phi \) (the phase imbalance) is given by \( \Phi = \beta L \). Given that both \( L \) and the effective index are likely to vary with temperature, the device could be used as an integrated optic temperature sensor.
CHAPTER 10

10.1 The main questionable approximation is that the overlap term \( < E_2, E_1 > \) is negligible, while assuming that the coupling term \( k_0^2 (n_T^2 - n_2^2) E_2, E_1 > \) is not. This is inconsistent - if the coupling term is to have a finite value, the two modes must overlap to a reasonable extent, so that \( < E_2, E_1 > \neq 0 \). In fact, numerical calculations show typically that \( < E_2, E_1 > \) lies in the range 0.1 - 0.2 for practical devices, compared with unity for \( < E_1, E_1 > \) and \( < E_2, E_2 > \). As a result, it cannot logically be possible to obtain exactly 100% efficiency from a synchronous coupler, or zero crosstalk in a switch - there must always be some light in the 'wrong' guide.

10.2 a) The contribution from all paths involving three scatterings is:
\[
A_{2L}^3 = (-j\kappa)^3 A_{10} \int_{z_1}^{L} \int_{z_2}^{L} \int_{z_3}^{L} dz_3 \, dz_2 \, dz_1 = (-j\kappa L)^3 / 3! \, A_{10}
\]
b) Paths contributing to \( A_{2L}^5 \):
\[
A_{2L}^5 = (-j\kappa)^5 A_{10} \int_{z_1}^{L} \int_{z_2}^{L} \int_{z_3}^{L} \int_{z_4}^{L} \int_{z_5}^{L} dz_5 \, dz_4 \, dz_3 \, dz_2 \, dz_1 = (-j\kappa L)^5 / 5! \, A_{10}
\]
Hence \( A_{2L} = -j A_{10} \{ \kappa L - (\kappa L)^3 / 3! + (\kappa L)^5 / 5! \, \ldots \} = -j A_{10} \sin(\kappa L) \)

10.4 Start with: \( dA_1/dz + j\kappa A_2 \exp(-j\Delta \beta z) = 0 \)  
Multiply Eqn (1) by \( A_1^* \Rightarrow A_1^* dA_1/dz + j\kappa A_1 A_2^* \exp(-j\Delta \beta z) = 0 \)  
Take complex conjugate of (2) \( \Rightarrow A_1 dA_1^*/dz - j\kappa A_1 A_2^* \exp(+j\Delta \beta z) = 0 \)  
Perform similar process with: \( dA_2/dz + j\kappa A_1 \exp(+j\Delta \beta z) = 0 \)  
Multiply Eqn (4) by \( A_2^* \Rightarrow A_2^* dA_2/dz + j\kappa A_2 A_1^* \exp(+j\Delta \beta z) = 0 \)  
Take complex conjugate of (5) \( \Rightarrow A_2 dA_2^*/dz - j\kappa A_2 A_1^* \exp(-j\Delta \beta z) = 0 \)  
Now sum Eqns (2), (3), (5), (6) \( \Rightarrow A_1^* dA_1/dz + A_1 dA_1^*/dz + A_2^* dA_2/dz + A_2 dA_2^*/dz = 0 \)  
Or: \( d(A_1 A_1^*)/dz + d(A_2 A_2^*)/dz = 0 \)

10.5 Assume that the coupling coefficient depends roughly exponentially on the interwaveguide gap, following e.g. \( \kappa = \kappa_0 \exp(-\alpha g) \). Now, \( \kappa = \pi/(2L_{ng2}) \). From the data given, we obtain:
\[
g = 4 \mu m, \ \kappa = 1.5708/5 \times 10^{-3} = 314.16 / m = \kappa_0 \exp(-\alpha x 4 \times 10^{-6})
\]
\[
g = 6 \mu m, \ \kappa = 1.5708/7 \times 10^{-3} = 224.39 / m = \kappa_0 \exp(-\alpha x 6 \times 10^{-6})
\]
\[
g = 7.5 \mu m, \ \kappa = 1.5708/9 \times 10^{-3} = 174.53 / m = \kappa_0 \exp(-\alpha x 7.5 \times 10^{-6})
\]
From results (1) and (2), we get \( \exp(-\alpha x 2 \times 10^{-6}) = 224.39/314.16 = 0.71425 \).
\( \Rightarrow \alpha = -(1/2 \times 10^{-6}) \log_2(0.71425) = 1.6826 \times 10^6 / m \)
Hence, \( \kappa_0 = 314.16 / \exp(-1.6826 \times 10^6 x 4 \times 10^{-6}) = 615.81 / m \).
We can check that this is consistent, by predicting the coupling coefficient when $g = 7.5 \:\mu m$. In this case, we get $\kappa = 615.81 \exp(-1.6826 \times 10^5 \times 7.5 \times 10^{-6}) = 174.4 /m$, close to 174.53.

When $g = 7 \:\mu m$, $\kappa = 615.81 \times \exp(-1.6826 \times 10^5 \times 7 \times 10^{-6}) = 189.64 /m$. Hence, the length required for 50% coupling efficiency is $L_{\pi/4} = \pi/(4 \times 189.64) \: m = 4.14 \: mm$.

10.6 a) The fabrication process may be controlled by monitoring the optical transmission through the fibre. Typically, the grinding and polishing will take place in liquid slurry. When sufficient material has been removed, the evanescent field of the fibre mode will extend into this slurry, and there will be considerable attenuation.

b) The theoretical cross-coupled power is $P_2 = \sin^2(\kappa L)$. This will be very small if $\kappa L \approx \nu \pi$. If the coupler is assembled, and the two blocks are translated across one another, $\kappa$ will vary, reaching a maximum when distance between the fibres is minimum. Consequently, if $\kappa L = 0$, there will be only a single maximum in $P_2$ at this point as the blocks are shifted, and this will be very small. The remedy is to remove more material from the fibre by further polishing. If $\kappa L = \pi$, on the other hand, there will be two maxima, displaced on either side of the previous point. In each case, the cross-coupling efficiency will be $\approx 100\%$. The solution in this case is to operate the coupler at one of these points. Higher values of $\kappa L$ are similar.

10.7 a) The contribution from all paths involving three scatterings is:

$$A_{B_0^3} = (-j\kappa)^3 A_{F_0} \left[ \int_0^L \int_0^{z_1} \int_0^{z_2} \int_0^L \: dz_3 \: dz_2 \: dz_1 \right] = (-j\kappa L)^3/3 A_{10}$$

b) Paths contributing to $A_{B_0^5}$:

$$A_{B_0^5} = (-j\kappa)^5 A_{10} \left[ \int_0^L \int_0^{z_1} \int_0^{z_2} \int_0^{z_3} \int_0^L \: dz_4 \: dz_3 \: dz_2 \: dz_1 \right] = (-j\kappa L)^5 2/15 A_{10}$$

Hence $A_{B_0} = -j A_{10} \{ \kappa L - (\kappa L)^3/3 + (\kappa L)^5 2/15 \ldots \} = -j A_{10} \tanh(\kappa L)$

10.8 Start with $dA_p/dz + j \kappa A_p = 0 \Rightarrow d^2A_p/dz^2 + j \kappa \: dA_p/dz = 0 \quad (1)$

Now, $dA_p/dz - j \kappa A_p = 0 \Rightarrow dA_p/dz = j \kappa A_p \quad (2)$

Substitute in Eqn. (1) $\Rightarrow d^2A_p/dz^2 - \kappa^2 A_p = 0 \quad (3)$

Assume the general solution:

$$A_p = A \cosh(\kappa z) + B \sinh(\kappa z) \quad (4)$$

The boundary conditions are: $A_p = 1$ at $z = 0 \Rightarrow A = 1 \quad (5)$

And: $A_p = 0$ at $z = L \Rightarrow dA_p/dz = 0$ at $z = L$, from Eqn. (2)

Hence: $\kappa \sinh(\kappa L) + \kappa B \cosh(\kappa L) = 0 \Rightarrow B = -\tanh(\kappa L) \quad (6)$

So that:

$$A_p = \cosh(\kappa z) - \tanh(\kappa L) \sinh(\kappa z)$$

$$= \{ \cosh(\kappa L) \cosh(\kappa z) - \sinh(\kappa L) \sinh(\kappa z) \} / \cosh(\kappa L)$$

$$= \cosh\{ \kappa(L - z) \} / \cosh(\kappa L) \quad (7)$$
Now, from Eqn. (1):
\[ A_B = \frac{j\kappa}{dA_p/dz} = -j \frac{\sinh(\kappa(L - z))}{\cosh(\kappa L)} \] (8)
\[ = -j \frac{\sinh(\kappa(L - z))}{\cosh(\kappa L)} \] (9)
Equations (7) and (9) then represent the desired solutions.

10.9 Reflection efficiency is \( R = \tanh^2(\kappa L) \)
\[ \Rightarrow L = \frac{1}{\kappa \tanh^{-1}(\sqrt{R})} = \frac{1}{200} \tanh^{-1}(\sqrt{0.99}) \text{ m} = 15 \text{ mm} \]
11.1 For maximum in the function \( \rho(\lambda_0) = (8\pi\hbar c/\lambda_0^5)/\{\exp(\hbar c/\lambda_0 kT) - 1\} \), we require \( d\rho/d\lambda_0 = 0 \). Hence:

\[
-(5/\lambda_0^6)/\{\exp(\hbar c/\lambda_0 kT) - 1\} + (1/\lambda_0^5)(\hbar c/\lambda_0^2 kT)\exp(\hbar c/\lambda_0 kT)/\{\exp(\hbar c/\lambda_0 kT) - 1\}^2 = 0
\]

Or:

\[
-5 \{\exp(\hbar c/\lambda_0 kT) - 1\} + (\hbar c/\lambda_0 kT) \exp(\hbar c/\lambda_0 kT)/\{\exp(\hbar c/\lambda_0 kT) - 1\}^2 = 0
\]

If we now define \( \beta = \hbar c/\lambda_0 kT \), this equation can be written as: \( \beta/5 + \exp(-\beta) - 1 = 0 \). This is a transcendental equation, with a single solution for \( \beta \). Consequently, \( \lambda_0 T = \hbar c/k\beta = \text{constant} \).

By successive approximation, the solution can be found as \( \beta = 4.96511 \), which yields:

\[
\lambda_0 T = 6.62 \times 10^{-34} \times 3 \times 10^8 / (1.38 \times 10^{-23} \times 4.96511) = 2.898 \times 10^{-3} \text{ m K.}
\]

11.2 \( \lambda = h/\sqrt{(2mE)} \) so that \( V = h^2/(2me^2) \). Hence:

\[
V = (6.62 \times 10^{-34})^2 / \{2 \times 9.1 \times 10^{-31} \times 1.6 \times 10^{-19} \times (0.05 \times 10^{-9})^2\} = 601.98 \text{ V.}
\]

\[
p = \sqrt{(2mE)} = \sqrt{(2 \times 9.1 \times 10^{-31} \times 1.6 \times 10^{-19} \times 601.98)} = 1.3239 \times 10^{-33} \text{ Ns.}
\]

11.3 For \( z \leq 0 \), the wavefunction comprises incident and reflected waves, so assume the trial solution:

\[
\Psi = \Psi_0 \exp(jE t/\hbar) \{\exp(-jk_1 z) + R \exp(+jk_1 z)\} \quad \text{where} \quad k_1 = \sqrt{(2mE/\hbar^2)}
\]

For \( z \geq 0 \), the wavefunction comprises transmitted wave only, so assume the solution:

\[
\Psi = \Psi_0 \exp(jE t/\hbar) T \exp(-jk_2 z) \quad \text{where} \quad k_2 = \sqrt{2m(V - E)/\hbar^2}
\]

Boundary conditions:

Match \( \Psi \) at \( z = 0 \) \( \Rightarrow 1 + R = T \)

Match \( d\Psi/dz \) at \( z = 0 \) \( \Rightarrow k_1 \{1 - R\} = k_2 T \)

Solve the two simultaneous equations above to determine \( R \) and \( T \), giving:

\[
R = \frac{k_1 - k_2}{k_1 + k_2}
\]

\[
T = 2k_1 / \{k_1 + k_2\}
\]

The probability of reflection is \( |R|^2 = \left| \frac{k_1 - k_2}{k_1 + k_2} \right|^2 \)

If \( \exists < V \), \( k_2 \) is complex, so \( k_2 = -jk_2' \) where \( k_2' = \sqrt{2m(V - \exists)/\hbar^2} \)

In this case, the probability of reflection is \( |R|^2 = \left| \frac{k_1 + jk_2'}{k_1 - jk_2'} \right|^2 = 1 \).

11.4 \( \lambda_g = h/(eE_g) = 6.62 \times 10^{-34} \times 3 \times 10^8 / (1.6 \times 10^{-19} \times E_g) \)

\[
g = 1.24 \times 10^{-6} / E_g \text{ m} = 1.24/E_g \text{ µm.}
\]

For Ge:

\( \lambda_g = 2.12 \text{ µm} \)

For Si:

\( \lambda_g = 1.107 \text{ µm} \)

For GaP:

\( \lambda_g = 0.563 \text{ µm} \)

For InP:

\( \lambda_g = 0.918 \text{ µm} \)

For GaAs:

\( \lambda_g = 0.837 \text{ µm} \)

For C:

\( \lambda_g = 0.227 \text{ µm} \)

For SiO_2:

\( \lambda_g = 0.155 \text{ µm} \).
Choose Ge for the detector, and SiO$_2$ for the transparent window at $\lambda_0 = 1.55$ $\mu$m. For Ga$_{0.47}$In$_{0.53}$As, $\lambda_g = 1.698$ $\mu$m, so this would also be suitable as a detector material. The advantage over Ge would be improved sensitivity near $\lambda_g$, as it is a direct-gap material.

11.5 \[ E_{\text{Fi}} = \frac{E_c + E_v}{2} + \frac{(3kT/4) \log_e(m_{h^*}/m_{e^*})}{2} = E_v + \frac{E_g}{2} + \frac{(3kT/4) \log_e(m_{h^*}/m_{e^*})}{2} \]

Hence, for intrinsic InP:
\[ E_{\text{Fi}} = E_v + 1.35/2 + 3 \times 0.0259/4 \log_e(0.64/0.077) = E_v + 0.716 \text{ eV}. \]

11.6 At a temperature $T$, $n_i = \sqrt{(N_c N_v)} \exp(-E_g/(2kT))$.

Hence, for the three materials given, we get:
Ge: \[ n_i = \sqrt{(1.04 \times 10^{25} \times 6 \times 10^{24}) \exp{-0.66/(2 \times 0.0259)}} = 2.31 \times 10^{19} \text{ m}^{-3} \]
Si: \[ n_i = \sqrt{(2.8 \times 10^{25} \times 1.04 \times 10^{24}) \exp{-1.12/(2 \times 0.0259)}} = 6.95 \times 10^{15} \text{ m}^{-3} \]
GaAs: \[ n_i = \sqrt{(4.7 \times 10^{23} \times 7 \times 10^{24}) \exp{-1.42/(2 \times 0.0259)}} = 2.25 \times 10^{12} \text{ m}^{-3} \]

At a temperature $T + \Delta T$, $n'_i = \sqrt{(N_c N_v) \exp{(-E_g/(2k(T + \Delta T))}}$

If $\Delta T$ is small, $n'_i \approx \sqrt{(N_c N_v) \exp{(-E_g/(2kT)(1 - \Delta T/T)}} \approx n_i \exp(+E_g \Delta T/(2kT))$

Hence, if we write $n'_i = 2n_i$, then $\Delta T = 2kT^2 / E_g \log_e(2)$
So that $\Delta T = (2 \times 0.0259 \times 300 / 0.66) \log_e(2) = 16.32$ K.

11.7 Be: Gp IIA $\Rightarrow$ acceptor
C: Gp IVB, replaces As atom $\Rightarrow$ acceptor
Cd: Gp IIB $\Rightarrow$ acceptor;
Mg: Gp IIA $\Rightarrow$ acceptor
S: Gp VIB $\Rightarrow$ donor
Se: Gp VIB $\Rightarrow$ donor
Sn: Gp IVB, replaces Ga atom $\Rightarrow$ donor
Te: Gp VIB $\Rightarrow$ donor
Zn: Gp IIB $\Rightarrow$ acceptor.

For Zn-doping, the material is p-type.
\[ E_{FP} = E_v + KT \log_e(N_n/N_p) \]
\[ = E_v + 0.0259 \log_e(7 \times 10^{-24}/10^{18}) \text{ eV} = E_v + 0.408 \text{ eV}. \]

For complete ionization of the dopant atoms, the majority carrier density is $p \approx N_n = N_i = 10^{18} \text{ m}^{-3}$. The minority carrier density is then $n = n_i^2/p = (2.25 \times 10^{12})^2 / 10^{18} = 5 \times 10^6 \text{ m}^{-3}$.

11.8 For acceptors, the probability of ionization is the same as the probability of finding an electron at the acceptor level to perform the ionization. Hence we may write:
\[ F(E_a) = 1 / \{1 + \exp((E_a - E_F)/kT)\}. \]

Now: $E_a - E_F = (E_F - E_v) - (E_v - E_a) = 0.031 - 0.408 = -0.377$ eV.
Hence: $F(E_a) = 1 / \{1 + \exp(-0.377 / 0.0259)\} = 0.9999995$.
This is so close to unity that the assumption of complete ionization is a good one.
Sn is an n-type dopant. For donors, the probability of ionization is same as the
probability of not finding an electron at the donor level, i.e. \{1 - F(E_d)\}. Hence, \(N_d^+ = N_d \{1 - F(E_d)\}\).

11.9. The relationship between the Einstein A and B coefficients may be derived by
comparing the expression \(\rho(\omega) \, d\omega = A / [B_{12} \exp\{(\hbar c\omega)/kT\} - B_{21}] \, d\omega\) with the
black-body radiation law
\(\rho(\omega) \, d\omega = [(h\alpha n^3/\pi^2 c^3) / \{\exp(\hbar c\omega/kT) - 1\}] \, d\omega\).

Clearly, the term \(-B_{21}\) (which is associated with stimulated emission) in the former
translates to \(-1\) in the latter. If this process can be neglected, we would expect the
black body radiation law to modify to:
\(\rho(\omega) \, d\omega = (h\alpha n^3/\pi^2 c^3) / \exp(h\alpha c\omega/kT) \, d\omega \approx (h\alpha n^3/\pi^2 c^3) \exp(-h\alpha c\omega/kT) \, d\omega.\)

This relation was originally proposed by Wien in 1896 before the idea of stimulated
emission was suggested; it is a reasonable approximation to the correct result when
\(\exp(\hbar \omega/kT) \ll 1\) (i.e. for large values of \(\omega\) or low temperatures).

If \(\hbar \omega/kT \gg 1\), on the other hand, the black body radiation law may be approximated
as
\(\rho(\omega) \, d\omega \approx (h\alpha n^3/\pi^2 c^3) / [1 + (\hbar \omega/kT) - 1] \, d\omega \approx n^3 \omega^2 kT/\pi^2 c^3 \, d\omega.\)

This is the Rayleigh-Jeans law, originally derived using classical arguments. Its
failure as \(\omega \to \infty\) (a difficulty referred to as the ultraviolet catastrophe, since it
implies the existence of infinite energy at short wavelengths) was one of the main
spurs to the development of quantum theory.

11.10 In the steady-state, \(\partial \phi/\partial z = \{G(n - n_0)/v_g\} \phi = g_p \phi\), where \(g_p\) is the optical power gain
coefficient. Hence, \(g_p = G(n - n_0)/v_g\).

For a given power gain \(g_p\), we require an electron density of \(n \approx n_0 + g_p v_g/G\).
Assuming a refractive index of 3.5, we may approximate \(v_g\) by the phase velocity,
\(c/3.5\). For \(g_p = 10^6 \text{ m}^{-1}\), \(n_0 = 10^{24} \text{ m}^{-3}\) and \(G = 10^{-12} \text{ m}^3/\text{s}\) we therefore obtain
\(n \approx 10^{24} + 10^6 \times 3 \times 10^8 / (3.5 \times 10^{-12}) \approx 8.7 \times 10^{25} \text{ m}^{-3}\).
CHAPTER 12

12.1 The built-in voltage across a p-n junction is \( V_{bi} = (kT/e) \log_e \{N_a N_d /n_i^2\} \). Assuming that the intrinsic carrier concentration is \( n_i = 1.4 \times 10^{16} \text{ m}^{-3} \):

\[
V_{bi} = 0.0258 \log_e \{10^{24} \times 10^{23} / 2 \times 10^{32}\} = 0.873 \text{ V}.
\]

Similarly, the depletion layer width is \( w(0) = \sqrt{(2 \varepsilon V_{bi} /e) (N_a + N_d) / N_a N_d} \) with no bias voltage. Assuming that \( \varepsilon = 11.7 \varepsilon_0 \) for Si, we obtain:

\[
w(0) = \sqrt{(2 \times 11.7 \times 8.85 \times 10^{-12} \times 0.873 / 1.6 \times 10^{-19}) (10^{24} + 10^{23}) / (10^{24} \times 10^{23})} = 1.11 \times 10^{-7} \text{ m}, \text{ or } 0.111 \mu\text{m}.
\]

Under a forward-bias of \( V \) volts, the width becomes \( w(V) = \sqrt{1 - V/V_{bi}} \); hence, under a reverse bias of 5 V, \( w = 0.111 \sqrt{1 + 5/0.873} = 0.288 \mu\text{m} \).

Assuming that all the voltage is dropped across this highly resistive region, the average electric field across the depletion layer is then:

\[
E = (5 + 0.873) / (0.288 \times 10^{-6}) = 20.39 \times 10^6 \text{ V/m}.
\]

The drift velocity resulting from a given electric field is \( v_d = \mu E \). Hence, for electrons, \( v_d = 0.135 \times 20.39 \times 10^6 = 2.75 \times 10^6 \text{ m/s} \).

The transit time across the depletion layer is therefore:

\[
\tau = w/v_d = 0.288 \times 10^{-6} / 2.75 \times 10^6 = 0.105 \times 10^{-12} \text{ sec}, \text{ or } 0.105 \text{ psec}.
\]

For holes, \( v_d = 0.978 \times 10^6 \text{ m/s} \), and \( \tau = 0.294 \text{ psec} \).

The implication for photodiodes is that any photo-generated carriers will be swept out of the depletion region before they can recombine.

12.2. The current through the diode is:

\[
I = I_s \{\exp(eV/kT) - 1\}, \text{ where } I_s = A x J_s = \{D_e n_p /L_p + D_h p n /L_n\}.
\]

Since \( N_a \gg N_d \), the current will be carried almost entirely by electrons, so we may write \( J_s = D_e n_p /L_p \). The diffusion coefficient may be calculated using the Einstein relation, \( D_e = \mu_e kT/e \). Since \( \mu_e = 0.135 \text{ m}^2/\text{Vs} \), \( D_e = 0.135 \times 0.0258 = 3.48 \times 10^{-3} \text{ m}^2/\text{s} \).

The minority carrier concentration on the p-side is \( n_p = n_i^2 /N_a \); since the intrinsic carrier concentration in Si is \( 1.4 \times 10^{16} \text{ m}^{-3} \), \( n_p = 1.96 \times 10^{11} \text{ m}^{-3} \). Hence:

\[
J_s = 3.48 \times 10^{-3} \times 1.6 \times 10^{19} \times 1.96 \times 10^{11} / 1 \times 10^{-7} = 1.09 \times 10^{3} \text{ A/m}^2.
\]

Hence, for a cross-sectional area of \( 10^{-6} \text{ m}^2 \), \( I_s = 1.09 \text{ nA} \). For a forward voltage of 0.5 V, we then obtain \( I = 1.09 \times 10^{-9} \times \{\exp(0.5/0.0258) - 1\} = 0.284 \text{ A} \), or 284 mA.

For a lifetime \( \tau_e = 1 \mu\text{s} \), the diffusion length is:

\[
L_e = \sqrt{D_e \tau_e} = \sqrt{\{3.48 \times 10^{-3} \times 10^{-6}\}} = 5.9 \times 10^{-5} \text{ m}, \text{ or } 59 \mu\text{m}.
\]

Since this is much greater than \( L_p \), the thin-diode approximation is valid.

12.3 From Chapter 9, the index change in an electro-optic modulator is \( \Delta n_{eff} \approx -n^3 r \Gamma V/2g \).

Taking the breakdown electric field to be \( V/g = 3 \times 10^7 \text{ V/m} \), assuming that \( n = 3.2 \), that the electro-optic coefficient is \( r_{41} = -1.4 \times 10^{-12} \) and that the overlap factor is \( \Gamma = 0.5 \), we obtain:

\[
\Delta n_{eff} = 3.2^3 \times 1.4 \times 10^{-12} \times 3 \times 10^7 \times 0.5/2 = 3.5 \times 10^{-4}.
\]
From Chapter 3, the index change obtained by carrier injection is:
\[ \Delta n = -\Delta N e^2 \lambda_0^{-2} / 8\pi^2 c^2 m^* e_n, \]
where \( \Delta N \) is the density of injected carriers and \( m^* \) is their effective mass.

Since \( m^e_0 > m^e * \), only electron injection will be useful. Assuming that \( m^e_0 / m \) is 0.077, and that \( \lambda_0 = 1.5 \mu m \), we obtain:
\[ |\Delta n| = \Delta N x (1.6 x 10^{-19} x 1.5 x 10^{-6})^2 / (8 \pi^2 x (3 x 10^8)^2 x 0.077 x 9.1 x 10^{-31} x 8.85 x 10^{-12} x 3.2) = \Delta N x 4 x 10^{-27}. \]

Taking the upper limit of the injected electron density to be \( \Delta N \approx 10^{24} m^{-3} \), we then get \( |\Delta n| \approx 4 x 10^{-3} \). Carrier injection can therefore yield an index change an order of magnitude larger than the value obtained by the electro-optic effect.

12.4 Since the photocurrent generated in the PIN diode is defined as \( I_p = \eta P \lambda / h c \), the quantum efficiency is \( \eta = (h c / e) (I_p / P \lambda_0) \). This may be written as \( \eta = 1.24 (I_p / P \lambda_0) \) when \( \lambda_0 \) is measured in \( \mu m \).

For a diode of cross-section 0.1 mm\(^2\) and an illuminating beam of 1 mW/cm\(^2\) intensity, \( P = 0.1 x 10^6 x 10^{-3} x 10^4 = 10^{-6} \) W. The quantum efficiency is then:
\( \eta = 1.24 x 0.996 x 10^{-6} / (1.3 x 10^{-6}) = 0.95. \)

Assuming constant quantum efficiency and illumination intensity, \( I_p \) is proportional to \( \lambda_0 \). At \( \lambda_0 = 1.5 \mu m \), we would then expect \( I_p = 0.996 x 1.5 / 1.3 = 1.15 \mu A \). At \( \lambda_0 = 1.7 \mu m \), however, we would expect \( \eta = 0 \) and \( I_p = 0 \), since \( \lambda_0 > \lambda_g \) for Ga\(_{0.47}\)In\(_{0.53}\)As.

12.5 The power launched into the optical fibre is \( P_L = \eta e \eta c P' \), where \( P' \) is the internally generated power (given by \( P' = \eta' h c \tau_0 e \lambda_0 \)). \( \eta_e \) is the external efficiency, and \( \eta_c \) is the coupling efficiency into the optical fibre. The internal efficiency is then found as \( \eta' = P_e \lambda_0 / h c I' \eta c \), or as \( \eta' = P L \lambda_0 / 1.24 I' \eta c \), if \( \lambda_0 \) is measured in \( \mu m \).

Now, \( \eta_e = 1/n(n + 1)^2 \), so for GaAs (for which \( n = 3.5 \)) we obtain \( \eta_e = 0.0141 \).

Similarly, \( \eta_c = \{4n/(n' + 1)^2\} NA^2 \), so for an optical fibre (for which \( n' = 1.5 \)) with an NA of 0.3, we obtain \( \eta_c = 0.96 x 0.3^2 = 0.0864 \).

We therefore obtain \( \eta' = 50 x 10^6 x 0.82 / (1.24 x 50 x 10^3 x 0.0141 x 0.0864) = 0.543 \). Since \( \eta' = \tau / \tau_0 \), the radiative recombination time is \( \tau_0 = \tau / \eta' = 10 / 0.543 = 18.42 \) nsec.

When the LED is modulated, the ratio of AC to DC optical power generated is
\[ |P'/P| = 1 / \sqrt{1 + \omega^2 \tau}_c^2 \]. When this ratio has fallen to 3 dB below its DC value,
\[ |P'/P| = 1/2, \] so that \( \omega = \sqrt{3/\tau}_c \).

Hence, the 3 dB modulation bandwidth is \( \sqrt{3/(2\pi x 10 x 10^9)} \) Hz, or 27.6 MHz.

12.6 The results of Chapter 4 suggest that a medium with complex relative dielectric constant \( \varepsilon = \varepsilon'_i + j\varepsilon''_i \) will have a propagation constant \( k = k' + jk'' \). When \( \varepsilon''_i \ll \varepsilon'_i \), these quantities may be written in terms of a complex refractive index \( n = n' + jn'' \) as \( k' = 2\pi n' / \lambda_0 \) and \( k'' = k'n'' / n' \).
Now, $k''$ is the amplitude gain coefficient $g$, equal to $g_p/2$. We therefore deduce that a power gain $g_p$ will follow from the existence of an imaginary index $n''$ equal to $gn'/k'$ (or $gn'/k'$).

Similarly, the results of Chapter 9 show that the change in the real part of the effective index of a guided mode due to an index perturbation $\Delta n'$ can be found as an overlap between the index perturbation and the normalised transverse modal field, i.e. as $\Delta n'_{\text{eff}} = < \Delta n', |E|^2 >$.

We would therefore expect that an imaginary index perturbation $n''$ would alter the imaginary part of the effective index by $\Delta n''_{\text{eff}}$, found by analogy with the above as $\Delta n''_{\text{eff}} = < n'' , |E|^2 > = n'/k' < g , |E|^2 >$.

As a result of this change in effective index, the propagation constant $\beta'$ should alter to $\beta' + j\beta''$, where $\beta'' = \frac{2\pi \Delta n'_{\text{eff}}}{\lambda_0} = < g , |E|^2 >$. $\beta''$ is clearly the effective amplitude gain $g_{\text{eff}}$ and is found as an overlap between the normalised transverse field and the gain distribution. Similarly, the effective power gain is $g_{p\text{eff}} = 2g_{\text{eff}} = < g_p , |E|^2 >$.

Including a non-zero $\Gamma$, the rate equations modify to:

$$\frac{d\phi}{dt} = \frac{8n}{\tau_p} + \Gamma G \phi(n - n_0) - \phi/\tau_p \text{ and}$$
$$\frac{dn}{dt} = \frac{I/e}{n/\tau_e - \Gamma G \phi(n - n_0)}.$$

Solving these gives the conventional answer, namely that the power out of the cavity is $P = \frac{hc(1 - I)}{e\lambda_g}$ in the lasing regime. This time, however, not all the power is coupled out of the end mirrors; some is lost by absorption.

Including loss in both core and cladding, the threshold gain is:

$$\Gamma g_p = \Gamma \alpha_{\text{core}} + (1 - \Gamma) \alpha_{\text{clad}} + \frac{1}{L} \log_e \left( \frac{1}{R_1 R_2} \right).$$

In fact, the gain is clamped to this value throughout the lasing regime; this shows that the gain goes to make up absorption and output coupling losses, in the proportions shown.

The power lost from the cavity is split into power absorbed and power coupled out through the mirrors, in a similar way. The useful fraction therefore defined by the efficiency:

$$\eta = \left\{ \frac{1}{L} \log_e \left( \frac{1}{R_1 R_2} \right) \right\} / \Gamma \alpha_{\text{core}} + (1 - \Gamma) \alpha_{\text{clad}} + \frac{1}{L} \log_e \left( \frac{1}{R_1 R_2} \right)$$

and the usable output is $P = \eta hc(1 - I)/e\lambda_g$.

A sketch of the data suggests that the threshold is between 20 and 30 mA. Using the second and third points, we get $dP/dI = \frac{(2 - 0.5)}{(35 - 30)} = 0.3$ A/W. When lasing, $P = dP/dI \{1 - I\}$. Hence, the threshold current is $I_t = I - P / (dP/dI) = 30 - 0.5 / 0.3$ mA, or 28.33 mA.

The threshold current density is then $J_t = I/A$, where $A$ is the area through which the current flows. Here, we get $J_t = 28.33 \times 10^3 / (250 \times 3 \times 10^{-12}) = 0.3$ A/m$^2$.

Similarly, the output from one mirror is $P = \eta hc(1 - I)/2e\lambda_g$.
This can be written as $P = 0.62\eta (1 - I)/\lambda_g$ if $\lambda_g$ is measured in µm.
Differentiating this, we obtain \( \frac{dP}{dI} = 0.62 \frac{\eta}{\lambda g} \), so that \( \eta = (\lambda_g/0.62) \frac{dP}{dI} \). Hence, the efficiency is \( \eta = 0.82 \times 0.3 / 0.62 = 0.4 \).

Defining the average loss as \( \alpha_{av} = \Gamma \alpha_{core} + (1 - \Gamma) \alpha_{clad} \), the efficiency is \( \eta = \frac{1}{1 + L \log_e(1/R_1 R_2)} \). Hence, \( \alpha_{av} = (1/\eta - 1) \frac{1}{L \log_e(1/R_1 R_2)} \).

Taking \( R_1 = R_2 \approx 0.555 \) (normal incidence on a GaAs/air interface), \( 1/L \log_e(1/R_1 R_2) = 4710 \) m\(^{-1}\), so:

\[
\alpha_{av} = \frac{1}{0.4 - 1} \times 4710 = 7065 \text{ m}^{-1} .
\]

12.8 Start with the photon rate equation

\[
\frac{d\phi}{dt} = \beta n/\tau_r + G\phi(n - n_0) - \phi/\tau_p \tag{1}
\]

Now assume \( \phi(t) = \phi' + \phi''(t), n(t) = n' + n''(t) \) where \( \phi' \) and \( n' \) are DC values. Ignoring all second-order terms, Equation 1 becomes:

\[
\frac{d\phi''}{dt} = \beta(n' + n'')/\tau_r + G(\phi' + \phi'')(n - n_0) + G\phi'n'' - (\phi' + \phi'')/\tau_p \tag{2}
\]

However, Equation (1) implies that the DC terms satisfy:

\[
0 = \beta n'/\tau_r + G\phi'(n - n_0) - \phi'/\tau_p \tag{3}
\]

So that Equation 2 reduces to:

\[
\frac{d\phi''}{dt} = \beta n''/\tau_r + G\phi''(n - n_0) + G\phi'n'' - \phi''/\tau_p \tag{4}
\]

Now, the photon lifetime \( \tau_p \) is defined by \( 1/\tau_p = G(n - n_0) \). Using this, and ignoring the contribution of spontaneous emission, Equation 4 reduces to:

\[
\frac{d\phi''}{dt} = G\phi'n'' \tag{5}
\]

Now turn to the electron rate equation:

\[
\frac{dn}{dt} = I/ev - n/\tau_e - G\phi(n - n_0) \tag{6}
\]

Assume that \( I(t) = I' + I''(t) \), where \( I' \) is a DC value. Ignoring second-order terms, Equation 6 becomes:

\[
\frac{dn''}{dt} = (I' + I'')/ev - (n' + n'')/\tau_e - G(\phi' + \phi'')(n - n_0) - G\phi'n'' \tag{7}
\]

However, Equation 6 implies that the DC terms also satisfy:

\[
0 = I'/ev - n'/\tau_e - G\phi'(n - n_0) \tag{8}
\]

Using this, Equation 7 reduces to:

\[
\frac{dn''}{dt} = I''/ev - n''/\tau_e - G\phi''(n - n_0) - G\phi'n'' \tag{9}
\]

Using \( \tau_p \) once again, Equation 9 may be written as:

\[
\frac{dn''}{dt} = I''/ev - n''/\tau_e - \phi''/\tau_p - G\phi'n'' \tag{10}
\]

Eqns 5 and 10 may then be rearranged to yield a single second-order equation in \( \phi'' \):

\[
\frac{d^2\phi''}{dt^2} + 2\zeta\omega_c \frac{d\phi''}{dt} + (\omega^2_c + \phi''/\tau_p)\phi'' = G\phi''/ev \tag{11}
\]

Defining a resonant frequency \( \omega_c \) and damping factor \( \zeta \) as

\[
\omega_c = \sqrt{(G\phi'/(\tau_p)), \quad \zeta = (1/\tau_e + \tau_p\omega_c^2)/2\omega_c}
\]

Equation 11 may be written as

\[
\frac{d^2\phi''}{dt^2} + 2\zeta\omega_c \frac{d\phi''}{dt} + \omega_c^2\phi'' = \tau_p \omega_c^2 I''/ev \tag{12}
\]
Now, the photon flux out of the cavity is \( \Phi(t) = \phi(t)/\tau_p \), and the optical power output is \( P(t) = h\Phi(t)/\lambda_g \). Defining \( P(t) = P' + P''(t) \), where \( P' \) is the DC output, we obtain \( P' = h\nu\phi(1)/\lambda_g \tau_p \) and \( P''(t) = h\nu\phi''(t)/\lambda_g \tau_p \).

Substituting the former into Equation 12, we can rewrite the resonant frequency as \( \omega_r = \sqrt{(G\lambda_g P'/h\nu)} \).

Similarly, substituting the latter into Equation 13, we obtain:

\[
d\frac{d^2 P''}{dt^2} + 2\zeta\omega_r \frac{dP''}{dt} + \omega_r^2 P'' = \{P'' + \frac{hc\omega_r^2}{e\lambda_g}\}I''.
\]

12.9. Assume that the current step is described by \( I'(t) = 0 \) for \( t \leq 0 \), \( I'(t) = i' \) for \( t > 0 \).

Begin by finding the complementary function (CF), the solution to:

\[
d\frac{d^2 P''}{dt^2} + 2\zeta\omega_r \frac{dP''}{dt} + \omega_r^2 P'' = 0
\]

This can be found by assuming that \( P''(t) = p \exp(\gamma t) \), where \( \gamma \) is the solution of the auxiliary equation \( \gamma^2 + 2\zeta\omega_r \gamma + \omega_r^2 = 0 \), i.e. \( \gamma = -\zeta\omega_r \pm \omega_r \sqrt{\zeta^2 - 1} \).

For light damping, \( \gamma \approx -\zeta\omega_r \pm j\omega_r \), so the CF may be written as \( P''(t) = p'' \exp(-\zeta\omega_r t) \cos(\omega_r t + \psi) \), where \( p'' \) and \( \psi \) are constants.

Now find the particular integral (PI), the solution to the differential equation:

\[
d\frac{d^2 P''}{dt^2} + 2\zeta\omega_r \frac{dP''}{dt} + \omega_r^2 P'' = \{P'' + \frac{hc\omega_r^2}{e\lambda_g}\}i''.
\]

Clearly, the PI is \( P'' = \{P'' + \frac{hc\omega_r^2}{e\lambda_g}\}i'' \).

The total solution is then found as the sum of the PI and the CF, with any unknown constants chosen to satisfy the boundary conditions (BCs). We may therefore write \( P''(t) = \{P'' + \frac{hc\omega_r^2}{e\lambda_g}\}i' + p'' \exp(-\zeta\omega_r t) \cos(\omega_r t + \psi) \).

For the BCs, we may assume that \( \phi'' = 0 \) and \( n'' = 0 \) at \( t = 0 \). Note that Equation 5 in the solution to Question 12.8 implies that the latter condition reduces to \( d\phi''/dt = 0 \).

Hence, \( P'' = dP''/dt = 0 \) at \( t = 0 \), so that:

\[P''(t) = p''\{1 - \exp(-\zeta\omega_r t) \cos(\omega_r t)\} \text{ where } p'' = (hc\omega_r^2/e\lambda_g)\,i''.\]

A plot of the solution is shown below, for \( \zeta = 0.05 \). The output settles to its final value \( p'' \) only after a number of oscillations. There is also considerable overshoot.
12.10 For an N-element Y-junction-coupled array laser, the threshold gains of the transverse modes are given by

\[ g_\nu = \frac{1}{2L} \log_e \left( \frac{1}{f_\nu R_1 R_2} \right) \]

where \( \nu \) is the mode number, \( R_1 \) and \( R_2 \) are the end-facet reflectivities, \( L \) is the cavity length, and \( f_\nu = \frac{1}{2} \{ 1 + \cos[\nu \pi/(N+1)] \} \) is the transmission of the \( \nu \)th mode through the Y-junction network.

For \( N = 9 \), \( L = 500 \times 10^{-6} \) m, and \( R_1 = R_2 \approx (3.5 - 1)/(3.5 + 1) = 0.555 \) (assuming normal reflection at a GaAs/air interface), we obtain:

<table>
<thead>
<tr>
<th>( \nu )</th>
<th>( f_\nu )</th>
<th>( g_\nu )</th>
<th>( \nu )</th>
<th>( f_\nu )</th>
<th>( g_\nu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9755</td>
<td>1200 /m</td>
<td>6</td>
<td>0.3455</td>
<td>2238 /m</td>
</tr>
<tr>
<td>2</td>
<td>0.9045</td>
<td>1276 /m</td>
<td>7</td>
<td>0.2061</td>
<td>2755 /m</td>
</tr>
<tr>
<td>3</td>
<td>0.7939</td>
<td>1406 /m</td>
<td>8</td>
<td>0.0955</td>
<td>3524 /m</td>
</tr>
<tr>
<td>4</td>
<td>0.6545</td>
<td>1599 /m</td>
<td>9</td>
<td>0.0245</td>
<td>4886 /m</td>
</tr>
<tr>
<td>5</td>
<td>0.5000</td>
<td>1868 /m</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The mode with \( \nu = 1 \) clearly has the lowest threshold and will lase preferentially.
CHAPTER 13

13.2 The polynomial equation

\[ a(x, y) = xy a_{\text{GaP}} + x(1 - y) a_{\text{GaAs}} + (1 - x)y a_{\text{InP}} + (1 - x)(1 - y) a_{\text{InAs}} \]

may be rewritten as

\[ a(x, y) = \alpha + \beta x + \gamma y + \delta xy, \]

where

\[ \alpha = a_{\text{InAs}}, \quad \beta = (a_{\text{GaAs}} - a_{\text{InAs}}), \quad \gamma = (a_{\text{InP}} - a_{\text{InAs}}) \] and \[ \delta = (a_{\text{GaP}} - a_{\text{GaAs}} + a_{\text{InAs}} - a_{\text{InP}}). \]

It is then easy to show that compounds lattice-matched to InP are defined by the locus

\[ y = \frac{1 - (\beta/\gamma) x}{1 + (\delta/\gamma) x}. \]

Since \[ \beta = -0.4054 \text{ Å}, \quad \gamma = -0.1894 \text{ Å} \] and \[ \delta \ll \beta/\gamma. \]

We may then put:

\[ y \approx \frac{1 - (\beta/\gamma) x}{1 + (\delta/\gamma) x}, \]

or \[ y \approx 1 - 2.14x. \]

When \[ y = 0, \quad x = 1/2.14 = 0.47, \]

so lattice-matched indium gallium arsenide has the composition \[ \text{In}_{0.53} \text{Ga}_{0.47} \text{As}. \]

13.3 Evaluating the relevant derivatives of the solution given, we obtain:

\[ \frac{\partial C}{\partial x} = \left\{ \frac{S}{\sqrt{4 \pi D t}} \right\} \left( -x/2Dt \right) \exp\left[-x^2/(4Dt)\right] \]

\[ \frac{\partial^2 C}{\partial x^2} = \left\{ \frac{S}{\sqrt{4 \pi D t}} \right\} \left\{ x^2/4D t^2 - 1/2Dt \right\} \exp\left[-x^2/(4Dt)\right] \]

\[ \frac{\partial C}{\partial t} = \left\{ \frac{S}{\sqrt{4 \pi D t}} \right\} \left\{ x^2/4D t^3 - 1/2t^2 \right\} \exp\left[-x^2/(4Dt)\right] \]

Hence, by inspection, \[ \frac{\partial C}{\partial t} = D \nabla^2 C \] as required.

A normalised Gaussian distribution has:

\[ f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left[-x^2/(2\sigma^2)\right] \]

and \[ \int_{-\infty}^{\infty} f(x) \, dx = 1. \]

Since the function is symmetric about \( x = 0 \), it is clear that \[ \int_{-\infty}^{\infty} f(x) \, dx = 1/2. \]

Now, the concentration function \( C(x, t) \) may be written as

\[ C(x, t) = 2S \left\{ \frac{\sqrt{\pi D t}}{\sqrt{4 \pi D t}} \right\} \exp\left[-x^2/(4Dt)\right]. \]

By comparison with the previous result, we see that \[ \int_{-\infty}^{\infty} C(x, t) \, dx = 1/2 \left( 2S \right) = S. \]

The quantity \( S \) represents the surface density of dopant at \( t = 0 \).

13.4 The diffusion length is \( L_D = \frac{2}{\sqrt{\pi}} \left( \sqrt{Dt} \right) \), where \( D = D_0 \exp\left(-E_a/kT\right) \).

As given, \( D_0 = 10^{-1} \text{ cm}^2 \text{ s}^{-1} = 10^{-5} \text{ m}^2 \text{ s}^{-1} \).

At \( T = 900^\circ \text{C} = 1173 \text{K}, \)
\[ D = 10^{-1} \exp\left(-3 \times 1.6 \times 10^{-19} / (1.38 \times 10^{-23} \times 1173) \right) = 1.32 \times 10^{-18} \text{ m}^2 \text{ s}^{-1}. \]

Hence, \( L_D = \frac{2}{\sqrt{\pi}} \left\{ 1.32 \times 10^{-18} \times 8 \times 3600 \right\} = 3.9 \times 10^{-7} \text{ m}, \) or \( 0.39 \mu\text{m}. \)

Similarly, at \( T = 300^\circ \text{C} = 1273 \text{K}, \)
\[ D = 1.36 \times 10^{-17} \text{ m}^2 \text{ s}^{-1} \text{ and } L_D = 1.25 \mu\text{m}. \]

13.5 i) Evaluating the relevant derivatives of the solution given, we obtain:

\[ \frac{\partial C}{\partial x} = \left\{ \frac{S}{4\pi D t} \right\} \left( -x/2Dt \right) \exp\left[-(x^2 + y^2)/4Dt\right] \]

\[ \frac{\partial^2 C}{\partial x^2} = \left\{ \frac{S}{4\pi D t} \right\} \left\{ x^2/4D t^2 - 1/2Dt \right\} \exp\left[-(x^2 + y^2)/4Dt\right] \]

\[ \nabla^2 C = \left\{ \frac{S}{4\pi D t} \right\} \left\{ (x^2 + y^2)/4D t^2 - 1/Dt \right\} \exp\left[-(x^2 + y^2)/4Dt\right] \]

\[ \frac{\partial C}{\partial t} = \left\{ \frac{S}{4\pi D t} \right\} \left\{ (x^2 + y^2)/4D t^3 - 1/2t^2 \right\} \exp\left[-(x^2 + y^2)/4Dt\right] \]

Hence, by inspection, \[ \frac{\partial C}{\partial t} = D \nabla^2 C \] as required.
The solution represents the concentration distribution arising from a line source parallel to the z-axis, for times much greater than that required to diffuse the source material into the substrate.

ii) Diffusion from a strip of Ti metal may be modelled by the superposition of a distribution of line sources spaced along the line x = 0 from y = -w/2 to y = +w/2.

Using the solution above for a line source, we therefore obtain:

\[ C(x, y, t) = \left( \frac{S}{4\pi Dt} \right) \int_{-w/2}^{w/2} \exp\left\{-\frac{x^2 + (y - y')^2}{4Dt}\right\} \, dy' \]

\[ = \left( \frac{S}{4\pi Dt} \right) \exp\left(-\frac{x^2}{4Dt}\right) \int_{-w/2}^{w/2} \exp\left\{-\frac{(y - y')^2}{4Dt}\right\} \, dy' \]

Now, the error function is defined by:

\[ \text{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u \exp(-u^2) \, du \]

Putting \( u = \frac{(y' - y)}{L_D} \), where \( L_D = 2\sqrt{Dt} \), we obtain:

\[ C(x, y, t) = C \exp\left(-\frac{x^2}{L_D^2}\right) \left[ \text{erf}\left(\frac{w + 2y}{2L_D}\right) + \text{erf}\left(\frac{w - 2y}{2L_D}\right) \right] \]

where \( C \) is a constant.

Assuming now that the index change resulting from diffusion is proportional to the local Ti concentration, we obtain the index distribution of a diffused channel guide as:

\[ n(x, y) = n_s + \Delta n \exp\left(-\frac{x^2}{L_D^2}\right) \left[ \text{erf}\left(\frac{w + 2y}{2L_D}\right) + \text{erf}\left(\frac{w - 2y}{2L_D}\right) \right] \]

13.6 Referring to the figure below: by simple trigonometry:

\[ w_1 = r / \cos(\phi), \text{ and } w_2 = w_1 - s \tan(\phi) = r / \sin(\theta) - s / \tan(\theta) \]

For the <111> Si planes, \( \theta = 54.74^\circ \).

For a fibre diameter of 125 \( \mu m \), \( r = 62.5 \mu m \); for a clearance of 5 \( \mu m \), \( s = 2.5 \mu m \).

Hence, \( w_2 = 74.77 \mu m \). The groove width is then \( 2w_2 = 149.55 \mu m \).

The groove depth is \( h = r / \sin(\phi) - s = r / \cos(\theta) - s = 105.76 \mu m \).

13.7 The code given will generate the central section of a 5 mm-long directional coupler, as shown below. The guides are specified on mask level #1. The guide widths are both 7 \( \mu m \), and the inter-waveguide gap is also 7 \( \mu m \). The electrodes are specified on mask level #2 and are of width 50 \( \mu m \). Since the electrodes overlap the guides, the vertical electric field component can be used to control the device, so that z-cut, y-propagating LiNbO\(_3\) would be appropriate.
13.8 The figure below shows part of one end of the S-bend geometry, which is symmetric about the device centreline.

To satisfy the design constraints, we must have:
\[ R \{1 - \cos(\theta)\} = R\theta^2 / 2 = (150 - 7) \times 10^{-6} / 2 = 71.5 \times 10^{-6} \text{ m}. \]

Hence, the bend angle is \( \theta \approx \sqrt{71.5 \times 10^{-6} \times 2 / 30 \times 10^{-3}} = 0.069041 \text{ rad}. \)

The transition length is then:
\[ L = 2R\sin(\theta) \approx 2R\theta = 2 \times 30 \times 10^{-3} \times 0.069041 = 4.14 \times 10^{-3} \text{ m}, \text{ or } 4.14 \text{ mm}. \]

13.9 Assuming that negative electron resist and subtractive processing are used to fabricate the mask set, the shapes defined by the code will emerge as chrome blocks on the glass mask.

A typical Ti:LiNbO\textsubscript{3} process involving subtractive definition of the waveguide patterns and additive definition of the electrodes might then follow steps 1 - 9 below for guide fabrication, 10 - 15 for electrode fabrication and 16 - 19 for packaging.

1. Acid clean LiNbO\textsubscript{3} substrate
2. Coat with e.g. 800 Å Ti metal by evaporation
3. Spin-coat with negative photoresist
4. Bake to drive off resist solvent
5. Expose to mask #1 in mask aligner
6. Develop resist
7. Wet etch to remove unwanted Ti metal
8. Remove remaining photoresist
9. Diffuse Ti metal into LiNbO$_3$, e.g. for 8 hours at 1050°C in furnace
10. Spin-coat with positive photoresist
11. Bake to drive off resist solvent
12. Expose to mask #2 in mask aligner
13. Develop resist
14. Coat with e.g. 1000 Å Al metal
15. Lift off resist, together with unwanted Al metal
16. Cut and polish device ends
17. Package device
18. Bond electrode connections
19. Fibre pigtail

13.10 i) Preform volume = $1 \times \pi \times (20 \times 10^{-3}/2)^2 \ m^3 = \pi \times 10^{-4} \ m^3$.
Preform mass = $\pi \times 10^{-4} \times 2.6 \times 10^3 \ kg = 0.817 \ kg$.
Deposition time = $0.817 / 2 \times 10^{-3} \ min = 6.81 \ hours$.

ii) Conservation of mass in the fibre drawing process suggests that the preform and fibre velocities and diameters must be related by $v_p \pi (d_p/2)^2 = v_f \pi (d_f/2)^2$, so that

\[ d_f = d_p \times \sqrt{v_p/v_f}. \]

Hence:
\[ d_f = 20 \times 10^{-3} \times \sqrt{3.5 \times 10^{-3} / 60 \times 1.5} \ m \approx 125 \ \mu m. \]

The fibre length is then found as:
\[ L_f = \pi \times 10^{-4} / \{ \pi \times (125 \times 10^{-6} / 2)^2 \} = 25600 \ m, \ or \ 25.6 \ km. \]
CHAPTER 14

14.1 Assuming an acoustic velocity of $6.57 \times 10^3$ m/s, the acoustic wavelength at 500 MHz is $\Lambda = 6.57 \times 10^3 / 500 \times 10^6 = 1.314 \times 10^{-3}$ m, or 13.14 µm.

The K-vector associated with the acoustic grating then has magnitude:

$$K = 2\pi/\Lambda = 0.4782 \times 10^6 \text{ m}^{-1}.$$ 

For an effective index of 2.2 and an optical wavelength of 0.633 µm, the propagation constant is $\beta = 2\pi / 2.2 \times 0.633 \times 10^{-6} = 21.837 \times 10^6 \text{ m}^{-1}$.

Referring to the figure below: for Bragg diffraction of the 0th order beam into the +1th diffraction order, the transducer tilt angle should be chosen so that:

$$2\beta \sin(\phi) = K,$$

i.e. so that:

$$\phi = \sin^{-1}(K / 2\beta) = \sin^{-1}(0.4782 / (2 \times 21.837)) = 0.627^\circ.$$

For a transducer of 200 MHz bandwidth, the extreme operating frequencies are 400 and 600 MHz. Performing a similar calculation at (say) 400 MHz, we obtain:

$$\Lambda' = 16.425 \mu\text{m}, K' = 0.3825 \times 10^6 \text{ m}^{-1}.$$ 

The optimal transducer angle is now $\phi' = \sin^{-1}(0.3825 / (2 \times 21.837)) = 0.502^\circ$.

If the transducer is fixed at the angle $\phi$ found above, the optical beam will now be off-Bragg by an angle $\Delta\theta = 0.627 - 0.502 = 0.125^\circ$, or $2.18 \times 10^{-3}$ rads.

The angular bandwidth of a transmission grating of thickness $d$ is $\Lambda/d$, so the maximum allowable acoustic beamwidth is:

$$d = \Lambda / \Delta\theta = 16.425 \times 10^{-6} / 2.18 \times 10^{-3} = 7.5 \times 10^{-3} \text{ m}, \text{ or } 7.5 \text{ mm}.$$ 

14.2 The resistance of the conductor before scaling is $R = \rho L/(w \times t)$, where $\rho$ is the resistivity. After scaling down all dimensions by $\alpha$, the new resistance is:

$$R' = \rho(L/\alpha) / ((w/\alpha) \times (t/\alpha)) = \alpha R.$$ 

Regarding the conductor as one plate of a parallel-plate capacitor, the capacitance to the ground-plane before scaling is approximately $C = \varepsilon Lw/h$, where $\varepsilon$ is the dielectric constant of the insulator. After scaling, the new capacitance is $C' = \varepsilon(L/\alpha) \times (w/\alpha) / (h/\alpha) = C/\alpha$. Hence, $R'C' = RC$, so the time constant is unchanged by scaling.

This result is disadvantageous, because gate operating delays are expected to reduce with scaling, leaving interconnect delays as the limit on performance. Guided wave optical devices do not suffer from RC time constants, as they are transmission-line devices with real impedance. They are also immune to mutual interference.
14.3 The Gray codes for all two-bit numbers are:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>00</td>
</tr>
<tr>
<td>1</td>
<td>01</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

A two-bit integrated optic A-to-D converter with a Gray-coded output may be constructed using two Mach-Zehnder interferometers. In contrast to a binary-coded device, each MZI should have the same electrode length, and be fitted with additional electrodes designed to apply a static phase shift of $\pi$ (for the LSB MZI) and $\pi/2$ (for the MSB). The MZI outputs should then appear as the full lines in the figure below. After detection and thresholding, the outputs should be as shown by the shaded blocks.

14.4 The phase change accumulated in propagating along the fibre is $\phi = 2\pi n_{\text{eff}} L / \lambda_0$. Differentiating the above, we obtain $d\phi / dT = 2\pi / \lambda_0 \{ n_{\text{eff}} dL / dT + L d n_{\text{eff}} / dT \}$. Hence, $\Delta \phi = 2\pi / \lambda_0 \{ n_{\text{eff}} (1/L) dL / dT + d n_{\text{eff}} / dT \} L \Delta T$

Since $\alpha = (1/L) dL / dT$, $\Delta \phi = 2\pi / \lambda_0 \{ n_{\text{eff}} \alpha + d n_{\text{eff}} / dT \} L \Delta T$

Assuming that $L = 10$ cm, $\lambda_0 = 0.85 \mu$m, $n_{\text{eff}} \approx 1.458$, $\alpha = 0.004 \times 10^{-4} / ^\circ \text{C}$ and $d n_{\text{eff}} / dT = 1.1 \times 10^{-5} / ^\circ \text{C}$ we obtain:

$$d\phi / dT = (2\pi / 0.85 \times 10^{-6}) \{ 1.458 \times 0.004 \times 10^{-4} + 1.1 \times 10^{-5} \} \times 10^{-1} = 8.562 \text{ rad/}^\circ \text{C}.$$  

$d n_{\text{eff}} / dT$ is the dominant effect, contributing $\approx 95\%$ of the induced phase shift.

14.5 From the analysis of Section 14.6, we may express the output power from the lower arm of the interferometer as $P = P_0 \sin^2(\Phi/2)$, where $P_0$ is the input power and $\Phi$ is the phase delay between the two arms.

For a time delay $\tau$, $\Phi = \omega \tau$, where $\omega$ is the angular frequency, so that $P = P_0 \sin^2(\omega \tau/2)$. Identical outputs are thus obtained for time delays $\tau' = 2\nu \pi/\omega \pm \tau$.

The normalised sensitivity of the device to changes in $\tau$ is $(1/P_0) dP / d\tau = \omega \phi / 2 \sin(\omega \tau)$. Hence the sensitivity is zero when $\omega \tau = \nu \pi$. The sensitivity is maximum when $\omega \tau = \pi/2 + \nu \pi$.

14.6 Assuming for simplicity that the input signal is described by a continuous linear ramp in frequency, and ignoring any common phase delay, the output amplitude obtained from a Mach-Zehnder sensor in a linear FM system may be written as:

$$A(t) = 1/2 \{ s(t) - s(t - \tau) \}$$  
where $s(t) = A_0 \exp \{ j2\pi f(t) t \}$ and $f(t) = \{ f_1 + (f_2 - f_1) t / T \}$. 

R.R.A.Syms and J.R.Cozens  
Optical Guided Waves and Devices  
47
A(t) is therefore found as the difference between the linear FM input (full-line in (a) below) and a delayed copy of the input (dashed). Now, \( s(t - \tau) = s(t) \exp(-j2\pi\Delta f_1 t) \), where the beat frequency \( \Delta f_1 \) is defined by:
\[
\Delta f_1 = (f_2 - f_1)\tau/T,
\]
so that \( A(t) = \frac{1}{2} s(t) \{1 - \exp(-j2\pi\Delta f_1 t)\} \).

The output power is then
\[
AA^* = A_0^2/4 \{1 - \exp(-j2\pi\Delta f_1 t)\}\{1 - \exp(+j2\pi\Delta f_1 t)\} = P_0/2 \{1 - \cos(2\pi\Delta f_1 t)\}.
\]

This implies that the output is now time varying, at a frequency \( \Delta f_1 \) that is linearly dependent on the imbalance \( \tau \). Measurement of \( \Delta f_1 \) may therefore yield \( \tau \) without ambiguity, and with a constant sensitivity.

For a periodic sawtooth frequency modulation, the two combining signals are as shown in (b) below, and the beat frequency oscillates between the two values \( \Delta f_1 \) and \( \Delta f_2 = (f_2 - f_1) (T - \tau)/T \) as shown in (c) below.

14.7 From the analysis of Section 14.7, the time-of-flight difference is \( \Delta \tau = \Omega \tau D/c \), where \( \tau = L/c \), i.e. \( \Delta \tau = \Omega LD/c^2 \).

Assuming \( \Omega = 10^5 \) rad/sec, \( L = 10^3 \) m and \( D = 20 \times 10^{-2} \) m, we obtain:
\[
\Delta \tau = 10^5 \times 10^3 \times 20 \times 10^{-2} / (3 \times 10^8)^2 = 2.222 \times 10^{-20} \text{ sec}.
\]

This time delay would not be directly measurable. The corresponding phase difference is \( \Delta \Phi = \omega \Delta \tau = 2\pi \Delta \tau / \lambda_0 \).

Assuming that \( \lambda_0 = 0.85 \) \( \mu \)m, we obtain:
\[
\Delta \Phi = 2\pi \times 3 \times 10^8 \times 2.222 \times 10^{-20} / 0.85 \times 10^{-6} = 4.927 \times 10^{-5} \text{ rads}.
\]

14.8 i) The response \( P_n(\lambda_0) = \text{sinc}^2\{\pi(\lambda_0 - \lambda_n) / \Delta \lambda\} \) is shown in (a) below. \( \Delta \lambda \) represents the approximate bandwidth of the filter, or, alternatively, \( |\lambda_0 - \lambda_n| = \Delta \lambda \) defines the positions of the first zeros in the filter response.
ii) For $\Delta \Lambda \geq \Delta \lambda$, the largest interchannel crosstalk will occur when the channel separation is such that the $n+1^{th}$ channel coincides with the first sidelobe in the response of the $n^{th}$ filter, as shown in (b) below.

At this point, $\Delta \Lambda = 3\Delta \lambda/2$, and the likely crosstalk is:
\[
sinc^2\left(\frac{3\pi}{2}\right) = \left(\frac{2}{3\pi}\right)^2 = 0.04503, \text{ or } -13.46 \text{ dB}
\]

14.9 The directional coupler will be in the cross state only near zero volts, when the two guides are synchronous. For a frequency of 1 GHz, the time between zero-crossings is $T = 1 / (2 \times 10^9) = 0.5 \times 10^{-9}$ sec, or 0.5 nsec.

Near zero volts, the drive voltage varies linearly with time. In this region, the cross-coupled output will therefore vary following the directional coupler switch characteristic, and will appear as a short pulse. For a CW optical input, the output will be as shown below.

![Diagram showing directional coupler and optical output](image-url)
Assuming that the drive signal may be written as \( V(t) = V_0 \cos(\omega_m t) \), where \( V_0 = 100V \), the voltage gradient near the zero crossings has magnitude \( \frac{dV}{dt} = V_0 \omega_m \). The time taken to reach the switch voltage of \( V_s = 10V \) is thus:

\[
t = \frac{V_s}{\left(\frac{dV}{dt}\right)} = \frac{10}{100 \times 2\pi \times 10^9} = 1.592 \times 10^{-11} \text{ sec.}
\]

The pulse width is therefore \( \Delta t = 2t = 31.83 \) psec. More generally, the device acts as a periodic sampling 'gate', open for a short time at regular intervals. Several such devices, arranged to open at different times could therefore be used to separate the channels in an optical TDM communications system.

14.10 The requirement for a 1 mW signal at the detector places a constraint of 10 dB maximum loss in the system, assuming a source power of 10 mW.

The power budget is:

- 25\% launch efficiency
- 10 splices, each contributing 0.1 dB loss
- Propagation through 50 km of 0.2 dB/km SM fibre

Total 6 dB loss

1 dB loss

10 dB loss

17 dB loss

The optical amplifier must therefore have a gain of \( \geq 7 \) dB. Assuming a gain of 0.25 dB/m from the erbium-doped fibre, a 28 m fibre length is required.