Springer Series in Materials Science 159

Xavier Marie
Naci Balkan *Editors*

Semiconductor Modeling Techniques

Springer

Springer Series in Materials Science

Volume 159

For further volumes:
http://www.springer.com/series/856

The Springer Series in Materials Science covers the complete spectrum of materials physics, including fundamental principles, physical properties, materials theory and design. Recognizing the increasing importance of materials science in future device technologies, the book titles in this series reflect the state-of-the-art in understanding and controlling the structure and properties of all important classes of materials.

Xavier Marie · Naci Balkan
Editors

Semiconductor Modeling Techniques

Springer

Editors
Xavier Marie
Laboratoire de Physique et Chimie des Nano-Objets
INSA—Université de Toulouse
135 avenue de Rangueil
31077, Toulouse cedex
France

Naci Balkan
School of Computer Science
University of Essex
Essex CO4 3SQ
UK

ISSN 0933-033X
ISBN 978-3-642-27511-1 ISBN 978-3-642-27512-8 (eBook)
DOI 10.1007/978-3-642-27512-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012939128

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Advances in high performance semiconductor electronic and optoelectronic devices over the last 40 years are due in large measure to the developments in growth, fabrication and experimental techniques. However, predicting and understanding the operation of these novel devices would not have been possible without the parallel developments in theoretical and modelling techniques.

This book aims to provide a comprehensive account of the theoretical and modelling techniques used in semiconductor research and is suitable for researchers and postgraduate students. It covers the theoretical description of electronic and optical properties of semiconductors and devices, and techniques used to understand the electronic band structure and the band gap engineering where the strain and quantum confinement can be optimised for ideal device performance. It also covers the fundamental theory of lasers and semiconductor optical amplifiers, as well as the main computational techniques used to understand linear and nonlinear electronic transport.

The book is based on the lectures given by leading experts in the EU-COST Action MP0805 training school held in Izmir in April 2011.

Toulouse — Xavier Marie
Colchester — Naci Balkan

Contents

Chapter 1
Introduction to Semiconductor Heterostructures

R. Ferreira

Abstract We present an introduction to the physics of semiconductor nanostructures. We review the main assumptions of the $k \cdot p$ method applied to a bulk crystal and then focus on the envelope function method to describe the electronic states of semiconductor heterostructures.

1.1 Introduction

Semiconductor heterostructures have become in the last decades a cornerstone in applicative and fundamental researches on condensed matter. Progresses in the field of semiconductor nano-objects are actually the result of concomitant immense developments in various areas: growth and processing techniques, mastering of materials properties, impressive advances in experimental set-ups and spectroscopic tools, new concepts of structures and devices functioning, the understanding of new physical aspects regarding the coupling of electrons, photons and phonons in low-dimensional systems.

Most such developments are, so to speak, relatively recent, i.e. related to the last two or three decades. The seminal proposal by Esaki and Tsu for band gap engineering by stacking layers of different semiconductors, as well as the early sample realisations, date instead from the 1970s. In the 1980s, systematic optical studies in quantum wells and resonant transport through thin barriers were undertaken. At that time, the main concepts underlying the physics of electrons in semiconductor heterostructures were established. In particular, experimental and theoretical efforts

R. Ferreira (✉)
Laboratoire Pierre Aigrain, Ecole Normale Supérieure
24 rue Lhomond, Paris 75231 Cedex 05, France
e-mail: robson.ferreira@lpa.ens.fr

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_1,

have permitted to clearly establish quantum confinement and tunnel coupling as key concepts in the field. The appearance of new structures and concepts marked the 1990s. Amongst them we quote (non exhaustively): (i) fabrication and first theoretical modelling of quantum wires and dots, which pushed forward the field of lower (quasi one and zero) dimensional electronic systems; (ii) the exploitation of more and more elaborate stacks of alternating well and barrier layers, like in semiconductor superlattices and periodic structures for quantum cascade lasers (QCL) and (iii) the coming into play of dielectric light confinement altogether with electronic confinement, which launched the field of strong light–matter coupling in semiconductor heterostructures. Quantum dots (QDs), semiconductor microcavities and quantum cascade structures brought to the stage many new aspects of heterostructure physics: the realisation of quasi-atomic structures in a crystalline matrix, the ability to control the light–matter coupling and the study of new elementary quasi-particles (cavity polaritons) in condensed matter, and a new concept (unipolar scheme) for laser light generation in the mid-IR spectral domain. In the last decade the multiple developments of these fields can be observed, as well as the rapid growth of new ones, like nanoacoustics and nanophotonics. A key concept underlying the research efforts in this period is "coherence". Indeed, many efforts in the semiconductor nanostructure domain have been devoted, in parallel with worldwide studies in many fields, to the realisation of physical systems operating in the (ideally) pure quantum regime, as motivated by the potential implementation of future opto-electronic devices. This has in particular pushed the understanding of decoherence sources in the nano-objects, revealing both the profound influence of the nano-object environment on its electronic and optical properties, as well as many particular aspects of electron–phonon coupling in low-dimensional systems. Additionally, crossing domains have emerged and become mature fields by themselves, like: (a) studies of quantum dots in various kinds of optical (micro-) cavities; (b) studies of strong light–matter couplings in quantum-cascade-like structures (inter subbands polaritons); (c) use of superlattices and quantum dots for the optical generation of acoustic signals. It is finally worth stressing the continuous improvement of the theoretical description of electronic states of nanostructures, in parallel and jointly with advances in sample realisation and experimental studies (we shall come back to this point later on in this chapter; see also Paul Voisin's contribution in this volume).

Delimiting the field of semiconductor heterostructures is nonsense: in fact, semiconductor nano-objects are today present in numerous areas, and concepts and techniques from many different fields contribute to the advance of nanostructured semiconductor physics. Thus, interdisciplinary concepts and efforts have become a driving force of research nowadays. To illustrate this point, and without aiming to be exhaustive, one can quote: (i) the tremendous developments of the physics of an electron gas confined near a semiconductor heterojunction; (ii) the realisation of structures containing either semiconductor or metallic layers, in the field of spintronics, aiming at developing hybrid devices for the concomitant generation and detection of spin-polarised carriers; (iii) embedding semiconductor-based systems (like QDs and QCL structures) in photonic crystals, aiming at improving/controlling

light–matter coupling and developing light sources with specific characteristics; (iv) the impressive number of transport and optical studies done on carbon nanotubes.

This book aims at presenting an introduction to the physics of semiconductor heterostructures (for a comprehensive review see [1]; see also [2, 3]). This cannot be envisioned without recalling the physics of bulk semiconductors. Indeed, in a rough image, a semiconductor nanostructure is made of juxtaposed pieces of different semiconductors, and, as we will recall, their electronic levels and optical properties retain many fundamental aspects of the corresponding ones for the original materials. The first part is thus devoted to a quick review of bulk properties: the electron Bloch states and the effective mass description of near-edge conduction and valence bands states (see e.g. [4–6]). In the second part we introduce the envelope function method to describe electronic levels in perturbed semiconductors. Indeed, an overwhelming number of studies in semiconductor physics are intimately related to the presence of some kind of perturbation: doping with shallow impurities provide carriers; a d.c. bias triggers a current; the optical characteristics are the response of the crystal to an external e.m. excitation. The building of a theoretical framework allowing the description of electronic levels in perturbed crystals is thus naturally the object of immense efforts from the very beginning of semiconductor physics [7]. The envelope function method appeared as a versatile approach for many cases of interest, namely whenever the perturbation strength is "weak" enough (as compared to typical band energy widths or separations) and "slowly varying" in space (as compared e.g. to the lattice period). The simplest heterostructures are introduced in the third part. We also present an envelope function description of their one-electron states. As we shall see, although the nanostructuration process introduces as a rule a large (in energy) and abrupt (i.e., at the cell size scale) variation of the crystal potential, it can nevertheless be properly tackled by the envelope function method, provided some assumptions, to be discussed later on, are retained in the model.

1.2 Electronic States of Bulk Semiconductors

We review in the first part of this paragraph the fundamental concepts related to the formation of energy bands in a semiconductor and the principal characteristics of their electronic states (Bloch functions). The second part is devoted to the $\vec{k} \cdot \vec{p}$ method.

1.2.1 Fundamental Concepts

The key concept in a perfect crystal is *translational invariance*, which characterises the presence of a spatial order at the microscopic level. To describe translational invariance we define a set of lattice vectors: the Bravais ensemble $\{\vec{R}\}$. In this way, two points in space differing by a lattice vector are physically equivalent. As a consequence, any physical property (i.e. any observable) of an idealised crystal in

its ground state is invariant under translation by a lattice vector. For instance, the density of electronic charge is a periodic function on the crystal lattice. This is a very fundamental statement, to which any theoretical model aiming at describing the physical properties of a semiconductor should conform, and distinguishes from the very beginning a crystal from an atomic or molecular system in empty space. Strictly speaking, a real sample is finite in size. Nonetheless, the initial assumption of strict periodicity, which applies only for infinite size material, turns out to be a very robust concept in "large enough" systems. Deviations from this infinite-size model for a crystal introduce surface effects, but do not affect its *bulk* properties. Also, interface effects will be of primary importance in nanostructures. For the moment, we keep considering the assumption of translational invariance and shall have the opportunity to come back again to surface and interface effects later on.

From this single assumption, a full set of physical and theoretical results follows. At the outset, it is worth recalling the very useful mathematical concept of *reciprocal lattice*. Indeed, any periodic function in the real lattice can be decomposed in Fourier series within an ensemble of reciprocal lattice vectors $\{\vec{K}\}$ obeying $\vec{K} \cdot \vec{R} = 2\,\pi\,n$, with n an arbitrary integer. We thus possess two equivalent discrete sets of vectors to describe a perfect crystal: $\{\vec{R}\}$ and $\{\vec{K}\}$. The two ensembles share some important properties: (i) the vectors of the ensemble form a periodic lattice in either real or reciprocal spaces, so that each vector can be written in terms of elementary basis vectors $\vec{R} = i_1\vec{a}_1 + i_2\vec{a}_2 + i_3\vec{a}_3$ and $\vec{K} = j_1\vec{b}_1 + j_2\vec{b}_2 + j_3\vec{b}_3$, where $i_{1,2,3}$ ($j_{1,2,3}$) are integers and $\vec{a}_{1,2,3}$ ($\vec{b}_{1,2,3}$) are the generating basis in the real (reciprocal) space; (ii) the volume span by the basis vectors form a unitary cell, rigid translations of which within the lattice (i.e. by using the different combinations of $i_{1,2,3}$ or $j_{1,2,3}$) permits completely filling the space without overlap. The choice of the basis vectors, and thus also of the unitary cells is however arbitrary. There is nevertheless one particular choice that more clearly reflects the symmetry properties of the crystal: this is the Wigner-Seitz cell in real space and its corresponding reciprocal lattice counterpart, the Brillouin Zone (BZ). These particular unit cells, definitions and their constructions in the two spaces are presented and discussed in many workbooks (see e.g. [4, 5]). In the following, we shall implicitly assume the Wigner-Seitz cell in real space and first Brillouin cell in reciprocal space.

Our principal objective is to discuss a method for calculation of electronic states in a semiconductor. This task is of course impossible for a real system, i.e. constituted of a (virtually) infinite number of carriers of either charges: electrons and nuclei. Two approximations are usually invoked to overcome this difficulty: the Born-Oppenheimmer and the Hartree-Fock approximations. The first allows disentangling the nuclei and electron motions, whereas the second replaces the true N-electrons problem into an effective one-electron problem. These assumptions, exhaustively discussed in the literature, will not be detailed here. We shall only recall two major consequences of them. (i) The problem of finding the electronic eigenstates of a crystal fits a one-electron problem $H_{\rm cr} = T + V_{\rm cr}(\vec{r})$, where $T = \vec{p}^2/(2m_0)$ is the kinetic energy operator (m_0 the bare electron mass) and $V_{\rm cr}(\vec{r})$ the effective crystal potential. (ii) However, the true expression of $V_{\rm cr}(\vec{r})$ is hardly known: although the

bare interactions are coulombic-like in the original N-body problem, the simplifications brought about by the Born-Oppenheimer and Hartree-Fock schemes considerably influence the profile of the effective crystal potential. A lot of considerable theoretical effort has been put in evaluating its precise form or, equivalently, different matrix elements involving $V_{\rm cr}(\vec{r})$. We shall here instead adopt a more pragmatic strategy, namely, exploit as much as possible the *symmetry properties* of $V_{\rm cr}(\vec{r})$ and push as far as possible the description of the crystal eigenstates. As we shall see in the following, we will end up with a semi-phenomenological description of the electronic states in terms of effective parameters, which are ultimately related to different matrix elements involving the eigenstates of $H_{\rm cr}$.

Accounting thus for the two aforementioned approximations, we shall look in the following for the stationary eigenstates of one electron in a static potential $V_{\rm cr}(\vec{r})$. Although not analytically known, the effective potential should comply with translational invariance: $V_{\rm cr}(\vec{r}+\vec{R}) = V_{\rm cr}(\vec{r})$ for any $\vec{R}$. Mathematically speaking, this constraint on $V_{\rm cr}(\vec{r})$ classes the possible one-electron eigenstates of $H_{\rm cr}$ into a very particular kind of solutions: those fulfilling the Bloch condition, namely, any solution $\Psi(\vec{r})$ verifying $\Psi(\vec{r}+\vec{R}) = \exp\{i\theta(\vec{R})\}\,\Psi(\vec{r})$, where $\theta(\vec{R})$ is an $\vec{R}$-dependent phase. Note that $\theta(\vec{R})$ should be real to ensure wavefunction normalisation and crystal invariance of any physical property depending upon the electron charge density $|\Psi(\vec{r})|^2$. It can also be readily checked that the term $\exp\{i\theta(\vec{R})\}$ is simply the eigenvalue of the operator $T_{\vec{R}}$ performing a translation by $\vec{R}$ in the real space: $T_{\vec{R}}\ f(\vec{r}) = f(\vec{r}+\vec{R})$ for any function $f(\vec{r})$. More information about this phase can be obtained by the fact that two translation operations in space commute, and thus $\theta(\vec{R}_1 \pm \vec{R}_2) = \theta(\vec{R}_1) \pm \theta(\vec{R}_2)$. These results can be cast in the form $\theta(\vec{R}) = \vec{k}\cdot\vec{R}$, where $\vec{k} = (k_x, k_y, k_z)$ is a constant (and so far unspecified) three-dimensional real vector. The $\vec{k}$ wavevector can thus be used to label the eigenstates: $\Psi_{\vec{k}}(\vec{r})$.

In order to proceed, we note that a Bloch state can be written in the form

$$\Psi_{\vec{k}}(\vec{r}) = e^{i\,\vec{k}\cdot\vec{r}}\, u_{\vec{k}}(\vec{r}) \tag{1.1}$$

where $u_{\vec{k}}(\vec{r})$ is a periodic function over the crystal lattice: $u_{\vec{k}}(\vec{r}+\vec{R}) = u_{\vec{k}}(\vec{r})$. It is evident that this form fulfils the Bloch property $\Psi(\vec{r}+\vec{R}) = e^{i\,\vec{k}\cdot\vec{r}}\,\Psi(\vec{r})$ imposed by the existence of translational invariance. Also, one can easily show that the periodic part of the Bloch state is solution of the stationary eigenvalue problem

$$[\,H_{\rm cr} + \hbar\vec{k}\cdot\vec{p}/m_0\,]\,u_{\vec{k}}(\vec{r}) = [\,E(\vec{k}) - \hbar^2\vec{k}^2/(2m_0)\,]\,u_{\vec{k}}(\vec{r}) \tag{1.2}$$

The determination of the total eigenstate $\Psi_{\vec{k}}(\vec{r})$ is replaced by the determination of the periodic solutions $u_{\vec{k}}(\vec{r})$. Of course, this cannot be accomplished without knowledge of $V_{\rm cr}(\vec{r})$. However, it is important to realise that such an eigenvalue problem admits more than one solution. Actually, there should be infinity of solutions for a given $\vec{k}$ value. We need correspondingly a new label (or a new set of labels)

to fully determine the eigenstates. This label is the *band* index, n. The reason for this name is the following: the eigenfunctions read now as $\Psi_{n,\vec{k}}(\vec{r})$, whereas the corresponding energies $E_n(\vec{k})$ become (quasi-) continuous functions of the (quasi-continuous) variable $\vec{k}$, so that $E_n(\vec{k})$ describes an energy band when n is kept constant and $\vec{k}$ is varied inside the first BZ.

Let us now stress the fact that the wavevector labelling the crystal eigenstates can be restricted to the first Brillouin zone. Indeed, two wavevectors differing by a reciprocal lattice vector $\vec{K}$ have the same Bloch phase (as can be immediately shown by using $\vec{K} \cdot \vec{R} = 2\,\pi\,n$). The values that $\vec{k}$ can assume are usually specified by imposing the so-called Born – von-Karm boundary conditions to the problem: the idealised infinite crystal is replaced by a finite-size sample with periodic eigenstates, i.e.

$$\Psi_{n,\vec{k}}(x + L_x, y, z) = e^{i\,k_x\,L_x}\,\Psi_{n,\vec{k}}(\vec{r}) \equiv \Psi_{n,\vec{k}}(\vec{r}) \tag{1.3}$$

for a translation along the Ox direction over the whole sample size $L_x = N_x a_x$, where a_x is the lattice period along the Ox direction and N_x the number of crystals sites along this same direction. This restricts the k_x wavevectors to the ensemble of N_x values: $\{2\pi n_x/N_x\}, 0 \leq n_x < N_x$. Note that there is thus as many allowed k_x values as unity cells along the Ox direction, and that k_x values outside this interval are redundant since they lead to the same Bloch phase: indeed, for $k'_x = k_x + 2\pi i_x/L_x$ there is $\exp\{ik'_x L_x\} = \exp\{ik_x L_x\}$ for any integer $i_x \neq 0$. The ensemble $\{2\pi i_x/L_x\}$ is equal to the previously defined reciprocal space wavevectors (here in one dimension). Thus, it results that k_x can be restricted to values inside the first Brillouin Zone. The same results hold for the two other real space directions, defining analogously k_y and k_z. In conclusion, the two principal results that follow the Born – von-Karm assumption are that:

(i) the non-redundant wavevectors can be restricted to the first Brillouin zone of the crystal lattice;
(ii) there are as many $\vec{k}$ allowed wavevectors (and thus one-electron eigenstates *per band*, without considering spin degeneracy) as unity cells in the crystal.

Note that the first result does not follow straightforwardly from the effective eigenvalue problem defining $u_{\vec{k}}(\vec{r})$. However, it can be demonstrated that the set of eigensolutions obtained when replacing $\vec{k}$ by $\vec{k} + \vec{K}$ in (1.2) is actually $\vec{K}$-independent (see e.g. [4, 5]): $u_{n,\vec{k}+\vec{K}}(\vec{r}) = u_{n,\vec{k}}(\vec{r})$ (except possibly by an absolute constant phase); $E_n(\vec{k} + \vec{K}) = E_n(\vec{k})$. As a consequence, both the periodic functions and the energy dispersions (i.e. energy variation as a function of the wavevector) related to Bloch states are periodic functions in the reciprocal space.

Note additionally that the second result (ii) is essential, since it allows recovering extensibility for average physical properties regarding a finite-size sample. It could, actually, surprise us, owing to the sample-dependent nature of the result. Nonetheless, the sample-to-sample variations are meaningless in practice, since for a fixed band n the energy difference between states related to two consecutive $\vec{k}$ values is very small (it is actually roughly inversely proportional to the squared inverse of the

characteristic sample size), and thus impossible to be observable in any actual (i.e., macroscopic) sample. That implies that any small (as compared to the BZ) volume Δ_k in reciprocal space contains a large number of states: $\Delta_k \Omega_{\rm cr}/(2\pi)^3$, where $\Omega_{\rm cr}$ is the crystal volume (a result that can be easily demonstrated by recalling that one state occupies the length $2\pi/L_x$ of the first BZ of a one-dimensional crystal, and thus a volume $(2\pi/L_x)\ (2\pi/L_y)\ (2\pi/L_z)$ in a three-dimensional crystal). Finally, in any practical calculation involving the whole set of eigenstates of a given energy band, we will be authorized to replace any summation over $\vec{k}$ as follows:

$$\sum_{\vec{k}} F(\vec{k}) \ \rightarrow \ \frac{\Omega_{\rm cr}}{(2\pi)^3} \int d\vec{k}\, F(\vec{k}) \tag{1.4}$$

with $d\vec{k} = dk_x\, dk_y\, dk_z$, provided the function $F(\vec{k})$ varies "slowly enough" with $\vec{k}$ (its Fourier transform varies "slowly enough" in real space as compared to the lattice parameter).

Let us now come back to the periodic functions $u_{n,\vec{k}}(\vec{r})$. As mentioned, we cannot calculate $u_{n,\vec{k}}(\vec{r})$ without the knowledge of $V_{\rm cr}(\vec{r})$. However, similar to the existence of Bloch states (and the natural introduction of wavevectors as well as the existence of energy bands) that directly follows from the translational symmetry, one can proceed further and obtain a deeper insight into the crystal eigenstates by considering another very general symmetry property of $V_{\rm cr}(\vec{r})$, namely, the fact that $V_{\rm cr}(\vec{r})$ should equally reflect the *local* invariance of the crystal lattice. As far as rotational symmetry is considered, of course, there is no infinitesimal rotational invariance in a crystal, as one finds for an isolated atom (in this latter case the invariance is related to the isotropy of space, whereas a crystal medium is intrinsically anisotropic and physical properties are isotropic only in certain limits and/or under certain conditions). However, it is quite intuitive that there is angular isotropy inside a crystal for *finite* rotations in space. For instance, a cubic lattice with a single atom per site is invariant under an $n\pi/2$ rotation around any of its principal axis, with n an arbitrary integer. The stationary eigenstates should reflect this symmetry. Such rotations are better dealt with within the group theory formalism. Without going into the details, one may say, on very general grounds, that the Bloch eigenstates related to high symmetry points in the Brillouin zone should be invariant under a certain number of local transformations that leave invariant the crystal lattice structure. Such local transformations include both finite rotations as well as eventual reflexions through given plans. By high symmetry points we mean particular points (values of $\vec{k}$) of the Brillouin zone, and at these points only the wavefunctions fulfil specific symmetry requirements.

For the sake of this review, we shall concentrate on semiconductors with a particular lattice structure: the so-called zinc-blend lattice. This particular structure includes a large variety of semiconductors that will be considered in the following: III-V (GaAs, InAs, AlAs, ...) and II-VI (CdTe, CdSe, ...) materials. Group IV materials (Ge, Si) have a diamond structure, which corresponds to the zinc-blend structure with two identical atoms per unity cell. Conversely, the zinc-blend structure is the diamond one with two different atoms per unit cell.

It turns out that the $k = 0$ point of the BZ is particularly important in zinc-blend materials and has been given a particular name: the Γ point. It is a "high-symmetry" point, i.e. the solutions (eigenfunctions) related to this point display well-defined symmetry properties. Let us now quote two important results that will be used in the following:

(a) It is known from experiments that the fundamental band gap of the semiconductors we will be interested in (like GaAs) occurs at $k = 0$ and also that most of the optical and low bias transport characteristics of such materials involve electronic states (of the uppermost filled valence band and/or low lying empty conduction band) with small wavevectors, i.e. conduction and valence states near the Γ point. Moreover, other $k = 0$ edges are energetically far below the upper most filled valence band or far above the low-lying empty conduction band. As a consequence, in most semiconductors and semiconductor-based heterostructures we will mostly be concerned with a small part of the crystal band structure, in both $\vec{k}$ and energy axes, namely, with states around $k = 0$ and pertaining to the topmost valence and low-lying conduction bands. Note however that the remote bands cannot be neglected, as we will see below.
(b) Group theory analysis (not detailed here) indicates that the $k = 0$ states around the fundamental interband gap have the following symmetry characteristics (in absence of spin-orbit coupling, to be considered later): there is one low-lying conduction state with "S" orbital symmetry and three uppermost valence states with "P" orbital symmetry. The "practical" meaning of "S" and "P" orbital symmetries will be given later on.

In the following paragraph, we use these two results to discuss the $\vec{k} \cdot \vec{p}$ method, which is particularly versatile for the description of electronic states with energy around the fundamental band gap and $\vec{k}$ near the Γ point.

1.2.2 The $\vec{k} \cdot \vec{p}$ Method

As mentioned above, we are particularly interested in the description of the states pertaining to a small part of the crystal band structure, in both $\vec{k}$ and energy axes, namely, the states around $k = 0$ and related to the topmost valence and low-lying conduction bands. The $\vec{k} \cdot \vec{p}$ method is particularly versatile for the description of such electronic states. However, its starting point is actually much more general than focusing in the more interesting but rather small region in energy *versus* $\vec{k}$ space. Indeed, the structure of the effective eigenvalue problem (1.2) strongly suggests spanning the $\vec{k} \neq 0$ solutions on the basis of the k=0 solutions, assumed to be known:

$$u_{n,\vec{k}}(\vec{r}) = \sum_{n'} a_{n'}(\vec{k})\, u_{n',0}(\vec{r}) \tag{1.5}$$

where $H_{\rm cr}\, u_{n,0}(\vec{r}) = E_{n,0}\, u_{n,0}(\vec{r})$ with $u_{n,0}(\vec{r}) = u_{n,\vec{k}=0}(\vec{r})$ and $E_{n,0} = E_n(\vec{k}=0)$. This gives the matrix eigenvalue problem to solve:

$$\left\{\left[\, E_{n,0} - E_n(\vec{k}) + \hbar^2\,\vec{k}^2/(2m_0)\,\right]\delta_{n',n} + \hbar\,\vec{k}\cdot\vec{p}_{n',n}/m_0\right\}\; a_{n'}(\vec{k}) = 0 \qquad (1.6)$$

This method gives in principle exact crystal eigensolutions if one is able to provide a complete set of energy values ($E_{n,0}$) and matrix elements for the momentum operator ($\vec{p}_{n',n}$) associated with the $k = 0$ states. This is actually impossible to implement and a truncation of the $k = 0$ ensemble is unavoidable in actual calculations. Before considering such approximations, it is worth stressing at this point one important aspect of the $\vec{k}\cdot\vec{p}$ formalism. One could envision combining group theory analysis and experiments to obtain detailed information about the *parameters* $E_{n,0}$ and $\vec{p}_{',n}$: the theory allowing to ascertain which of the numerous matrix elements actually do not vanish, while an ideally complete set of experiments would provide the values of such matrix elements (at least their absolute values) as well as the energy edges $E_{n,0}$. In this sense, the $\vec{k}\cdot\vec{p}$ approach appears as a semi-phenomenological model, in which general considerations based on symmetry arguments are used to push as far as possible the theoretical description, whereas the final determination of ultimate parameters is ensured by experiments. This strategy allows circumventing the very first (and possibly the principal) difficulty underlying the modelling of the one-electron crystal eigenstates, namely, the lack of detailed knowledge of the actual crystal potential $V_{\rm cr}(\vec{r})$.

Different kinds of approximations follow from different truncation schemes and/or treatment of "remote" $k = 0$ edges (i.e. states with energy $E_{n,0}$ very far from the energy interval around the fundamental band gap we are interested in). The crudest approximation consists in retaining only the four states (the conduction "S" and the three valence "P" states) mentioned above in the basis: the simplified Kane model. In this case the 4×4 hamiltonian matrix (in the basis S, P_x, P_y, P_z) reads as:

$$\begin{array}{c|cccc} & |S\rangle & |P_x\rangle & |P_y\rangle & |P_z\rangle \\ |S\rangle & \lambda_S & \Pi k_x & \Pi k_y & \Pi k_z \\ |P_x\rangle & \Pi^* k_x & \lambda_P & 0 & 0 \\ |P_y\rangle & \Pi^* k_y & 0 & \lambda_P & 0 \\ |P_z\rangle & \Pi^* k_z & 0 & 0 & \lambda_P \end{array} \qquad (1.7)$$

where $\lambda_{S,P} = E_{S,P} + \hbar^2\vec{k}^2/(2m_0)$. $E_S = E_{c,0}$ and $E_P = E_{\rm v,0}$ are the edges of the conduction and valence band, respectively, which define the interband gap $E_G = E_S - E_P$. This matrix has been obtained by making explicit use of the aforementioned symmetry properties of the "S" and "P" orbitals, which in the present case states that only three interband matrix elements do not vanish and they are all equal: $\langle S|p_x|P_x\rangle = \langle S|p_y|P_y\rangle = \langle S|p_z|P_z\rangle = \Pi$, where capital P is used to label the three P-like orbitals, while small p denotes momentum operator. The four eigenenergies are readily obtained as:

$$\begin{array}{ll} (\lambda_P - e)^2 = 0 & \Rightarrow \ 2 \ \textit{solutions} \\ (\lambda_P - e)(\lambda_S - e) = |\Pi|^2 \, k^2 & \Rightarrow \ 2 \ \textit{solutions} \end{array} \tag{1.8}$$

Inserting the obtained solutions back in the matrix eigenproblem allows extracting the expansion coefficients and thus the searched periodic functions (and the full Bloch wavefunction) for any $\vec{k}$ value. Note that the model lies on two parameters: the interband matrix element of the momentum operator and the interband gap $E_G = E_S - E_P$. Their precise values cannot, of course, be inferred from the model itself, but can be extracted from measurements. To illustrate this point, let us consider the eigenstates for energies around the conduction band edge : $E \approx E_S$. To the lowest order in the inter-band coupling one has:

$$\begin{array}{l} E(\vec{k}) \approx E_S + \frac{\hbar^2 \vec{k}^2}{2 m_c^*}; \qquad \frac{1}{m_c^*} = \frac{1}{m_0} + \frac{|\Pi|^2}{E_G} \\ u_{c,\vec{k}}(\vec{r}) \approx S(\vec{r}) + \frac{\Pi^*}{E_G} \left\{ k_x P_x(\vec{r}) + k_y P_y(\vec{r}) + k_z P_z(\vec{r}) \right\} \end{array} \tag{1.9}$$

The dispersion is parabolic with effective mass $m_c^* < m_0$. As a consequence, information on $|\Pi|$ can be extracted from the electron effective mass (as measured e.g. in cyclotron resonance experiments), while E_G can be inferred from optical absorption experiments. It turns out that for various zinc-blend semiconductors $E_P \approx 23$ meV. For GaAs, $E_G = 1.5$ eV at low temperature and thus $m_c^*/m_0 \approx 0.07 \ll 1$.

As crude as it appears (we discuss below its main drawbacks), this model captures two essential features of the crystal electronic states: (i) the near edge dispersions are to a good approximation parabolic with effective masses governed by second order (interband) $\vec{k} \cdot \vec{p}$ couplings, and (ii) the Bloch functions for mobile electrons are admixtures of conduction and valence band solutions. If, on the one hand, the effect of $\vec{k} \cdot \vec{p}$ couplings on the effective masses cannot be disregarded ($m_c^*/m_0 \ll 1$), on the other hand the admixtures in the wavefunctions can in many circumstances be neglected, allowing to speak in terms of conduction and valence states as thought pure (i.e. non-admixed) in character: for instance, $u_{c,\vec{k}}(\vec{r}) \approx u_{c,0}(\vec{r}) = S(\vec{r})$.

The principal drawback of the simplified Kane model is the occurrence of two valence band dispersions with positive effective mass (and equal to the bare electron one). The corresponding eigenstates are pure valence-band states, thus unaffected by the interband couplings. One can explain this result by invoking a general quantum mechanical argument: the problem of one discrete state (the conduction one) coupled to the states of a degenerate ensemble (the valence ones) can always be reduced to a two levels problem. Indeed, it is always possible to properly hybridize (i.e. linearly combine) the degenerate states in such a way that amongst the states of the new (degenerated and orthonormalized) ensemble only one posseses a nonzero matrix element with the discrete state, all others remaining uncoupled. Actual band structure models account for three essential ingredients not present in the simplified Kane model: spin degeneracy, spin-orbit coupling and interband coupling to states ouside the $\{S, P_x, P_y, P_z\}$ subspace, keeping nonetheless the initial semi-phenomenological strategy of the $\vec{k} \cdot \vec{p}$ method. We illustrate this point with the Luttinger Hamiltonian that

describes the topmost valence states in diamond-like semiconductors (in this case the low-lying conduction band is treated as a remote band). First point: the consideration of spin enlarges the $k = 0$ basis to six states: $P_x\uparrow$; $P_x\downarrow$; $P_y\uparrow$; $P_y\downarrow$; $P_z\uparrow$; $P_z\downarrow\}$. Second point: the consideration of the spin-orbit coupling (of the same physical origin as isolated atoms) within this new basis splits the 6-fold degenerate state into a quadruplet (Q) and a doublet (D):

state			*energy shift*
$\|Q_1\rangle$	$\|+3/2\rangle$	$\frac{1}{\sqrt{2}}\|(P_x + iP_y)\uparrow\rangle$	$+\Delta/3$
$\|Q_2\rangle$	$\|-1/2\rangle$	$\frac{-1}{\sqrt{6}}\|(P_x - iP_y)\uparrow\rangle - \sqrt{\frac{2}{3}}\|P_z\downarrow\rangle$	$+\Delta/3$
$\|Q_3\rangle$	$\|+1/2\rangle$	$\frac{1}{\sqrt{6}}\|(P_x + iP_y)\downarrow\rangle - \sqrt{\frac{2}{3}}\|P_z\uparrow\rangle$	$+\Delta/3$
$\|Q_4\rangle$	$\|-3/2\rangle$	$\frac{-1}{\sqrt{2}}\|(P_x - iP_y)\downarrow\rangle$	$+\Delta/3$
$\|D_1\rangle$	$\|SO;+1/2\rangle$	$\frac{1}{\sqrt{3}}\|(P_x + iP_y)\downarrow\rangle + \frac{1}{\sqrt{3}}\|P_z\uparrow\rangle$	$-2\Delta/3$
$\|D_2\rangle$	$\|SO;-1/2\rangle$	$\frac{-1}{\sqrt{3}}\|(P_x + iP_y)\uparrow\rangle + \frac{1}{\sqrt{3}}\|P_z\downarrow\rangle$	$-2\Delta/3$

(1.10)

where Δ is the spin-orbit energy involving the $\{P_x, P_y, P_z\}$ states. The states of the lower energy doublet are called "split-off" (SO) states. The linear combinations of orbital and spin components displayed in the previous table are the same as obtained for the three P states of an isolated atom in the presence of spin-orbit coupling, whose orbital wavefunctions possess well-defined orbital angular momentum components along a given axis, whereas the total wavefunctions possess well-defined total angular momentum components along this same axis. In a crystal the $k = 0$ orbital states are said to be eigenstates of a pseudo-angular momentum operator, whereas the quadruplet and triplet states are eigenstates of a pseudo-total angular momentum operator. Third point: the inclusion of remote $k = 0$ states are necessary to obtain a fair description of the energy dispersions (i.e. dependence of the energies with k). Indeed, the $\vec{k} \cdot \vec{p}$ matrix elements amongst the restricted basis of P-like states vanish. Such couplings involve, as previously mentioned, a number of interband matrix elements and interband energy gaps. However, symmetry analysis shows that only a small number of parameters "effectively" govern the valence band dispersions. To illustrate this point, let us further restrict ourselves to the description of the sole uppermost states (i.e. associated with the quadruplet of $k = 0$ states). In this case the four dispersions are eigensolutions of the 4×4 Luttinger matrix

	$\|+3/2\rangle$	$\|-1/2\rangle$	$\|+1/2\rangle$	$\|-3/2\rangle$
$\|+3/2\rangle$	H_{hh}	c	b	0
$\|-1/2\rangle$	c^*	H_{lh}	0	$-b$
$\|+1/2\rangle$	b^*	0	H_{lh}	c
$\|-3/2\rangle$	0	$-b^*$	c^*	H_{hh}

(1.11)

where

$$
\begin{aligned}
H_{hh}(\vec{k}) &= \frac{-\hbar^2}{2m_0}\left[(\gamma_1 - 2\gamma_2)k_z^2 + (\gamma_1 + \gamma_2)(k_x^2 + k_y^2)\right] \\
H_{lh}(\vec{k}) &= \frac{-\hbar^2}{2m_0}\left[(\gamma_1 + 2\gamma_2)k_z^2 + (\gamma_1 - \gamma_2)(k_x^2 + k_y^2)\right] \\
c(\vec{k}) &= \frac{\sqrt{3}\,\hbar^2}{2m_0}\left[\gamma_2\,(k_x^2 - k_y^2) - 2i\,\gamma_3\,k_x k_y\right] \\
b(\vec{k}) &= \frac{-i\sqrt{3}\,\hbar^2}{m_0}\,\gamma_3\,(k_x - ik_y)k_z
\end{aligned}
\tag{1.12}
$$

are functions of only three adimensional parameters, $\gamma_{1,2,3}$: the Luttinger parameters. They incorporate all non-diagonal $\vec{k}\cdot\vec{p}$ couplings of the valence states with remote bands. Their values have been obtained for a series of materials. For instance, for GaAs there is: $\gamma_1 = 6.85$; $\gamma_2 = 2.1$; $\gamma_3 = 2.9$. The labels "hh" and "lh" correspond, respectively, to the "heavy" and the "light" dispersions: for propagation along the Oz axis ($k_{x,y} = 0$), the non-diagonal terms of the Luttinger matrix vanish and one readily obtains two dispersions, related respectively to the $\pm 3/2$ and to the $\pm 1/2$ states. Since the hh-related mass ($m_0/(\gamma_1 - 2\gamma_2)$) is larger than the lh-related one ($m_0/(\gamma_1 + \gamma_2)$), the $\pm 3/2$ states are usually called "heavy" states and the $\pm 1/2$ ones "light" states. Note however that for $k_{x,y} \neq 0$, the non-diagonal terms do not vanish and the corresponding valence state wavefunction contains both heavy and light components.

1.3 Envelope Function Model

Let us now introduce the envelope function method to describe electronic levels in perturbed semiconductors. Indeed, as previously mentioned, an overwhelming number of studies in semiconductor physics is related to the modification of the crystal properties due to the presence of some kind of perturbation: doping with shallow impurities provides carriers; a d.c. bias triggers a current; the optical characteristics are the response of the crystal to an external e.m. excitation. The envelope function method is a versatile approach for many cases of interest, namely whenever the perturbation strength is "weak" enough (as compared to typical band energy widths or separations) and "slowly varying" in space (as compared e.g. to the lattice period) [5]. However, our main objective is to show that the envelope function description can be used to describe the electronic states of heterostructures. This is actually not evident *per se*, since, as we shall see, the nanostructuration process introduces as a rule large (in energy) and abrupt (i.e. at the cell size scale) variation of the crystal potential.

For definiteness, let us assume a position-dependent perturbation to the crystal Hamiltonian $H = H_{\rm cr} + W(\vec{r})$. As a starting point, we take advantage of the fact that the Bloch states (eigenstates of $H_{\rm cr}$) form a complete basis ($\langle\Psi_{n,\vec{k}}|\Psi_{n',\vec{k}'}\rangle = \delta_{n,n'}\,\delta_{\vec{k},\vec{k}'}$) that can be used to diagonalise the perturbation term. We expand correspondingly the perturbed eigenfunctions within this basis:

$$|\Psi\rangle \;=\; \sum_{n,\vec{k}} a_n(\vec{k}) \left|\Psi_{n,\vec{k}}\right\rangle \tag{1.13}$$

The coefficients are solutions of the Hamiltonian matrix

$$\sum_{n',\vec{k}'} \left\{ \left[\; E_{n'}(\vec{k}') \;-\; E \; \right] \delta_{n,n'}\, \delta_{\vec{k},\vec{k}'} \;+\; W^{n,\vec{k}}_{n',\vec{k}'} \right\} \; a_{n'}(\vec{k}') \;=\; 0 \tag{1.14}$$

The sum runs over the whole set of Bloch states. Using the definition of Bloch function, the matrix elements of the perturbation reads as:

$$W^{n,\vec{k}}_{n',\vec{k}'} \;=\; \int d\vec{r}\; e^{i\,(\vec{k}'-\vec{k})\cdot\vec{r}}\; W(\vec{r})\; u^{*}_{n,\vec{k}}(\vec{r})\; u_{n',\vec{k}'}(\vec{r}) \tag{1.15}$$

The main difficulty for the evaluation of this integral (for a given W, and besides the fact that the u's are unknown functions) comes from the fact that it presents terms with completely different spatial behaviours: the product of the two u's is rapidly varying (actually this product is periodic on the crystal lattice), whereas the plane wave is slowly varying over a unit cell (for wavevectors near the Γ point, which is of particular interest here). However, this drawback can be turned into an advantage, and serve as the starting point for an important approximation, which is a fundamental aspect of the envelope function method. To illustrate this point, we consider a simpler, one-dimensional version of this same problem:

$$I = \int dx \quad F(x) \quad u(x) \tag{1.16}$$

where $F(x)$ ($u(x)$) is a slowly (periodic) function.

$$\begin{aligned} I &\equiv \sum_{p} \int_{p} dx\; F(x)\; u(x) \\ &\approx \sum_{p} F(x_p) \int_{p} dx\; u(x) = \left[\sum_{p} F(x_p) \right] \int_{0} dx\; u(x) \\ &\approx \left[\int dx\; F(x) \right] \frac{1}{a_0} \int_{0} dx\; u(x) \end{aligned} \tag{1.17}$$

In the first step, the integral over the whole space is exactly replaced by a sum over contributions coming from the different cells (labelled by the integer p and of spatial period a_0). In the second line we take profit of the slow variation of $F(x)$ over the p-th unit cell (centred around x_p). Since $u(x)$ is periodic, its integral is the same over any cell, which permits taking its value (evaluated for instance at the central cell) out from the summation. Then, in the last step we use the very notion of an integral

as a limiting of a summation to (approximately) transform the discrete sum into an integral over the whole space. This same procedure can be employed for the original three-dimensional problem:

$$W^{n,\vec{k}}_{n',\vec{k}'} \approx \left[\int d\vec{r}\, e^{i\,(\vec{k}'-\vec{k})\cdot\vec{r}}\, W(\vec{r}) \right] \int \frac{d\vec{r}}{\Omega_0}\, u^*_{n,\vec{k}}(\vec{r})\, u_{n',\vec{k}'}(\vec{r}) \tag{1.18}$$

Note that we have explicitly assumed that the perturbation potential varies slowly inside one unit cell. This assumption is quite good for a large class of perturbation potentials, for which additionally interband mixings ($W^{n,\vec{k}}_{n',\vec{k}'}$; $n' \neq n$) can be neglected (at least in a first order description of perturbed states). As examples, we quote: (i) weak static electric or magnetic field, which leads to intraband motion of electrons, and (ii) bound and scattering states in the presence of shallow impurities (e.g. donors). Low frequency (as compared with the interband one) electromagnetic fields can also be tackled by this approach (which can be generalised to account for time-dependent effects). On the contrary, an optical excitation near the interband gap does not, of course, belong to this class: however weak, an interband excitation effectively couples valence and conduction band states. Also, the effect of a perturbation on the valence band states deserves a particular treatment, because of the important mixing of the states at $k \neq 0$ (as e.g. the previously mentioned heavy-light mixings). For the sake of simplicity, let us here restrict ourselves to the simplest one-band description (a multi-band generalisation of the following developments can be found in the literature). In this case, the band index n is fixed.

The procedure leading to the last equation allows transforming the original integral into the product of two integrals. Its enormous advantage is that it allows decoupling the slowly and rapidly varying terms. Moreover, we can show that the second integral can, for many purposes, be set equal to unity:

$$\int_{\Omega_0} \frac{d\vec{r}}{\Omega_0} u^*_{n,\vec{k}}(\vec{r}) u_{n',\vec{k}'}(\vec{r}) \approx \delta_{n,n'} + O^{(2)}(\vec{k}-\vec{k}') \tag{1.19}$$

where the second term is a second order correction in the one-band model. Finally, one gets

$$W^{n,\vec{k}}_{n',\vec{k}'} \approx \delta_{n,n'}\, \tilde{W}(\vec{k}'-\vec{k}) \tag{1.20}$$

where the tilde refers to Fourier transform. We endup with the eigenvalue problem:

$$\sum_{\vec{k}'} \left\{ \left[E_n(\vec{k}') - E \right] \delta_{\vec{k},\vec{k}'} + \tilde{W}(\vec{k}'-\vec{k}) \right\} \alpha_n(\vec{k}') = 0 \tag{1.21}$$

Solving this matrix Hamiltonian is equivalent to solving the operator equation

$$\begin{aligned} &\left[E_n(\vec{k} \to -i\vec{\nabla}) + W(\vec{r}) \right] F(\vec{r}) = E F(\vec{r}) \\ &F(\vec{r}) = \sum_{\vec{k}} \alpha_n(\vec{k})\, e^{i\vec{k}\cdot\vec{r}} \end{aligned} \tag{1.22}$$

where $\vec{\nabla}$ is the gradient operator. This can be demonstrated by (i) formally expanding $E_n(\vec{k})$ in powers of $\vec{k}$ and using $(-i\vec{\nabla})^p\, e^{i\vec{k}\cdot\vec{r}} \to (\vec{k})^p\, e^{i\vec{k}\cdot\vec{r}}$, and (ii) multiplying on the left by $e^{-i\vec{k}'\cdot\vec{r}}$ and integrating over the whole space. In the simplest case of a parabolic dispersion there is:

$$\begin{aligned} &E_n(\vec{k}) = E_n(0) + \tfrac{\hbar^2\vec{k}^2}{2m_n^*} \\ &\left[-\tfrac{\hbar^2\vec{\nabla}^2}{2m_n^*} + W(\vec{r}) \right] F(\vec{r}) = [E - E_n(0)]\, F(\vec{r}) \end{aligned} \tag{1.23}$$

Finally, the steps leading to this last equation resume in the replacement

$$-\frac{\hbar^2\vec{\nabla}^2}{2m_0} + W(\vec{r}) + V_{\mathrm{cr}}(\vec{r}) \to -\frac{\hbar^2\vec{\nabla}^2}{2m_n^*} + W(\vec{r}) + E_n(0) \tag{1.24}$$

On the left-hand side of this expression, one has the original problem of one electron moving in the crystal with its bare mass and in the presence of the perturbing potential. In the envelope function model the crystal potential is no longer explicitly present but its effect is contained in two effective parameters, the effective mass m_n^* and the energy edge $E_n(0)$, both related to a given band. This constitutes the essence of the envelope function method, namely, incorporating the role of the crystal potential into a few (two in the simple one-band model) effective parameters that can be either separately calculated (via more accurate theories) or inferred from experiments. In this sense, it extends to a perturbed crystal the strategy of the $\vec{k}\cdot\vec{p}$ method developed to calculate the electronic states of a perfect one.

As previously mentioned, this model has been applied with success to many physical situations of practical interest (calculations of low field electron transport, shallow impurity levels, cyclotron resonance effects). In the last part of this chapter, we indicate how it can also be implemented to describe the electronic levels of nanostructured systems.

In the following, we describe the electronic levels of nanostructured systems within the envelope function approach. To this end, we shall return to the one-band envelope function Hamiltonian and consider an "unperturbed" crystal: $W(\vec{r}) = 0$. On very general grounds, a nanostructure is obtained by the juxtaposition of regions of different semiconductors. For simplicity, we shall focus on the case of multilayers, i.e. a nanostructure formed by juxtaposing layers of different semiconductors (we shall not enter into the details of the growth process). Figure 1.1 shows some examples of such structures. The simplest one is, of course, the simple heterojunction (uppermost scheme), where two semi-infinite layers are brought into contact through a common surface (the heterojunction interface). Before discussing the envelope

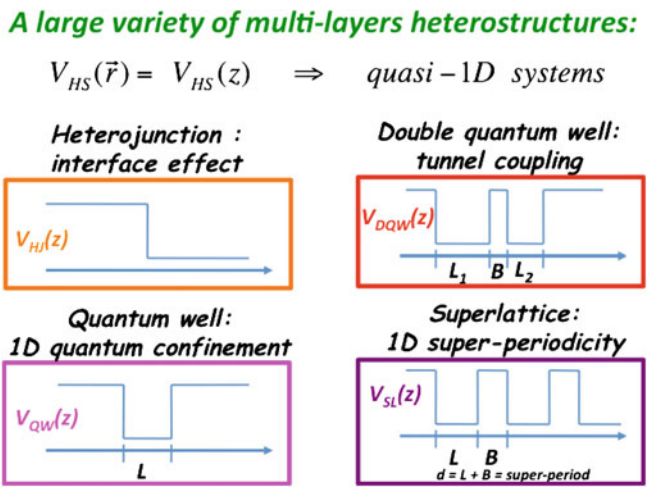


Fig. 1.1 Schematic representation of the band alignments related to the formation of four 1D heterostructures: a single heterojunction, a quantum well, a double quantum well and a superlattice

function formalism, it is worth pointing out two important aspects: (i) the two materials are often chosen "not so different", namely, they usually share the same crystalline structure and energy band sequence and (ii) the interface is assumed to be "ideal", namely, the materials on either sides of the interface conserve their bulk-like characteristics (crystal symmetry, interatomic bounds). Departures from either assumption do actually lead to new physics in the nanostructures; the consideration of such fine effects is nevertheless beyond the scope of these introductory notes.

The key point in the envelope function analysis is the fact that the energy edges $E_n(0)$ of different semiconductors have different values. Calculation of the alignment of such bands is a very complex problem. One has, nevertheless been able to infer from experiments the relative alignment of, e.g., the conduction band edges. As a practical example, let us consider the binary GaAs and the ternary $Al_xGa_{1-x}As$ (other systems will be discussed in the forthcoming chapters). The ternary has a larger fundamental bandgap than the binary and it is nowadays well established how the energy difference between the two conduction band edges (at the Γ point) varies with the Al content in the ternary. In a multilayer, the edge $E_n(0)$ related to the same band becomes a position-dependent function, $E_n(z)$. More generally, in a nanostructure the edge variation writes $E_n(\vec{r})$ and the envelope function Hamiltonian for the heterostructure becomes:

$$\left[-\frac{\hbar^2}{2}\vec{\nabla}\frac{1}{m_n^*}\vec{\nabla} + E_n(\vec{r})\right] F(\vec{r}) = EF(\vec{r}) \tag{1.25}$$

The kinetic energy term has been rewritten in order to comply with the requirement of current conservation through a given interface (another condition is the continuity of

the envelope wavefunction $F(\vec{r})$). Note that the difference in mass introduces much weaker effects, as compared to those due to the band edge offsets; correspondingly, the effective mass is very often taken to be position-independent.

The different structures in Fig. 1.1 correspond to model systems, which have been extensively studied in the framework of the envelope function method, with the help of the last equation. They allow envisioning many different fundamental physical effects, as the quantum confinement of carriers in a quantum well of nanometric width, the interwell (tunnel) coupling through a thin barrier in a double quantum well system, and the formation of a super-periodicity in superlattice structures. The physics of such structures, as well as of many others, will be discussed in detail in the later chapters. Additionally, more accurate methods will be presented and discussed, allowing a deeper understanding and finer description of the electronic states of electrons in nano-objects and their interactions. Also, other effects will be presented and discussed, as the response of different heterostructures to both external (applied electric bias and electromagnetic fields) and internal (coupling of electrons to lattice ions motions, strains) perturbations. In conclusion, nanostructuration represents nowadays a whole field of research in semiconductor heterostructures, leading to powerful concepts as band engineering, in particular in the field of optoelectronics and aiming at the development of new transport-based devices.

References

1. G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures* (Les Editions de Physique, Paris, 1988), p. 317
2. D. Grundman, M. Ledentsov, N.N. Bimberg, *Quantum Dot Heterostructures* (John Wiley & Sons, Chichester, 1998)
3. G. Bastard, J.A. Brum, R. Ferreira, Solid State Phys. **44**, 229–415 (1991)
4. C. Kittel, Introduction to Solid State Physics (John Wiley & Sons Inc, New York, 1996)
5. N.W. Ashcroft, N.D. Mermin, *Solid State Physics* (Holt, Rinehart and Winston, New York, 1976)
6. C. Kittel, *Quantum Theory of Solids* (Wiley, New York, 1987)
7. J.M. Luttinger, W. Kohn, Phys. Rev. **97**, 869 (1955)

Chapter 2
Theory and Modelling for the Nanoscale: The *spds** Tight Binding Approach

R. Benchamekh, M. Nestoklon, J.-M. Jancu and P. Voisin

Abstract The potential of the extended-basis tight binding method for quantitative modelling in nanosciences is discussed and illustrated with various examples. We insist on the method's ability to account for atomistic symmetries and to treat all the energy scales of electronic structures (from sub-meV quantities such as spin splittings to full-band properties like the optical index) using a single set of material parameters.

This chapter does not deal with the mathematics of the tight binding theory: there are excellent references—in particular, the celebrated text books by P. Yu and M. Cardona [1] and W. A. Harrison [2], and seminal papers [3–5] that the interested reader can use to dig into the technical aspects of the method. Here, we shall focus

R. Benchamekh · M. Nestoklon · P. Voisin (✉)
Laboratoire de Photonique et de Nanostructures,
CNRS, Route de Nozay,
91460 Marcoussis, France
e-mail: paul.voisin@lpn.cnrs.fr

R. Benchamekh
IPEST, route Sidi Bou Said,
2075 La Marsa, Tunisia

M. Nestoklon
Ioffe Institut RAS, Polytekhnicheskaya 26,
St Petersburg, Russia 194021

J.-M. Jancu
FOTON, Université Européenne de Bretagne,
INSA Rennes, 20 Avenue des Buttes de Coësmes,
35708 Rennes, France

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_2,

on conceptual aspects in connection with the importance of atomistic symmetries in the emerging field of nanosciences.

2.1 Introduction: A Snapshot View of Theoretical Methods for Nanosciences

As the solid-state community goes deeper and deeper into the exploration of the "nanoworld", a need for new modelling approaches taking into account bond-length scale variations of chemical composition and strain distribution becomes more and more apparent. Accuracy of the theoretical prediction is one issue, in particular when very large confinement energies come into play. But more importantly, atomistic symmetry breakings usually missed by "continuum" approaches contribute to qualitatively important features, in particular for discrete quantum systems such as impurity or quantum dot states. Another aspect of the modeling problem is the extraordinarily large number of objects populating the nanoworld, each having specificities that are as many modeling problems. Hence, flexibility of the theoretical methods is an essential issue. The need to take into account atomistic details of the nanostructure obviously points towards computational methods based on atomistic description of the electronic properties. On the other hand, a real risk exists that such methods provide accurate modeling results but do not explain them: confrontation of computational results with simple (eg effective mass) models remains essential to understanding. In this chapter, we shall discuss theoretical methods from the point of view of modeling experimental situations.

The landscape of theoretical methods can be separated into first-principle methods on one side, and methods using empirical parameters, on the other side. Among the latter, one can further distinguish atomistic and non-atomistic methods. The k.p (or envelope function) theory is a typical example of a non-atomistic method where the underlying crystal structure is represented through symmetries of bandedge Bloch functions that have empirically determined energies and momentum matrix elements between them [6–8]. Conversely, tight-binding and atomistic pseudopotential methods start with the atomic texture of the crystal potential or wave functions and keep them explicitly in the formal development. In the last 30 years, the k.p theory has had many successes and it remains by far the most popular method in semiconductor physics, due to its (relatively) simple formalism that can be mastered up to excellent levels by experimentalists. A distinctive feature is that the k.p theory is a toy model where complexity can be introduced progressively in the form of new basis states and new couplings in the Hamiltonian. Typically, one can start with the simple effective mass concept, then introduce non-parabolicity, spin-splittings, etc. Conversely, atomistic methods tend to operate in a "nothing or all" mode, complexity comes as a whole and cannot be decomposed in a perturbative spirit. In the following paragraph, we briefly introduce and compare the atomistic pseudopotential and the tight binding methods. The former was developed by A. Zunger and co-workers [9] since

the early 1990s. An analytical shape of potential around each atomic site is considered, depending on a few parameters that will be determined empirically. Then the crystal (or possibly, nanostructure) potential is obtained as the sum of the atomistic potentials and eigensolutions are expanded on a plane wave basis. The most difficult part of the game is to fit the coefficients of the atomistic pseudopotentials (APP) in order to reproduce as precisely as possible the band structure of parent materials (for instance, to reproduce binary bulk material band structure). Depending on which set of constraints is used, one may get rather different pseudopotentials. The major avantages of the technique are its ability to treat chemical discontinuities (another material is only another set of pseudopotentials) and strain, and the expansion in a naturally complete plane wave basis. However, nothing guarantees the transferability of parameters (say, the As pseudopotential in AlAs may differ considerably from the As pseudopotential in GaAs), and in its present implementation, the method suffers from parametric poverty that hampers precise full-band representation. Conversely, the tight binding method has its historical root in the early 1930s, when chemists and physicists were trying to formulate a quantum theory of the covalent bond. The leading idea was that electrons are "tightly" bound to individual atoms but can occasionally visit neighboring sites by tunnelling through the potential barrier. Hence the theory relies on "on-site" energies of various orbitals and "hopping" matrix elements between adjacent sites. These simple concepts were used by Slater and Koster [3] to formulate a description of crystal band structure where parameters of a tight binding model are considered as adjustable parameters whose values are empirically fixed in order to reproduce experimental features. Since in general, hopping between distant orbitals depends not only on distance but also on relative orientations, the empirical tight binding method has unavoidable parameter richness. From a practical point of view, this is both convenient (parametric flexibility) and inconvenient (multidimensional optimisation required for any parameter determination). Finally, it is important to note that the model uses on-site orbitals whose radial dependency is completely unknown: writing and diagonalising the Hamiltonian does not require any assumption on local wavefunctions, but their symmetry properties. There is a similar situation with the k.p theory that ignores the spatial dependencies of the zone centre Bloch functions.

2.2 The Empirical Tight Binding Formalism

While a number of physicists and chemists contributed to the emergence of the method, the present form of the Empirical Tight Binding (ETB) formalism was set in a seminal paper by Slater and Koster in 1954 [3]. The formalism is based on the (mathematically demonstrated) existence of a set of orthonormal orbitals (named the Löwdin orbitals) $\phi_m(\vec{\mathrm{r}} - \vec{r}_{jl})$ localised in the vicinity of the atomic sites, where r_{jl} denotes the position of the lth atom of the jth unit cell and m labels the different atomic-like orbitals. Translational invariance allows the introduction of Bloch sums:

$\Phi_{ml\vec{k}}(\vec{r}) = \frac{1}{\sqrt{N}} \sum_j \exp(i\vec{r}_{jl}.\vec{k})\, \phi_m(\vec{r} - \vec{r}_{jl})$, where N is the number of primitive cells. Bloch sums form a complete basis for the crystal eigenstates (Bloch functions) that can be expressed as $\Psi_{\vec{k}}(\vec{r}) = \sum_{m,l} C_{ml} \Phi_{ml\vec{k}}(\vec{r})$.

Injecting this expression into the Schrödinger equation $H\Psi_{\vec{k}} = E_{\vec{k}}\Psi_{\vec{k}}$, and using explicitly the orthogonality of Löwdin orbitals, one gets a set of linear equations for the C_{ml} coefficients: $\sum_{m',l'} \left(H_{ml,m'l'} - E_{\vec{k}}\delta_{m,m'}\delta_{l,l'}\right) C_{m'l'} = 0$, where the Hamiltonian matrix elements are:

$$H_{ml,m'l'} = \frac{1}{N} \sum\nolimits_{j,j'} \exp\ i\vec{k}.(\vec{r}_{jl} - \vec{r}_{j'l'})\ \left\langle \phi_m(\vec{r} - \vec{r}_{jl}) \right| H \left| \phi_{m'}(\vec{r} - \vec{r}_{j'l'}) \right\rangle \quad (2.1)$$

Since the spatial forms of the Löwdin orbitals (and the potentiel..) are unknown, these matrix elements cannot be calculated: in the empirical approach, they are treated as adjustable parameters and fitted in order to reproduce supposedly known features of the band structure (e.g. fit to experimental data and/or ab initio calculations). The number of matrix elements depends obviously on the number of orbitals per atom, and on the range of interactions. As shown by Jancu et al. [5], a model with s, p, d and s^* orbitals and limited to nearest neighbour interactions is actually "numerically complete" over a large energy range (15 eV) sufficient for a nearly perfect representation of bulk semiconductor band structure up to the two first conduction bands. An extremely important result of Ref. [5] is that free electron states (corresponding to Bloch functions of an "empty crystal") can be very well represented into this $spds^*$ basis. This implies that the method can be used to calculate features of surface physics.

Although the method itself and the rigorous classification of matrix elements was available since 1954, one had to wait until the mid 1970's for the first practical implementation of the method by Chadi and Cohen [10], due to the computational difficulty of the "inverse problem": while calculating the band structure for a given set of parameters is relatively easy, devising a strategy to get parameters fitting a known band structure is a very difficult task.

2.3 Band Structure of Bulk Materials: From sp^3 to $sp^3d^5s^*$

When restricted to the nearest neighbour interaction, the model using the simplest basis formed by the s and p orbitals of the anion and cation requires 9 parameters. This model is still widely used because it accounts qualitatively for the main features of covalent bonding. In particular, it gives a fair account of the valence band structure and band gap energy of direct gap semiconductors. However, the energies of L and X valleys cannot be fitted correctly. This motivated the work of P. Vogl et al. [4] who introduced a new state in the basis, corresponding to an empty, upperlying state of S symmetry, called the s^* state. The minimum model now involves 13 parameters, and the energies of L and X valleys can be reasonably well fitted. The transferability of on-site parameters is also significantly improved. However, the dispersion near

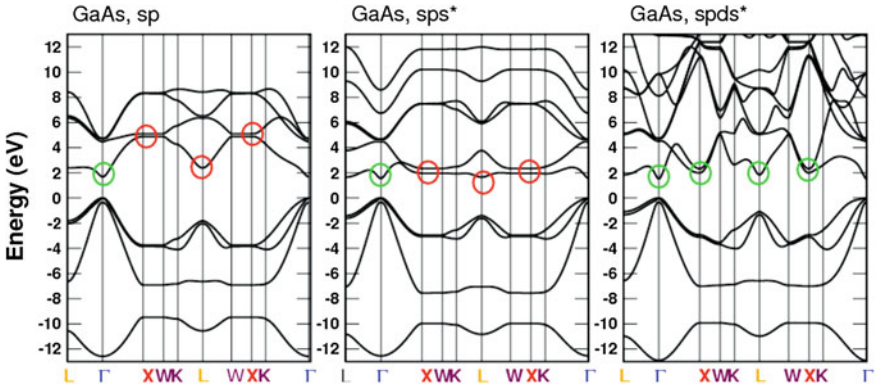


Fig. 2.1 Best fit of GaAs band structure in the sp^3, sp^3s^* and $sp^3d^5s^*$ nearest neighbor tight binding models. *Red* (*green*) circles indicate the regions of discrepancy (agreement) with experiment

the X point is still impossible to fit. The sp^3s^* model has been widely used in the 1980s to calculate III-V quantum well properties [11, 12], but the methodological difficulties finally led to a decline of the method's popularity in the early 1990s. In 1998, J.-M. Jancu et al. [5] finally extended the basis to the full shell of empty d states, and this extension of the basis allowed at last a nearly perfect modelling of semiconductor band structure, including Silicon, using essentially transferable on-site parameters. With the advent of the *spds** model, tight binding becomes a fully quantitative method, at the expense of a significant increase of the minimum number of parameters from 13 to 33. The best fit of GaAs band structure in the three models is shown in Fig. 2.1.

One should keep in mind that as long as the basis is not complete, interactions with other orbitals, either deep atomic states or upperlying free electron states are somehow included in the empirical approach as a "renormalisation" of the empirical parameters. For instance, in the sp^3s^* model, the matrix elements of the s^* orbital stand for the interactions with all the upperlying energy states, and as a consequence, the fitted on-state energy and two-centre integrals differ considerably from the values obtained for the s^* orbital in the $sp^3d^5s^*$ model. The latter are close to the free electron limit, which is physically sound. The neglected interactions with the deep atomic states (namely the d orbitals of the n = 3 shell for Ga and As) would manifest themselves mostly as a shift of the absolute energies of band extrema: this is introduced in the modeling of heterostructures as a "band offset" parameter. To the best of our knowledge, no attempt was made so far to predict band offsets in the tight binding theory by introducing explicitly the coupling to deep atomic states, but this should be technically feasible and in line with the current theory of band offsets [13, 14]. A most remarkable success of the *spds** model for bulk semiconductors is the perfect description of the band structure of Silicon [15], including the values of

valence band Luttinger parameters and effective masses of the X conduction valleys, for which extremely precise experimental data exist.

Tight binding is also suitable for calculations of optical matrix elements, either by introducing optical matrix elements between Löwdin orbitals as additional empirical parameters, or by using a k-space formulation [16] of the kinetic momentum operator as $\vec{p}(\vec{k}) = \frac{m_0}{\hbar}\nabla_{\vec{k}} H$. This approach has the merit of being gauge invariant, and gives excellent practical results without introducing any new parameter. Yet, it should be mentioned that there is still active theoretical discussions [17, 18] concerning the fact that this formulation misses intra atomic contributions to the optical properties.

2.4 Strain Effects: The Tight Binding Point of View

The effect of uniaxial stress on the band structure of semiconductors has been a major theoretical and experimental topic for many years. With the development of strained-layer epitaxy, it has also become an important issue in modern material science and device physics. Elastic deformations are ubiquitus in nano-objects. In their seminal approach, half a century ago, Bir and Pikus [19] established the strain Hamiltonian [19, 20] using the theory of invariants. It depends on a number of deformation potentials describing the shifts and splittings of the various band extrema. For instance, for a given band near the Brillouin zone centre, it reads as:

$$\begin{aligned} H_{\varepsilon}^{i} = &-a^{i}(\varepsilon_{xx}+\varepsilon_{yy}+\varepsilon_{zz}) - 3b^{i}[(L_z^2 - \frac{1}{3}L^2)\varepsilon_{zz} + cp] \\ &-\sqrt{3}d^{i}[(L_xL_y + L_yL_x)\varepsilon_{xy} + cp] \end{aligned}$$

where ε_{ij} are the components of the strain tensor $\boldsymbol{\varepsilon}$, $\boldsymbol{L}$ is the angular momentum operator and cp refers to circular permutations with respect to the axes x, y and z. a^i is the hydrostatic deformation potential describing the energy shift of band i, while b^i and d^i are the tetragonal and rombohedral (or trigonal) deformation potentials accounting for the eventual splitting of band i under the effect of corresponding uniaxial strain, respectively [001] or [111]. Other deformation potentials are involved at different high symmetry points of the Brillouin zone like X and L.

Within the tight binding formalism, strain effects are mainly determined by scaling the two-centre integrals (or transfer integrals) with respect to bond-length alterations, while bond-angle distortions are automatically incorporated via the phase factors in the Hamiltonian matrix elements (2.1). This leaves a more than sufficient number of strain-dependent parameters to fit the deformation potentials at the Brillouin zone centre. However, when trying to fit simultaneously the splittings of the zone-edge conduction valleys, one encounters difficulties [5, 21]. These have been overcome by considering that on-site energies of "quasi-free electron" states s^* and d must be shifted hydrostatically according to the change in unit cell volume, and, more importantly, d states split according to strain symmetry. With this approach,

absolutely general strain tensor can be handled, which is of utmost importance for many nanostructures, like self-assembled quantum dots or nanowires. Fitting deformation potentials clearly introduces a number of new parameters, but this is not a theoretical problem as long as the convergence of parameter-search routines remains good [21]. Again, it is important to check that the values of parameters coming out of blind-research procedures are compatible with the physical origin of these parameters.

2.5 Tight Binding as a Parameter Provider: Inversion Asymmetry and Parameters of the 14-Band k.p Model

Recent years have seen a strong interest in "semiconductor spintronics", following the paradigmatic idea of manipulating spin currents using electrostatic gates. A major issue in spin physics is the spin splitting of dispersion relations, which is due to the combined effects of spin–orbit interaction and inversion asymmetry. Spin splittings govern spin dynamics. They have historically been introduced in the k.p theory using the theory of invariants, with empirical material coefficients for the Bulk Inversion Asymmetry (or Dresselhaus term) and for the Strutural Inversion Asymmetry (or Rashba term). However, these terms are related to gaps and zone centre k.p matrix elements and can be deduced comprehensively from band structure calculations. Alternatively, spin splittings are "naturally" obtained in the tight binding calculation. The minimum framework to obtain spin-splitted dispersion relations in the k.p theory is a 14-band model including explicitly the anti-bonding p-type conduction band Γ_{8c}, and the momentum matrix element P' coupling Γ_{8c} and Γ_{6c}, which is allowed by inversion asymmetry. However, the original 14-band model introduced by Hermann and Weisbuch [22] missed the existence of another term allowed by inversion asymmetry, the off-diagonal spin-orbit , named Δ^-, coupling between the bonding (valence) and anti-bonding (conduction) p-type bands Γ_{8v} and Γ_{8c}, which was introduced by M. Cardona et al. [23] few years later. The model was further developed and applied to quantum wells by Pfeffer and Zawadski [24]. The scheme of band coupling is illustrated in Fig. 2.2a. It is interesting to point that this historical, progressive enrichment of the k.p theory has endorsed somewhat arbitrary values of the parameters, because it was constantly supposed that the main contribution to inversion asymmetry was the momentum matrix element P'. This topic was revisited by J.-M. Jancu et al. [25] who derived a new set of k.p parameters by fitting the 14-band k.p band structure to the spd*s** tight binding band structure.

The salient result of this re-examination is that new P' (resp. Δ^-) values are much smaller (resp. much larger) than old ones, to the extent that in new parameterization, the off-diagonal spin–orbit coupling Δ^- is by far the dominant contribution to the Dresselhaus constant γ_c, for all III–V semiconductors. It is also found that Δ^- is proportional to the difference between anion and cation spin–orbit constants, and hence vanishes for centro-symmetric semiconductors like Ge and Si. This new

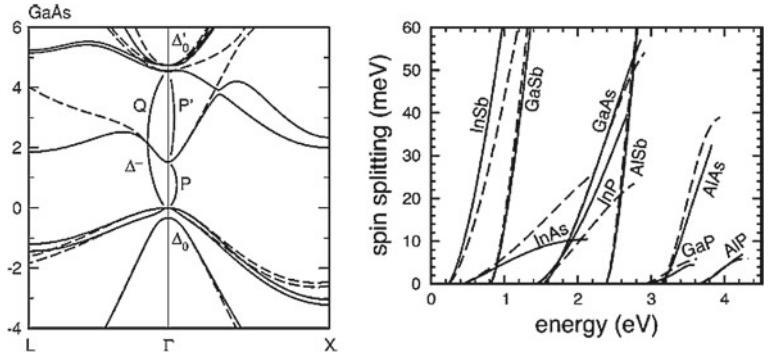


Fig. 2.2 *Left*: Band structure of GaAs as calculated with the 40-band TB model (*solid lines*) and the 14 x 14 $\mathbf{k} \cdot \mathbf{p}$ model using the new set of parameters. *Right*: Calculated conduction band spin splittings (*solid lines*: TB; *dashed lines*: k.p using new parameters) for main III–V semiconductors. From Ref. [25]

parametrization was later used by T. Nguyen-Quang [26] to calculate various properties of quantum well structures, like in-plane dispersion and spin splittings of valence bands or electric field-induced birefringence (Pockels effect) spectra, and yielded unprecedented agreement with experiments.

2.6 Quantum Confinement and Atomistic Symmetries: Interface Rotational Symmetry Breakdown

The merit of atomistic approaches is that lattice symmetries are automatically included in the Hamiltonian, which is not always the case for envelope function approaches. A remarkably simple example is the classical case of an interface between two semiconductors grown along the [001] axis. The presence of the interface obviously breaks the translational symmetry along the z axis, and the usual approach consists in writing the continuity relations for the "envelope" functions and their derivatives [6–8]. This can be done in the frame of a simple effective mass theory, or using a more elaborate multiband formalism. An immense majority of theoretical studies of quantum well structures was done along these lines, from the mid-1970s to the mid-1990s, assuming that an interface can be represented by a potential step Y(z) having full rotational symmetry. However, as illustrated in Fig. 2.3, the arrangement of chemical bonds at a (001) interface between a C_bA_b "barrier" material and a C_wA_w "well" material (C and A stand for cation and anion species) is such that all the $C_b - A_b$ bonds leaning backward in the barrier are in the 'horizontal' (-110) plane, and all the $A_b - C_w$ bonds leaning forward into the well are in the 'vertical' (110) plane. This evidences that the [110] and [-110] directions are not equivalent in the interface unit cell: in addition to breaking the

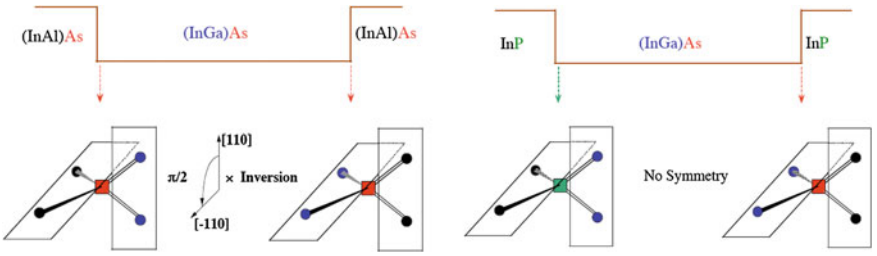


Fig. 2.3 Scheme of the atomic arrangement near successive interfaces in a Common Anion (*left*) or No Common Atom (*right*) quantum well. Here, we compare the technologically important (InGa)As / (AlIn)As and (InGa)As / InP systems. In, (InGa) and (AlIn) must be considered as three different effective cations

translational invariance, the presence of the interface breaks a rotational degree of freedom, namely the 4-fold roto-inversion symmetry 4^-. This implies that J_z (the z-component of the total angular momentum) is not a good quantum number. The point group symmetry of a single interface is C_{2v}. The main consequence is that heavy ($J_z = \pm 3/2$) and light ($J_z = \pm 1/2$) hole states are admixed by the interface potential. When considering a quantum well, one must combine the effects of two interfaces. It becomes necessary to distinguish between systems where the well and barrier materials share a common atom (in general the anion, like in GaAs/AlAs) and those where both the well and barrier anions and cations differ, or "No Common Atom" (NCA) systems, like InAs/GaSb or (GaIn)As/InP. In the former, one interface transforms into the other by the roto-inversion with respect to the well centre, and the resulting point group symmetry is D_{2d}, while the latter retains the C_{2v} point group symmetry of a single interface.

The important experimental consequence of C_{2v} symmetry is that the strength of optical transitions from valence to conduction band depends on (in-plane) polarisation of light. Krebs and Voisin [27], Ivchenko and co-workers [28] and B. Foreman [29] have independently shown how the envelope function theory (EFT) can be completed in order to include these symmetry considerations. Yet, valence band mixing by interfaces introduce specific material parameters that require atomistic information not provided by the theory. Conversely, tight binding calculations directly give quantitative account of these effects [30, 31], in good agreement with the measurement of absorption polarisation anisotropy in (GaIn)As/InP multi-quantum wells [30]. This is illustrated in Fig. 2.4. It is clear that the concept of interface rotational symmetry breaking is completely general and applies to any situation of chemical composition gradient, but an envelope function formulation based on the theory of invariants necessarily depends on the interface cristallographic orientation, that fixes the point group symmetry. Tight binding will model these effects in any situation, without introducing new parameters.

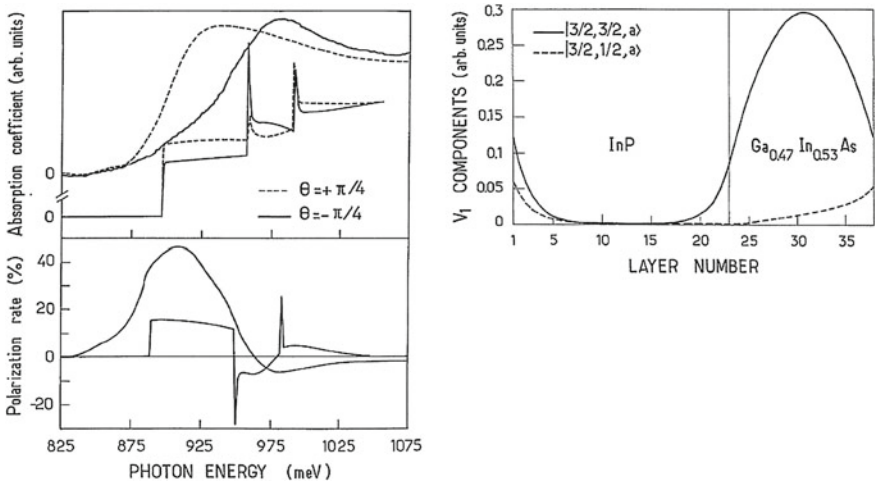


Fig. 2.4 Room-temperature absorption spectra of a 4.5/6.8 nm InGaAs/InP MQW for a photon polarisation along [110] ($\theta = \pi/4$) or along [-110] ($\theta = \pi/4$) (*left*, *upper panel*) and relative difference of these spectra, or 'polarisation spectrum' (*left*, *lower panel*). Calculated TB absorption curves are also shown. *Right panel* shows the projection of the valence band ground state on the anion $J_z = 3/2$ (heavy hole) and $J_z = 1/2$ (light hole) orbitals, evidencing the valence band mixing. From Ref. [30]

2.7 Quantum Confinement and Valley Mixing: X-valley and L-valley Quantum Wells

In some circumstances of large confinement, a direct gap semiconductor can transform into indirect due to the crossing of quantised levels at the zone centre with those associated with lateral valleys. This is in general due to the fact that effective masses at Γ are smaller than those at X or L, so confinement at Γ increases faster when well width decreases. The case of GaAs/AlAs where levels associated with the Z and X, Y valleys of AlAs become the ground state for narrow wells has been studied in great detail [32]. An anticrossing of Γ and Z valleys occurs due to coupling by the interface potential, and again EFT needs to introduce a specific parameter while TB naturally predicts the anticrossing. Here, we shall discuss the case of narrow GaSb/AlSb quantum wells grown along [001], where L-valley states of GaSb become the quantum well ground state for well widths smaller than 4.3 nm. The physical ingredients there are the confinement in an L valley, the coupling of pairs of L valleys projecting onto the same point of the two-dimensional Brillouin zone, e.g. (111) and (11-1), and the spin splitting of bulk L valleys for wavevectors perpendicular to the corresponding [111] axis. An EFT approach of this system was developped by Jancu et al. [33] and compared with TB modelling. In the basis spanned by the zeroth order wavefunctions $\Psi_{111}|\uparrow\rangle, \Psi_{111}|\downarrow\rangle, \Psi_{11-1}|\uparrow\rangle, \Psi_{11-1}|\downarrow\rangle$ the Hamiltonian reads as:

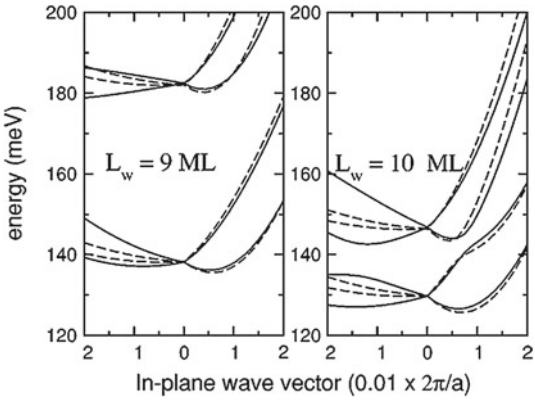


Fig. 2.5 Calculated in-plane dispersions for electrons in GaSb "L-valley QWs". In each panel, the *left part* refers to [1–11] direction and the *right part* to [110] direction. Solid lines show the tight-binding result and dashed lines the fitted 4 × 4 K.P model. From Ref. [33]

$$H = \begin{pmatrix} E_{2d} - A & C & V & iW \\ C* & E_{2d} + A & iW & V \\ V & -iW & E_{2d} + A & C \\ -iW & V & C* & E_{2d} - A \end{pmatrix}$$

where E_{2d} includes the quantized energy minimum and in-plane kinetic energy, A and C are related to the k-dependent matrix elements of the L-valley spin invariant $\alpha \mathrm{k} \cdot (\sigma \times \mathrm{n})$, and V and W are spin-conserving and spin-flip contributions of interface coupling. The calculation can be simplified using the infinite quantum well approximation, which leads to simple analytical forms for E_{2d} , A and C. The point that we wish to stress is that even this simplified approach contains three unknown material parameters (α, V and W). Again, the tight binding modeling directly gives the dispersion of electrons in these L-valley quantum wells without adjustable parameters. The fit of the EFT to the TB dispersion yields nice qualitative agreement, and gives the values of the missing parameters. It is also noteworthy that the EFT approach had to be developed for this specific problem of L valleys, while the tight binding code is exactly the same for any quantum well problematics Fig. 2.5.

2.8 Three-Dimensional Confinement: Symmetry Mistake in Current Theories of Impurity States

Another interesting example of atomistic symmetry importance can be found in the problem of substitutional impurities. The current theory of hydrogenic impurity states assumes that the electron (for the donor case) or hole (for the acceptor case) evolves under the effect of kinetic energy operator and the spherical electrostatic potential of an ionic charge. A quasi-Germanium model is implicitly used. This corresponds (for zinc-blende semiconductors) to solving a problem with O_h instead of T_d point

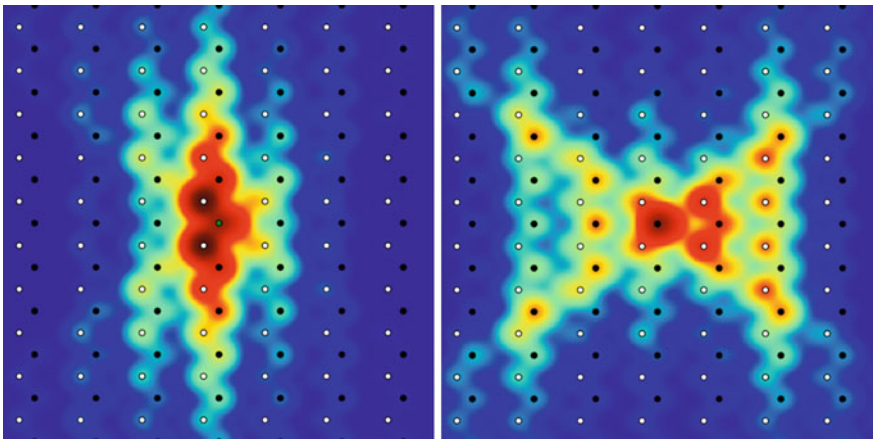


Fig. 2.6 (110) Plane cross-sections of a neutral acceptor state in Germanium in the impurity plane (*left*) and four atomic planes apart (*right*). The model includes Coulomb interaction and a δ-potential central-cell correction adjusted to produce a 100 meV binding energy. The positions of the Ge atoms in the section planes are shown by *black* and *white* circles. Atomic orbitals are represented by Gaussians with a 1.5 Å radius (M. Nestoklon et al. unpublished results)

group symmetry: indeed, nothing in the corresponding Hamiltonian accounts for the fact that the substitutional impurity potential is centred on an atomic site, and not at the inversion symmetry centre of the diamond structure. As pointed by Castner [34], correct Td symmetry can be restored by adding a tetrahedral (or octupolar) term in the Coulomb potential, that accounts for the atomic arrangement and corresponding non-spherical charge distribution near the impurity centre. Clearly, the octupolar correction is larger for deep impurity states, that are more sensitive to details of the "central cell" potential. Castner implemented this octupolar correction in the description of donors in silicon, for which valley degeneracy introduce another complexity. Here, we briefly discuss the case of acceptor states. There is abundant literature on the difficult problem of treating analytically Γ_8 degeneracy [35], but modern computing has allowed brute force solutions of the Luttinger kinetic operator in presence of a central force potential. Either in the case of very small or large spin–orbit splitting (resp. GaP and GaAs), it is found [36] that the cross-section of the local density of state (LDOS) of the 4-fold (resp. 6-fold) degenerate ground state in a (110) plane admits two planes of symmetry, the (001) and the (-110) planes. These symmetries are not compatible with T_d symmetry for which the reflection symmetry with respect to (001) plane does not exist. Conversely, as illustrated in Fig. 2.6, TB solutions of the same problem for a deep neutral acceptor state in Ge shows a large asymmetry with respect to (001) reflection. Clearly, the asymmetry of the impurity LDOS has nothing to do with crystal inversion asymmetry (which is absent in Ge), and simply reflects the tetrahedral environment of the impurity.

2.9 Alloys, beyond the Virtual Crystal Approximation: Dilute Nitrides

In the preceeding section, results of TB modelling of a three-dimensional object (actually, a supercell) are discussed. Calculations displayed in Fig. 2.6 correspond to diagonalisation of a 10,000 atom TB Hamiltonian, that is a $4 \cdot 10^5 \times 4 \cdot 10^5$ matrix. Much larger objects are actually accessible to computation, especially if only a few specific eigenstates are searched. The supercell frame can also be used to model realistic (random) alloys. This method was used to investigate the "giant bandgap bowing" of dilute nitride alloys. When a few percent of N atoms are substituted to As in GaAs, the bandgap *decreases* by a large amount (about 150 meV/%), instead of *increasing* by 198 meV/% as a linear interpolation between GaAs and GaN band gap would suggest. This intriguing experimental fact has raised considerable interest, due to the practical perspective of reaching Telecom wavelengths with materials epitaxied on a GaAs substrate. A simple heuristic model, the so-called band anticrossing model (BAC) [37] has accounted for the main observations, based on the consideration of an isoelectronic level associated with N, resonant with GaAs conduction band states. However, both the energy of the resonant level and the strength of its interaction with delocalised conduction states are free parameters of the BAC model. Jancu et al. (unpublished results) have applied TB to this problem, by distributing randomly the parent chemical species in a few thousand atom supercell, letting the atoms find their equilibrium positions through atomistic elasticity (using the Valence Force Field theory), and solving the resulting Hamiltonian. Figure 2.7 shows the successful comparison of TB calculations with experimental results.

Similar calculations were made for quaternary and quinary alloys InGaAsN and InGaAsSbN and yielded equally good agreement with experiments. These calculations also explained the unexpected trends observed when annealing the quaternary alloys. When annealing a GaInAsN alloy, only the first neighbors of an N atom (i.e. Ga and In) can rearrange, enriching the environment of N atoms with In as compared with the random distribution. Conversely, when annealing a GaAsSbN alloy, only the second neighbors of an N atom (i.e. As and Sb) can rearrange. Calculations show, in agreement with the astonishing experimental results, that annealing produces a larger increase of the bandgap in the second case!

For higher N-contents, the role of cluster states has to be considered. Lindsay and O'Reilly [38] successfully included in a TB model the interaction between the GaAs Γ-states and the full range of N-related levels present in the alloy.

While this example illustrates the predictive capability of the *spds** TB model, it is so far tantalising in the sense that TB gives the correct result but not the explanation of the result in terms of features of the parent hosts band structures and couplings to (or hybridisation with) substitutional species. However, we believe that it should be possible to dig this issue and end up with a general theory of alloys explaining why some systems (InAsSb, dilute nitrides) show giant bowings and some others (AlGaAs), almost none.

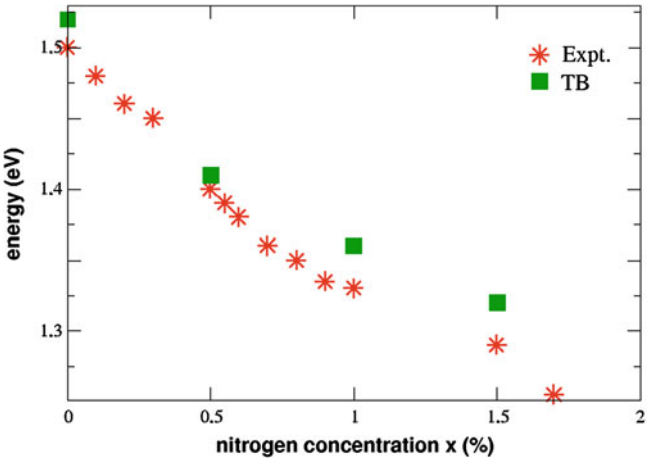


Fig. 2.7 Comparison of TB supercell calculations of random GaAsN alloys with experiment. From (J.-M. Jancu et al. unpublished results)

2.10 Full-Band Calculations: Dielectric Function and Piezo-Optical Constants

In this section, we shall focus on a completely different aspect of theoretical models, i.e. their ability to produce results valid throughout the Brilloin zone, and in a large energy range. k.p theory is a perturbative development valid in a narrow range of wave vectors in the vicinity of a high symmetry point, and is clearly disqualified when it comes to calculating full band properties such as a dielectric function. Current atomistic pseudopotentials lack parametric flexibility to reproduce full Brillouin zone: dielectric functions have been reported, but close examination reveals that a correct value of the optical index is obtained using very incorrect energy positions for the "E2" gap (R. Magri, unpublished results and private communications). Other methods are poorly compatible with heterostructures (nonatomistic pseudopotentials) or computationally very demanding (ab initio). Figure 2.8 compares the TB and experimental dielectric functions for GaAs. Although some significant discrepancy (which is currently attributed to excitonic effects) does exist near the so-called E2 gap near 5 eV, the overall agreement is excellent, and in particular, the value of the zero-frequency optical index (10.5) is very close to experimental value (10.9). Again, excitonic effects are not taken into account. From comparison with ab initio calculations, their influence on zero-frequency optical index is indeed expected to be in the 5% range. A similar agreement is found for all III–V semiconductors.

The effect of uniaxial strain on the optical index is also interesting because it tests the ability of a model to describe strain effect over the whole Brillouin zone. Under a uniaxial stress in the [001] direction, the bulk semiconductor becomes birefringent, with optical indices $n_x = n_y \neq n_z$. One defines the related piezo-optical constant

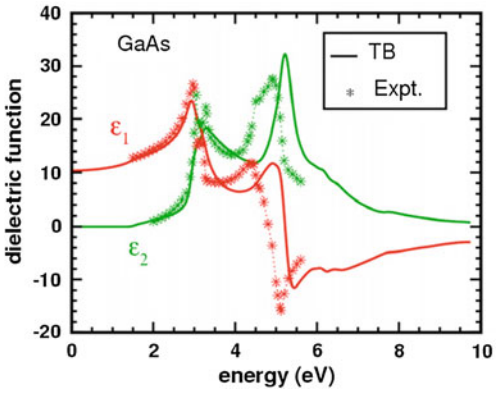


Fig. 2.8 Real and imaginary parts of the dielectric function. *Solid lines* show the tight binding result and *stars* are experimental data

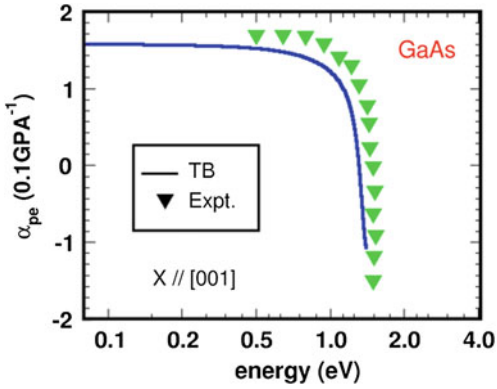


Fig. 2.9 Dispersion of the piezo-optic constant of GaAs under [001] stress

α_{001} as the relative difference $(n_x - n_z)/X$, where X is the applied stress [20]. The calculated dispersion of α_{001} in GaAs is compared with experimental results in Fig. 2.9. Equally satisfactory agreement is obtained for all semiconductors.

The successful calculation of spectral functions allow their extension to heterostructures. In principle, a superlattice grown along the [001] direction has the D_{2d} or C_{2v} symmetry which allows an anisotropic dielectric function with, respectively, $n_x = n_y \neq n_z$ or $n_x \neq n_y \neq n_z$. However, it is clear that breaking the tetrahedral symmetry around each atom will play a prominent role in the optical anisotropy, so uniaxial strain and interfaces are an essential ingredient. In order to investigate the feasibility of artificial semiconductors with significant optical anisotropy, Jancu et al. [39] have explored numerically the dielectric function of a number of configurations of ultrashort period superlattices (USPSL). The case of GaInAs/GaP or GaInAs/AlP is interesting because the free-standing USPSL is lattice-matched to a GaAs substrate, while the individual layers store considerable strain. GaInAs and GaP (or AlP) layers undergo, respectively, biaxial compression and biaxial tensile strain of about 3.6%. Calculated dielectric function for a GaInAs/GaP 3/3 USPSL is shown

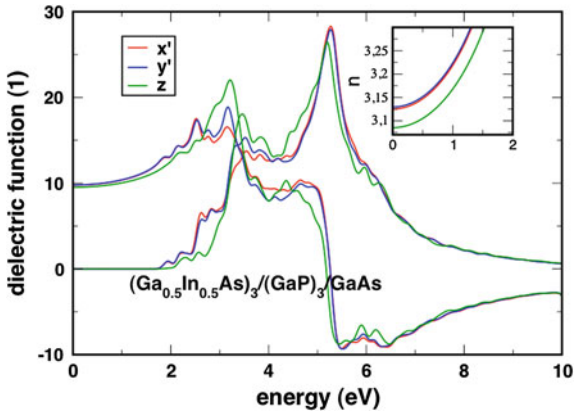


Fig. 2.10 Calculated dielectric function for a GaInAs/GaP 3/3 USPSL (from Ref. [39])

in Fig. 2.10. The observed optical anisotropy is in agreement with the C2v symmetry of this NCA material, but shows a rather complex spectral distribution. Strangely enough, an extremely simple zeroth order empirical rule could be inferred from the calculation of various cases: the birefringence of an USPSL is the atomic layer per atomic layer average of the piezo birefringences undergone by the various materials involved. As the piezo-optical constants of GaInAs and GaP are very different, one gets a situation where compressive and tensile strains compensate each other, but not the contributions to birefringence. Experimentally, the GaInAs/AlP (that gives similar theoretical spectra) could be grown by MBE and combined guided-wave optics and ellipsometric measurements revealed a material birefringence $n_x - n_z = 0.035$ of the same order of magnitude as the prediction, $n_x - n_z = 0.05$.

2.11 Surface Physics and Modeling of STM Images

Another domain where only few theoretical methods can be used is surface physics. There has been tremendous progress in this domain during the 1990s, thanks to the combination of ab initio calculations and scanning tunnelling microscopy (STM). In particular, the properties of the (110) natural cleavage surfaces of zinc blende semiconductors were thoroughly studied. Ab initio calculations have correctly predicted the elastic relaxation (the so-called buckling) of these surfaces, as well as the energy position and local density of states of surface electronic states. However, these methods cannot be used for the large supercells (>10,000 atoms) that are required to handle the situation of a sub-surface impurity state. The observation of single acceptor signature in STM images has raised enormous interest [40] because of the unpredicted shapes associated with resp. shallow and deep neutral acceptor states, respectively a triangle for GaAs: Be (binding energy 25 meV) or an asymmetric butterfly for GaAs: Mn or GaP: Cd (binding energy 115 meV). It was rapidly realised

that the local environment of the impurity should be important, and that tight binding should be a suitable method, but most contributors neglected surface physics effects and attempted to compare STM images with the cross-section of a bulk impurity state. Figure 2.10 (right panel) shows the experimental STM images measured for Mn acceptors respectively localised in the third, fourth and fifth sub-surface planes. Besides the striking 'butterfly' shape, one should notice that the atomic texture of the image shows only the rectangular lattice associated with the Ga surface atom, instead of the zig-zag chains of Ga and As atoms along the [1–11] direction that is present on a (110) surface. This atomic texture is also observed on naked (110) surface and explained in terms of the LDOS of specific surface states having a strong dangling bond character and, for this reason, extending much more into the vacuum than other crystal states. Hence, the schematic interpretation is that when the current flows from the STM tip to the semiconductor, one sees the lattice formed by empty dangling bonds on Ga atoms, while when it flows from the semiconductor to the tip, one should rather see doubly occupied dangling bonds on As atoms. Thus, atomic texture suggests that STM images are formed due to the hybridisation of surface and impurity states. In order to test this simple idea, Jancu et al. first calculated the situation of a 10,000 atom GaAs supercell containing a central Mn atom, including in the calculation Coulomb interaction and the hybridisation of Mn s, p and d-orbitals with neighboring As, but not the atomic exchange splitting among these d orbitals. A 4-fold degenerate neutral acceptor state with binding energy of about 100 meV comes out of the calculation. Cross-sections of the LDOS of this state in a (110) plane, respectively 3, 4 and 5 atomic planes above the impurity, are shown in Fig. 2.11 (left panel). While some similarities with the STM images can be argued, a striking discrepancy exists concerning the "orientation" of the butterfly asymmetry. Then, the electronic structure of the (110) surface was considered, taking into consideration relaxed atomic positions calculated by ab initio methods. This is possible only because free electron states can be reasonably well reproduced in the spd*s** model. Both the spectrum and local density of surface states of ab initio calculations were fairly well reproduced (without any extra parameter) by the TB calculation. For instance, the charge density of the "C3" conduction state at 2.2 eV, which has a prominent Ga dangling bond character, is identical in the two calculations. Finally, a supercell containing a free (110) surface and a Mn subsurface impurity at various depths was considered. A large splitting of the acceptor level due to surface strain (buckling) and a strong transfer of density probability from As to Ga atomic sites are the salient features of the results. The LDOS of the two-fold ground state in a (110) plane 2 Å above the surface, corresponding to simulated STM images, are shown in the central panel of Fig. 2.11. They agree rather well with the experimental images [41].

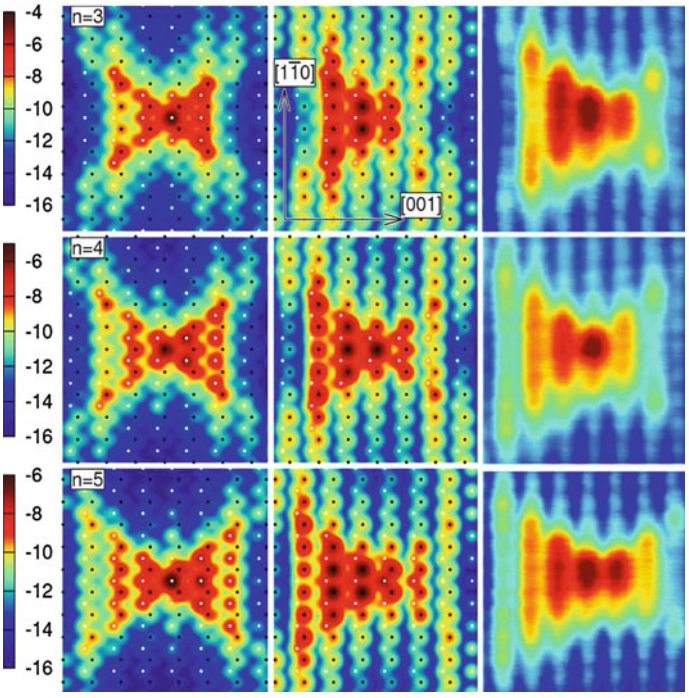


Fig. 2.11 Bulk impurity cross-section (BICS) (*left*), simulated STM images (SSTM) (*centre*), and experimental STM images (*right*) of an Mn neutral acceptor located n monolayers (n = 3 to 5) below the (110) surface. BICS is calculated in a (110) plane, n atomic planes away from the impurity, and SSTM 2 Å above the surface. SSTM LDOS is multiplied by 10^4 with respect to BICS. As (*white*) and Ga (*black*) positions on the surface are indicated. From Ref. [41]

2.12 Back to Theory: Local Wavefunction in the Tight Binding Approach

In the previous sections, we have evidenced that the *spds** tight binding approach is a powerful method to model single particle states in a variety of situations. In this final section, we come back to the more fundamental issue of calculating interactions between electronic states or quasi-particles. Compared to other methods, where single particle states are expanded in a complete basis of explicitly known functions (for instance, plane waves), tight binding suffers from the lack of information about the spatial dependencies of the Löwdin orbitals that form the basis but are never used explicitly in the formalism. This lack of knowledge on the local wavefunctions obviously hampers the calculation of interactions (in particular, short range interactions) between quasi particles. Benchamekh et al. (to be published), have recently attempted to solve this important theoretical issue. We start with local orbitals in the form of Slater orbitals that depend on adjustable "screening" parameters.

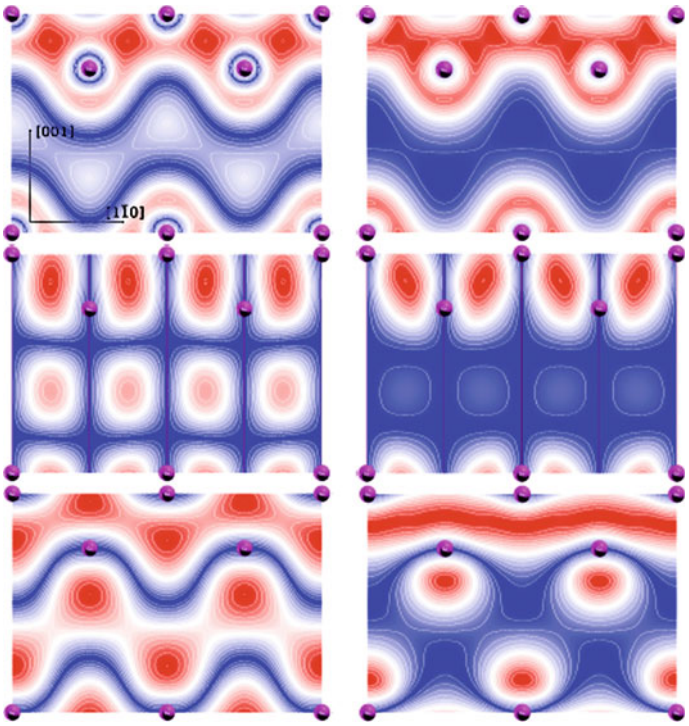


Fig. 2.12 Isodensity contours of the S, Y′ = X+Y and Z (from *top* to *bottom*) valence Bloch functions in the (1,1,0) plane, at the zone centre in bulk Ge. TB method as described in text (*left*) is compared with ABINIT calculations (*right*)

Once orthogonalised, these orbitals can be considered as trial functions for the Löwdin orbitals. They do reproduce the correct angular symmetry properties and expected long-distance behavior. Using this explicit basis, the momentum matrix elements between different bands at different points in the Brillouin zone can be calculated in real space and compared to those deduced from the derivation of Hamiltonian in momentum space. This procedure allows a fitting of the screening parameters, that is a self-consistent determination of approximate local wavefunctions. In Fig. 2.12, a representation of valence band states S_V, X+Y and Z at the zone centre obtained by this method for bulk Ge (neglecting spin-orbit interaction) is compared to equivalent ab initio calculations using the ABINIT package. The general agreement is very good, with minor discrepancies in the regions of low density. Similar agreement is obtained for conduction band states. While still in a preliminary stage, this approach opens a real perspective of combining the established potential of tight binding for single particle modelling with an ability to perform electron correlation calculations.

2.13 Conclusion

In this chapter, we have illustrated the potential of the extended basis *spds** tight binding model for quantitative modelling of III–V semiconductor structures. In fact, this model has also been applied successfully to many other materials, ranging from semiconductors to oxides and metals. A large range of nanostructures of major interest, for instance hybrid organic–inorganic nanostructures formed by semiconductor nanocrystals surrounded by organic ligands can be described with this method. Somehow, the *spds** TB model appears as a universal quantitative method for single particle states. Moreover, the recent attempt discussed in the preceeding section to derive self-consistently the local wavefunctions opens a route towards reconciling tight binding with many body physics. Yet, the method has some drawbacks that should also be stressed. First, it inherently requires a large number of empirical parameters, the determination of which is a most difficult task. Secondly, the computational difficulty is rather serious, and in absence of a well-tested open-source code, entering the tight binding community requires considerable effort. More fundamentally, as all computional methods, TB produces "numbers", but generally does not explain them in simple terms as the K.P theory usually does. Thus, the problem of qualitative understanding (which, in the authors' views, is the real intellectual challenge) tends to be decorrelated from that of obtaining quantitative results.

Acknowledgments This work was supported by "Triangle de la Physique" and CNRS-RAS international associate laboratory ILNACS

References

1. P.Y. Yu, M. Cardona, *Fundamentals of Semiconductors* (Springer, Berlin, 1996)
2. W.A. Harrison, *Electronic Structure and Properties of Solids* (Freeman, San Francisco, 1980)
3. J.C. Slater, G.F. Koster, Phys. Rev. **94**, 1498 (1954)
4. P. Vogl, H.P. Hjalmarson, J.D. Dow, J. Phys. Chem. Solids **44**, 365 (1983)
5. J.-M. Jancu, R. Scholtz, F. Beltram, F. Bassani, Phys. Rev. B **57**, 6493 (1998)
6. G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, (Les Editions de Physique, les Ulis, 1988)
7. E. L. Ivchenko, *Superlattices and Other Heterostructures: Symmetry and Optical Phenomena*, (Springer Series in Solid-state Sciences, 1997)
8. E. L. Ivchenko, *Optical Spectroscopy of Semiconductor Nanostructures*, (Alpha Science int, (2005)
9. K. Madër, A. Zunger, Phys. Rev. B **50**, 17393 (1994)
10. D.J. Chadi, M.L. Cohen, Physica Status Solidi B **68**, 405–419 (1975)
11. Yia-Chung Chang, J.N. Schulman, Phys. Rev. B **31**, 2069 (1985)
12. J.N. Schulman, Yia-Chung Chang, Phys. Rev. B **31**, 2056 (1985)
13. S.H. Wei, A. Zunger, Appl. Phys. Lett. **72**, 2011 (1998)
14. S.H. Wei, S.B. Zhang, A. Zunger, J. Appl. Phys. **87**, 1304 (2000)
15. F. Sacconi, J.-M. Jancu, M. Povolotskyi, A. Di Carlo, IEEE Trans. Electron Devices **54**, 3168 (2007)
16. T.B. Boykin, P. Vogl, Phys. Rev. B **65**, 35202 (2001)

17. B.A. Foreman, Phys. Rev. B **66**, 165212 (2002)
18. S.V. Goupalov, E.L. Ivchenko, Phys. Solid State **43**, 1867 (2001)
19. G.L. Bir, G.E. Pikus, *Symmetry and Strain-Induced Effects in Semiconductors* (Wiley, New York, 1974)
20. F.H. Pollak, M. Cardona, Phys. Rev. **172**, 816 (1968)
21. J.-M. Jancu, P. Voisin, Phys. Rev. B **76**, 115202 (1987)
22. C. Hermann, C. Weisbuch, Phys. Rev. B **15**, 823 (1977)
23. M. Cardona, N.E. Christensen, G. Fasol, Phys. Rev. B **38**, 1806 (1988)
24. P. Pfeffer, W. Zawadzki, Phys. Rev. B **53**, 12813 (1996) and references therein.
25. J.-M. Jancu, R. Scholz, E.A. Andrada e Silva, G.C. La Rocca, Phys. Rev. B **72**, 193201 (2005)
26. Q.T. Nguyen, Ph.D. Thesis, Ecole Polytechnique, (2006) http://tel.archives-ouvertes.fr
27. O. Krebs, P. Voisin, Phys. Rev. Lett. **77**, 1829 (1996)
28. E.L. Ivchenko, A.Y. Kaminski, U. Rossler, Phys. Rev. B **54**, 5852 (1996)
29. B.A. Foreman, Phys. Rev. Lett. **81**, 425 (1998)
30. O. Krebs, W. Seidel, J.P. André, D. Bertho, C. Jouanin, P. Voisin, Semiconduct. Sci. Technol. **12**, 938 (1997)
31. E.L. Ivchenko, M.O. Nestoklon, Phys. Rev. B **70**, 235332 (2004)
32. C. Gourdon, D. Martins, P. Lavallard, E.L. Ivchenko, Yun-Lin Zheng, R. Planel, Phys. Rev. B **62**, 16856 (2000), and references therein.
33. J.-M. Jancu, R. Scholz, G.C. La Rocca, E.A. de Andrada e Silva, P. Voisin, Phys. Rev. B **70**, 121306 (2004)
34. T.G. Castner Jr, Phys. Rev. B **79**, 195207 (2009)
35. A. Baldereschi, N.O. Lipari, Phys. Rev. B **8**, 2697 (1973)
36. C. Çelebi, P.M. Koenraad, A. Yu. Silov, W. Van Roy, A.M. Monakhov, J.-M. Tang, M.E. Flatté, Phys. Rev. B **77**, 075328 (2008)
37. W. Shan, W. Walukiewicz, J.W. Ager III, E.E. Haller, J.F. Geisz, D.J. Friedman, J.M. Olson, S.R. Kurtz, Phys. Rev. Lett. **82**, 1221 (1999)
38. A. Lindsay, E.P. O'Reilly, Phys. Rev. Lett. **93**, 196402 (2004)
39. J.M. Jancu, J.C. Harmand, G. Patriarche, A. Talneau, K. Meunier, F. Glas, P. Voisin, Comptes Rendus Physique **8**, 1174 (2007)
40. P.M. Koenraad, M.E. Flatté, Nature Mater. **10**, 91 (2011)
41. J.-M. Jancu, J.-C. Girard, M.O. Nestoklon, A. Lemaître, F. Glas, Z.Z. Wang, P. Voisin, Phys. Rev. Lett. **101**, 196801 (2008)

Chapter 3
Theory of Electronic Transport in Nanostructures

Eoin P. O'Reilly and Masoud Seifikar

Abstract As the first of three chapters on transport properties, we begin by explaining some of the key factors relevant to electron transport on a macroscopic scale. We then turn to address a range of novel nanoscale transport effects. These include the quantum Hall effect and quantised conductance, as well as the recent prediction and observation of quantised conduction associated with the spin quantum Hall effect in a topological insulator. We next consider graphene and the consequences of its unusual band structure before concluding with an overview of the potential use of "junctionless" transistors as one of the most promising approaches for future nanoscale electronic devices.

3.1 Introduction

There is considerable interest in transport at the nanoscale both from a fundamental perspective and also because of the requirements of current and future electronic devices. The invention of the transistor at Bell Labs in 1948 [1] enabled the development and widespread application of electronic devices. Since 1960, electronic devices have followed what is referred to as Moore's Law [2]: there has been an annual reduction of over 10% in the minimum feature size in electronic circuits, with the minimum feature size dropping from 10 μm in 1970 to 0.5 μm around 1990, and to 45 nm in 2010. As the feature size decreases it is no longer possible just to use macroscopic models to describe the electronic and transport properties—mesoscopic and

E. P. O'Reilly (✉) · M. Seifikar
Tyndall National Institute,
Lee Maltings, Dyke Parade, Cork, Ireland
e-mail: eoin.oreilly@tyndall.ie

E. P. O'Reilly · M. Seifikar
Department of Physics, University College Cork,
Cork, Ireland

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_3,

nanoscale effects and models need to be considered. So far the reduction in device feature size has been achieved through inventive solutions and detailed understanding, but scaling based on existing technology is now reaching its limits: if Moore's law continued to hold, then we would have subatomic scale devices by 2020! A range of new concepts are therefore being investigated to push towards and beyond the limits of Moore's Law. These include the use of new materials, such as graphene and carbon nanotubes, as well as the use of new device concepts based on existing materials, including the introduction of junctionless transistors [3].

As structure size scales down, quantum effects come into play. This can occur at very low temperatures in mesoscale devices, when thermal energies become comparable to quantum confinement effects, but can also be expected at higher temperatures in nanoscale devices, due to the larger quantum confinement effects in these structures. Many classical phenomena display measurable quantum character as the dimensions in which current can flow become restricted. The most widely known example is probably the quantum Hall effect for a 2-D electron gas, where carriers are confined in one dimension. In this case, the measured longitudinal resistance goes to zero at the same time as the Hall resistance becomes quantised in units of h/e^2. When carriers are further confined—in two dimensions—to form a quantum wire, the resistance of the wire itself can then become quantised, again in units of h/e^2.

Further effects can be observed by modifying the material band structure. It is reasonable to assume for most semiconductors that the electrons behave as particles with an effective mass, m^*, and then to treat carrier acceleration and transport as if the carriers were particles in free space with mass m^*. There are however at least two interesting classes of material where this assumption breaks down, namely graphene and so-called "topological insulators". In the case of graphene, which is a zero-gap semiconductor, the band dispersion varies linearly rather than quadratically with wavevector k, giving a band dispersion equivalent to that expected for photons or relativistic particles. As the name suggests, the band structure of a topological insulator is topologically different from that of a conventional semiconductor, with distinctly different gap states at the edges or surfaces of the material. This has surprising consequences, including the possibility to achieve a spin quantum Hall effect in zero magnetic field in suitably chosen samples.

3.1.1 Scope and Overview

This is the first of three chapters to address transport behaviour of semiconductor materials and heterostructures. In order to set up the framework for these three chapters, this chapter first presents an overview of some of the key factors relevant to electron transport on a macroscopic scale. This overview of macroscopic transport properties sets the scene for the more detailed consideration of transport in nanostructures in the remainder of this chapter, as well as providing relevant background for the discussion of hot electron transport in Chap. 4, and the introduction to Monte Carlo techniques in Chap. 5.

We begin in the next section by first providing an overview of macroscopic transport models. This begins with a definition of carrier effective mass and mobility, as well as a review of some of the key carrier scattering processes in semiconductor materials, and the use of Fermi's golden rule to calculate specific carrier scattering rates. We then introduce the detailed balance method as a high-level approach to estimating carrier mobility before going on to a more detailed model for carrier mobility based on solving the Boltzmann transport equation. We show that such models can account very well for the observed transport behaviour across a wide range of examples. We discuss briefly in Sect. 3.3 the need to go beyond Fermi's golden rule in cases where the scattering mechanisms strongly perturb the electronic properties, taking as example the strong scattering due to nitrogen-related resonant defect states in the dilute nitride alloy $GaAs_{1-x}N_x$. It has been proposed to include this alloy in GaAs-based multi-junction solar cells, to optimise the solar cell absorption efficiency. However, the intrinsically strong carrier scattering in the alloy has to date limited its usefulness in such applications.

Having established the key macroscopic models for carrier transport, we then turn to consider nanoscale transport effects. We start with the quantum Hall effect in Sect. 3.4, discussing the role both of quantisation and of edge states in a 2-D electron gas in an applied magnetic field. We then consider in Sect. 3.5 the spin quantum Hall effect, in which there has been considerable recent interest. In this case, the unusual band structure of a topological insulator can generate spin-polarised edge states, analogous to those observed in the quantum Hall effect, but leading, in this case, to the observation of quantised conductance associated with transport by identical spin carriers when no magnetic field is present. We continue in Sect. 3.6 with a more general discussion of the quantised conduction associated with current flow through quantum wires and quantum dots.

The overall interest in nanoscale transport is driven both by the novel fundamental phenomena which emerge at the nanoscale and also by the demands of future nanoscale devices, some of which may exploit these phenomena. There are a wide range of approaches being considered for future nanoscale devices. These include the development both of new materials with novel properties, as well as the investigation of novel device concepts. We overview in Sect. 3.7 graphene, one of the most interesting materials currently being investigated, and then turn in Sect. 3.8 to describe the junctionless transistor, as one of the more promising future device concepts currently being investigated, concluding with a brief summary in Sect. 3.9.

3.2 Macroscopic Transport Models

3.2.1 Carrier Effective Mass

From Bloch's theorem, we can associate a wavevector $\boldsymbol{k}$ with each energy state in a periodic solid, with the wavefunction $\psi_{n\boldsymbol{k}}$ of the nth state with wavevector $\boldsymbol{k}$ and energy $E_{n\boldsymbol{k}} = \hbar\omega_{n\boldsymbol{k}}$ being given by

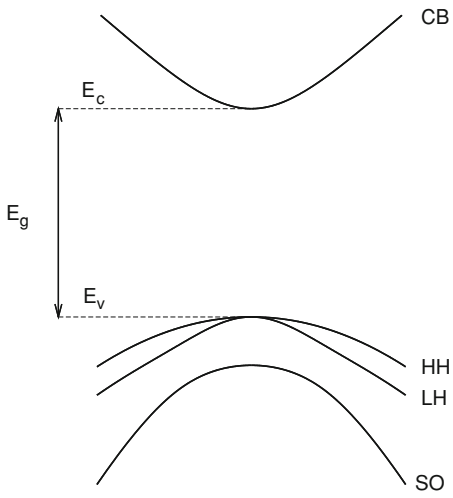


Fig. 3.1 Band structure of a direct gap semiconductor such as GaAs in the vicinity of the band gap and near the centre of the Brillouin zone. The lowest conduction state (at energy E_c) is separated by the band-gap energy E_g from the highest valence state, at energy E_v. The labels CB, HH, LH and SO indicate the conduction band, heavy-hole, light-hole and spin-split-off bands, respectively

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \exp(i\boldsymbol{k}\cdot\boldsymbol{r})u_{n\boldsymbol{k}}(\boldsymbol{r}) \tag{3.1}$$

where $u_{n\boldsymbol{k}}(\boldsymbol{r})$ is a function with the same periodicity as the lattice. Figure 3.1 shows as an example the band structure close to the energy gap between the filled valence and empty conduction band states of a direct gap semiconductor, such as GaAs. We have near the bottom of the conduction band that the energy $E_c(\boldsymbol{k})$ varies quadratically with wavevector $\boldsymbol{k}$, which we can write as

$$E_c(\boldsymbol{k}) = E_{c0} + \frac{\hbar^2 k^2}{2m_c^*} \tag{3.2}$$

where we use an effective mass, m_c^* to describe the conduction band dispersion. This concept of effective mass is very useful—we show below that we can treat an electron at the bottom of the conduction band as if it were a particle in free space with effective mass, m_c^*.

If we apply an external electric field $\boldsymbol{E}$, such that the electron experiences a force $\boldsymbol{F} = -e\boldsymbol{E}$ this implies the electron will move with an acceleration $\boldsymbol{a}$ given by $m_c^*\boldsymbol{a} = -e\boldsymbol{E}$, so that

$$\boldsymbol{a} = \boldsymbol{F}/m_c^* = -e\boldsymbol{E}/m_c^* \tag{3.3}$$

For small effective mass the electron then accelerates more rapidly in the solid than in free space. This is at first surprising, but reflects the fact that the electron is acted on not only by the external field $\boldsymbol{E}$ but also by the periodic field due to the lattice structure. If we were to take explicit account of both fields in discussing the dynamics of the electron it would exhibit its ordinary mass.

In order to derive the correct form for the electron effective mass, we consider an electron represented by a wave packet near the bottom of the conduction band, so

that the electron velocity is then given by the group or energy velocity, v_g, defined in terms of the variation of energy E with wavevector k as

$$v_g = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \frac{1}{\hbar}\frac{\mathrm{dE}}{\mathrm{d}k} \tag{3.4}$$

In the applied field $\boldsymbol{E}$, the electron experiences a force $\boldsymbol{F}$ such that its energy increases by

$$\begin{aligned} \delta \mathrm{E} &= F\delta x = F v_g \delta t \\ &= F\frac{1}{\hbar}\frac{\mathrm{dE}}{\mathrm{d}k}\delta t \end{aligned} \tag{3.5}$$

But we can also relate the energy change δE to the change in wavevector δk, as

$$\delta \mathrm{E} = \frac{\mathrm{dE}}{\mathrm{d}k}\delta k \tag{3.6}$$

By comparing the right-hand sides of (3.5) and (3.6) we find that

$$F = \hbar\frac{\mathrm{d}k}{\mathrm{d}t} \tag{3.7}$$

This holds irrespective of whether the electron is in free space or a periodic potential. We can use (3.4) to determine the electron acceleration a as

$$a = \frac{dv_g}{\mathrm{d}t} = \frac{1}{\hbar}\frac{\mathrm{d}^2\mathrm{E}}{\mathrm{d}k\mathrm{d}t} = \frac{1}{\hbar}\frac{\mathrm{d}^2\mathrm{E}}{\mathrm{d}k^2}\frac{\mathrm{d}k}{\mathrm{d}t} \tag{3.8}$$

Substituting (3.7) into (3.8) we find

$$a = \frac{1}{\hbar^2}\frac{\mathrm{d}^2\mathrm{E}}{\mathrm{d}k^2}F \tag{3.9}$$

By comparison with Newton's law (3.3), we see that the electron then behaves as if it has an effective mass, m^*_{eff} given by

$$\frac{1}{m^*_{\mathrm{eff}}} = \frac{1}{\hbar^2}\frac{\mathrm{d}^2\mathrm{E}}{\mathrm{d}k^2} \tag{3.10}$$

This broadens the definition of effective mass for a parabolic band in (3.2) to the more general case of a non-parabolic band dispersion.

3.2.2 Carrier Mobility

The current density J flowing in a bulk semiconductor with constant carrier density n due to an applied electric field E is given by

$$J = nev_d = \sigma E \tag{3.11}$$

where v_d is the average carrier (drift) velocity, e is the electron charge and σ is the conductivity of the semiconductor. The mobility μ describes how the drift velocity depends on applied electric field as:

$$v_d = \mu E \tag{3.12}$$

with the conductivity then depending on carrier density and mobility as

$$\sigma = ne\mu \tag{3.13}$$

When a distribution of electrons, initially in equilibrium, is subjected to an applied electric field, the electrons will start to accelerate, as described by (3.9). Their continued acceleration is however limited by a variety of scattering mechanisms, as discussed further below. If we assume that the electrons move with an average drift velocity, v_d, and assume an average carrier scattering relaxation time, τ_{rel}, then we can describe the evolution of the carrier distribution using the detailed balance method, with the rate of change of carrier momentum with time given by

$$\frac{\mathrm{d}(m^* v_d)}{\mathrm{d}t} = -eE - m^* v_d/\tau_{\text{rel}} \tag{3.14}$$

where (3.14) is a generalisation of (3.3) to take account of dissipation processes. Under steady-state conditions the left-hand side of (3.14) must equal zero, so that

$$eE = -m^* v_d/\tau_{\text{rel}}$$

or

$$\mu = \frac{e\tau_{\text{rel}}}{m^*} \tag{3.15}$$

For any scattering process, the scattering rate depends directly on the density of available final states—there can be no scattering if there are no final states available into which the electron can scatter. Hence, the relaxation time τ_{rel} is typically inversely proportional to the density of states. The density of states scales with dimension D as $(m^*)^{D/2}$, so that the relaxation time should scale as $(m^*)^{-1}$ in a 2-D electron gas, and as $(m^*)^{-3/2}$ in a bulk semiconductor. We see from (3.15) that the highest low-field mobility values may then be expected in materials with low carrier effective mass, with the expected low field mobility scaling with effective mass in a quantum well

structure as

$$\mu \propto (m^*)^{-2} \tag{3.16}$$

3.2.3 Carrier Scattering Mechanisms

We saw above that the electron acceleration is limited by carrier scattering, but did not specify what causes the carriers to be scattered. In practice, any perturbation that breaks the perfect periodicity of the crystal lattice can scatter electrons. Scattering mechanisms which can be important in typical semiconductor samples include:

- `Acoustic phonons`: these introduce long wavelength distortions of the lattice, where the displacement of neighbouring atoms is in phase. Because the phonon energy goes to zero as the phonon wavelength goes to infinity, acoustic phonon scattering can be of particular importance at low temperatures, where they provide the first excitations of the lattice;
- `Polar optic phonons`: in this case, neighbouring atoms vibrate out of phase. A polar optic phonon is a higher energy excitation than an acoustic phonon; polar optic phonon absorption or emission is therefore negligible at low temperature, but becomes an increasingly important inelastic scattering process as the temperature is increased;
- `Ionised impurities`: when an electron becomes unbound from a dopant atom and is free to move through the lattice, it leaves behind an ionised impurity centre, which acts as a Coulombic scattering centre;
- `Alloy fluctuations`: semiconductor alloys such as $In_xGa_{1-x}As$ or Si_xGe_{1-x} are used in many applications; the fluctuations due to the difference in potential associated with the different atom types in the alloy provide an additional scattering mechanism whose magnitude scales as $x(1-x)$;
- `Electron–electron`: due to the Coulomb repulsion between electrons as they propagate through the semiconductor;
- `Resonant defect levels`: within the conduction or valence band, provide an additional scattering path when the propagating electron energy is close to the resonant state energy;
- `Quantum well width fluctuations` lead to a position-dependent confinement energy, and tend to become increasingly important as the well width decreases.

For an indirect semiconductor such as unstrained bulk Si, there are six equivalent conduction band minima, each of which lies close to the X point along the six different Γ-X directions in the first Brillouin zone. The scattering in such an indirect gap material is generally considerably stronger than at the conduction band minimum in a direct gap semiconductor, both because of the larger density of states at the (X and L) points, and also because both intravalley and intervalley processes can be expected to contribute to the total scattering rate. Scattering between X valleys and between L valleys also become important at high electric fields in direct gap semiconductors:

as the electrons accelerate to higher energy in the Γ valley, they can acquire sufficient energy to become degenerate with and scatter into the L or X states. This will be discussed in more detail in Chap. 4.

3.2.4 Carrier Scattering Rates and Boltzmann Transport Equation

The rate of transition from an initial occupied state $|\psi_i\rangle$ to an empty final state $|\psi_f\rangle$ can be calculated from time-dependent perturbation theory, and is given by Fermi's golden rule as

$$R(i \rightarrow f) = \frac{2\pi}{\hbar} \left|\langle\psi_i|\Delta H|\psi_f\rangle\right|^2 D(E_f) \tag{3.17}$$

where ΔH in the matrix element is the perturbation to the Hamiltonian due to the scattering mechanism, $|\psi_i\rangle$ and $|\psi_f\rangle$ are both eigenstates of the unperturbed Hamiltonian H_0, the density of final states is $D(E_f)$ and $E_f = E_i$ for elastic scattering.

When we introduced the detailed balance approach in (3.14), we assumed that all carriers moved with the same average drift velocity and scattering time. In practice, the carriers are described by a distribution function $f(\boldsymbol{r}, \boldsymbol{p}, t)$, where f describes the probability of a state at position $\boldsymbol{r}$ and with momentum $\boldsymbol{p} = \hbar\boldsymbol{k}$ being occupied. In the absence of scattering, each particle follows the trajectory given by its group velocity $\boldsymbol{v}_g$ and the rate of change of momentum $\boldsymbol{p}$ is given by the external force $\boldsymbol{F}$ acting on the particle. (For external electric and magnetic, $\boldsymbol{B}$, fields acting on electrons, the force $\boldsymbol{F} = -e[\boldsymbol{E} + \boldsymbol{v}_g \times \boldsymbol{B}]$.) Thus, for short time-intervals Δt, those particles of momentum $\boldsymbol{p}$ which are at position $\boldsymbol{r}$ at time t were at position $\boldsymbol{r}–\boldsymbol{v}_g(\boldsymbol{p})\Delta t$ at time $t - \Delta t$ (for short Δt) and had momentum $\boldsymbol{p} - \boldsymbol{F}\Delta t$. Then, following the trajectories of all particles, we have that

$$\begin{aligned} f(\boldsymbol{r}, \boldsymbol{p}, t) &= f(\boldsymbol{r} - \boldsymbol{v}_g\Delta t, \boldsymbol{p} - \boldsymbol{F}\Delta t, t - \Delta t) \\ &= f(\boldsymbol{r}, \boldsymbol{p}, t - \Delta t) - [\boldsymbol{v}_g \cdot \nabla_{\boldsymbol{r}} f + \boldsymbol{F} \cdot \nabla_{\boldsymbol{p}} f]\Delta t \end{aligned} \tag{3.18}$$

where we have assumed that the distribution function f is smooth when deriving the second line of this equation. We then have, in the absence of scattering, that

$$\frac{\mathrm{d}f}{\mathrm{d}t} = -\boldsymbol{v}_g \cdot \nabla_{\boldsymbol{r}} f - \boldsymbol{F} \cdot \nabla_{\boldsymbol{p}} f \tag{3.19}$$

When we include scattering, we write

$$\frac{\mathrm{d}f}{\mathrm{d}t} = -\boldsymbol{v}_g \cdot \nabla_{\boldsymbol{r}} f - \boldsymbol{F} \cdot \nabla_{\boldsymbol{p}} f + \left[\frac{\mathrm{d}f}{\mathrm{d}t}\right]_{\text{scatt}} \tag{3.20}$$

where the last term on the right denotes the contribution of scattering, which can be written in detail as

$$\left[\frac{\mathrm{d}f}{\mathrm{d}t}\right]_{\text{scatt}} = \frac{1}{\hbar^3}\int \left[R(\boldsymbol{p}', \boldsymbol{p})f(\boldsymbol{p}')\{1 - f(\boldsymbol{p})\} - R(\boldsymbol{p}, \boldsymbol{p}')f(\boldsymbol{p})\left\{1 - f(\boldsymbol{p}')\right\}\right]\mathrm{d}^3\boldsymbol{p} \tag{3.21}$$

where $R(\boldsymbol{p}, \boldsymbol{p}')$ is the scattering rate of a particle of initial momentum $\boldsymbol{p}$ to final momentum $\boldsymbol{p}'$. The factors $[1 - f(\boldsymbol{p})]$ and $[1 - f(\boldsymbol{p}')]$ are there because of the Pauli exclusion principle, which prevents scattering into a state which is already occupied. Equation (3.21) emphasises that **scattering can only occur if there is an empty final state available**. This is usually the case in a bulk semiconductor, but we shall see below that some of the novel features observed in nanoscale transport arise precisely because scattering is suppressed due to there being no suitable final states available.

Equations (3.20) and (3.21) together give the full Boltzmann transport equation. The most difficult term to solve for is the scattering contribution of (3.21). It can be solved by various specialised numerical methods, including Monte Carlo simulation, as described in detail in Chap. 5. However, for many purposes in considering elastic alloy scattering, it is appropriate to use the relaxation time approximation, assuming that:

$$\left[\frac{\mathrm{d}f}{\mathrm{d}t}\right]_{\text{scatt}} = \frac{f(\boldsymbol{p}) - f_o(\boldsymbol{p})}{\tau(\boldsymbol{p})} \tag{3.22}$$

where $\tau(\boldsymbol{p})$ is the overall carrier relaxation time for carriers with momentum $\boldsymbol{p}$.

3.2.5 Mobility in Bulk Semiconductors and Heterostructures

Knowing the magnitude and temperature dependence of different scattering mechanisms, it is possible to provide a clear understanding of the overall temperature dependence of the mobility across a wide range of bulk semiconductors and semiconductor heterostructures. The data points in Fig. 3.2 show as an example the measured variation of mobility with germanium composition, x, in bulk silicon-germanium alloys, $Si_{1-x}Ge_x$ at room temperature. The upper dashed line shows the calculated mobility due to phonon scattering. It can be seen that this gives a good estimate of the mobility in Si and Ge, where it is the main scattering mechanism in low-doped samples, but significantly overestimates the mobility across a wide range of alloy compositions. The lower curve (solid line) shows the calculated mobility when alloy scattering is included. First-principles electronic structure methods were used to find the rates of intravalley and intervalley n-type carrier scattering due to alloy disorder in the alloys, with scattering parameters for all relevant Δ and L intravalley and intervalley alloy scattering processes being calculated [4]. It can be seen that the n-type carrier mobility, calculated from the scattering rate using the Boltzmann transport equation in the relaxation time approximation, is in excellent agreement with experiments across the full range of bulk, unstrained alloys.

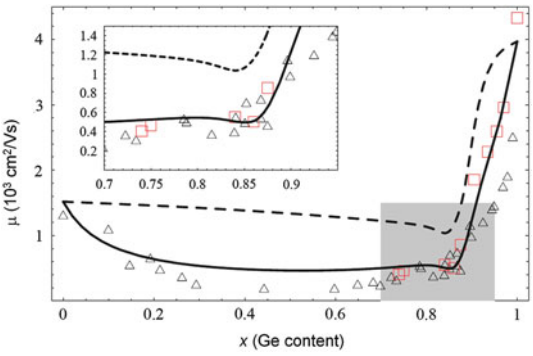


Fig. 3.2 Data points: measured variation of mobility with germanium composition, x, in bulk silicon-germanium alloys, $Si_{1-x}Ge_x$ at room temperature. *Dashed line* calculated scattering rate due to phonon scattering; *solid line* calculated mobility with alloy scattering included using first-principles calculation of scattering rates. The *inset* shows the mobility in the region near where the alloy X and L bands cross each other (following [4])

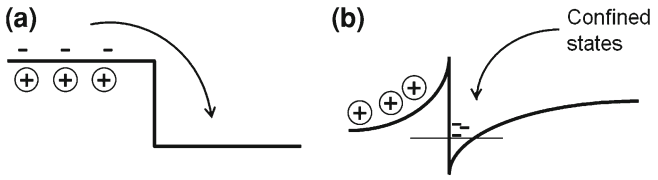


Fig. 3.3 **a** A modulation-doped heterojunction formed by doping a thin region of a wide-gap semiconductor close to the interface with a narrower gap material; **b** it is energetically favourable for the electrons to transfer into the narrower gap material, where they become confined at the heterojunction because of the electrostatic potential due to the positively charged ionised impurity sites

We saw from (3.13) that the conductivity, σ, in a bulk semiconductor, depends on the carrier density per unit volume, n, and the carrier mobility, μ, as $\sigma = ne\mu$. The obvious route to increasing conductivity, then is to increase the carrier density by increasing the doping density, N_d. However, this also increases the number of ionised impurity scattering centres ($=N_d$), thereby reducing the mobility, particularly at lower temperatures.

By contrast, the areal carrier density, n_s, can be increased in a low-dimensional system without significantly degrading the mobility. This can be achieved through modulation doping, where the dopant atoms are placed in a different layer from that in which conduction is occurring. This is illustrated in Fig. 3.3, where donor atoms are placed in the barrier layer adjacent to a layer with a lower conduction band edge energy. The excess donor electrons are transferred into the layer with lower band edge, leaving the ionised impurity centres in the barrier, typically over 10 nm from the conduction channel. At very low temperatures and in very pure materials, the electron mobility at GaAs/AlGaAs heterojunctions can exceed $10^7\,\mathrm{cm}^2/(\mathrm{Vs})$, four orders of magnitude larger than in low-doped bulk material, due to the virtual elimination of ionised impurity scattering. The effect is much less marked at room temperature, where other scattering mechanisms dominate, in particular scattering by

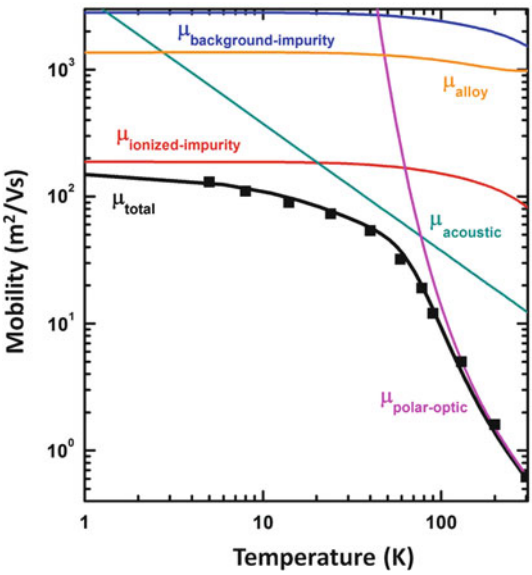


Fig. 3.4 Data points: experimentally measured variation of mobility with temperature in a GaAs/AlGaAs modulation-doped heterostructure, compared to calculated mobility, including temperature-dependent contributions from different scattering mechanisms. Experimental data from Ref. [5]; theoretical data courtesy of S. Birner, http://www.nextnano.de

polar-optic phonons. This is illustrated in Fig. 3.4, which shows as example the measured variation of mobility with temperature, as well as the calculated temperature dependence of the main scattering mechanisms. Although polar-optic phonon scattering becomes more important with increasing temperature, the room temperature mobility in modulation-doped heterojunction field effect transistors is nevertheless typically double that of the doped GaAs previously used in metal-gate field-effect transistors. This has two important consequences for the performance of high-speed transistors: first, the resistances are reduced, and with them the RC time constants, so that devices of a given size are faster and second, largely because of the reduced resistance, the levels of noise generated by the device (due to scattering processes) are also much reduced. The lowest noise transistors presently available are, therefore, based on modulation-doped heterojunctions, which find widespread application, for instance, in the amplifier circuits in satellite receivers and in mobile phones.

3.3 Scattering in Dilute Nitrides: Beyond Fermi's Golden Rule

We saw above how scattering in conventional semiconductor alloys can be well described using Fermi's golden rule to determine the carrier scattering rate. There are however a number of cases where a scattering centre introduces such a strong perturbation that it is necessary to go beyond Fermi's golden rule. This is the case for example when considering the strong scattering due to nitrogen-related resonant

defect states in the dilute nitride alloy $GaAs_{1-x}N_x$, a material which will be discussed further in Chap. 5. When a small fraction of arsenic atoms in GaAs is replaced by nitrogen the energy gap initially decreases rapidly, at about 0.1 eV per % of N for $x < \sim 0.03$ [6]. This behaviour is markedly different from conventional semiconductors, and is of interest both from a fundamental perspective and also because of its significant potential device applications. It has been proposed for instance to include GaInNAs lattice-matched to GaAs in multi-junction solar cells, to optimise the solar cell absorption efficiency. However, the intrinsically strong carrier scattering in the alloy has to date limited its usefulness in such applications.

The strong perturbation and large scattering cross-section due to an isolated N impurity in GaAs can be estimated using S-matrix theory (distorted Born wave approach). This was previously applied to successfully describe resonant scattering due to conventional impurities in GaAs [7, 8]. For a sufficiently localised perturbation, ΔV_N, the total scattering cross-section σ for an isolated impurity is given by

$$\sigma = 4\pi \left(\frac{m^*}{2\pi\hbar^2} \right)^2 |\langle \psi_{c1} | \Delta V_N | \psi_{c0} \rangle|^2 \Omega^2 \tag{3.23}$$

where m^* is the electron effective mass at the band edge and Ω is the volume of the region in which the wave functions are normalised. The state ψ_{c0} is the Γ-point conduction band Bloch wave function (in the absence of the N atom) and ψ_{c1} is the exact band-edge state in the presence of the N atom.

We note that the Born approximation is equivalent to setting $\psi_{c0} = \psi_{c1}$ in the required matrix elements. Although we saw above following (3.17) how this is perfectly adequate to describe conventional alloy and impurity scattering, it is entirely inadequate for the case of N defect scattering in GaAs.

Consider a perfect crystal for which the electron Hamiltonian is H_0 and the conduction band edge state has wave function ψ_{c0} and energy E_{c0}. When we introduce a single N atom into a large volume Ω of the otherwise perfect lattice, the new Hamiltonian, $H_1 = H_0 + \Delta V_N$, leads to a modified band edge state ψ_{c1} with energy E_{c1}. We can therefore rewrite the scattering matrix element of (3.23) as

$$\langle \psi_{c1} | \Delta V_N | \psi_{c0} \rangle = \langle \psi_{c1} | H_1 - H_0 | \psi_{c0} \rangle = (E_{c1} - E_{c0}) \langle \psi_{c1} | \psi_{c0} \rangle \tag{3.24}$$

Because $\langle \psi_{c1} | \psi_{c0} \rangle \rightarrow 1$ for sufficiently large Ω, it can be shown that at low impurity concentrations

$$\Omega \langle \psi_{c1} | \Delta V_N | \psi_{c0} \rangle = \frac{\mathrm{d}E_c}{\mathrm{d}n} \tag{3.25}$$

where E_c is the conduction band edge energy and n is the number of impurities per unit volume. Substituting (3.25) into (3.24), and noting that n is related to the concentration x by $n = 4x/a_0^3$, where a_0 is the GaAs unit cell dimension, the scattering cross-section for an isolated impurity is then given by

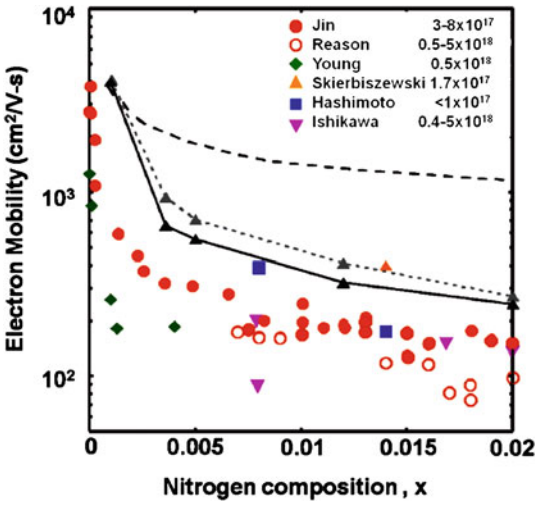


Fig. 3.5 Data points: measured variation of electron mobility with N composition x in $GaAs_{1-x}N_x$ (from [9]). The *uppermost dashed line* shows the calculated mobility, assuming scattering by isolated N atoms only (3.26) while the *lower lines* show the calculated mobility assuming a distribution of N states [10] and carrier density of 1×10^{17} cm^{-3} and 1×10^{18} cm^{-3} (*grey and black triangles*, respectively)

$$\sigma = \frac{\pi}{4}\left(\frac{m^*}{2\pi\hbar^2}\right)^2 \left[\frac{dE_c}{dx}\right]^2 a_0^6. \tag{3.26}$$

This result is key: it establishes a fundamental connection between the composition-dependence of the conduction band edge energy and the n-type carrier scattering cross-section in the ultra-dilute limit for semiconductor alloys, imposing general limits on the carrier mobility in such alloys.

We can see this by extending the isolated N result of (3.26) to the case of a dilute nitride alloy, $GaAs_{1-x}N_x$. The mean free path l of carriers depends in an independent scattering model on the scattering cross-section σ for a single defect and the number of defects n per unit volume as $l^{-1} = n\sigma$. The mobility μ is then related to the mean free path l as $\mu = e\tau/m^*$, with the scattering time $\tau = l/\bar{u}$, where $\bar{u}$ is the mean electron velocity.

The dashed line in Fig. 3.5 shows the estimated variation of the room temperature electron mobility with x in $GaAs_{1-x}N_x$, calculated using the two-level band-anticrossing model for $GaAs_{1-x}N_x$ [11, 12], which assumes all N resonant defect states to be at the same energy E_N. The electron mobility is estimated to be of the order of $1{,}000\,cm^2/(Vs)$ when $x = 1\%$, of similar magnitude to the highest values observed to date in dilute nitride alloys [13] but larger than that found in many samples, where $\mu \sim$ 100–300 $cm^2/(Vs)$, as shown by the data points in Fig. 3.5 (following [14, 9]). In practice, there is a wide distribution of N resonant state energies in $GaAs_{1-x}N_x$, associated with N–N nearest neighbour pairs and clusters [15, 16], with a significant number of these defect levels calculated to be close to the conduction band edge. The lower lines in Fig. 3.5 show the calculated mobility when scattering associated with this distribution of defect levels is included. It can be seen that inclusion of this distribution can largely account for the low measured electron mobility in this alloy system.

The intrinsically low electron mobilities in dilute nitride alloys have significant consequences for potential device applications. The low electron mobility, combined with the short non-radiative lifetimes observed to date, limit the electron diffusion lengths and efficiency achievable in GaInNAs-based solar cells. Further efforts may lead to increased non-radiative lifetimes, but are unlikely to see significant further improvements in the alloy-scattering-limited mobility [13].

3.4 Quantum Hall effect

The Hall effect provides a well-established technique to determine the mobility and carrier density per unit area in bulk semiconductor samples [17]. It was, therefore, an obvious technique to apply to low-dimensional semiconductor nanostructures. However, when such measurements were carried out at low temperatures on a 2-D electron gas, the results were completely unexpected [18]. The measured Hall resistance was quantised in units of h/e^2, where h is Planck's constant and e is the electron charge. As a consequence, a basic semiconductor experiment has become the standard for defining resistance and, more interestingly, has opened a wide field of fundamental research, some of which we discuss further in the following sections.

We consider first the classical Hall effect in a 2-D sample, with the current, I, given by

$$I = w n_s e v \tag{3.27}$$

where n_s is the areal carrier density, v the average carrier velocity and w the width of the sample. When a magnetic field, B, is applied perpendicular to the sample, it causes a force on each carrier, $F = e(v \times B)$, whose magnitude is then evB, directed towards the side of the sample. This leads to a build-up of charge on the two sides of the sample, until the induced electric field, E_H, exactly balances the magnetic force, $eE_H = evB$, with a measurable Hall voltage, V_H, across the sample then given by

$$V_H = E_H w = vBw \tag{3.28}$$

Combining (3.27) and (3.28), we can use the Hall voltage V_H to determine the areal carrier density n_s as

$$V_H = \frac{B}{n_s e} I \tag{3.29}$$

with the Hall resistance, R_H, defined as

$$R_H = \frac{V_H}{I} = \frac{B}{n_s e} \tag{3.30}$$

The Hall effect is widely used to measure the carrier density, n_s, and also the carrier mobility, μ, which can be determined knowing the current, I, carrier density and applied longitudinal voltage, V.

How then does the Hall effect become quantised in two dimensions? Two key factors need to be included in a description of the quantum Hall effect, namely, carrier quantisation and the existence of edge states in a finite sample.

We start by considering carrier quantisation. Close to the band edge, the energy levels in the ground state subband of a 2-D electron gas (2DEG) satisfy the relation

$$E(k_x, k_y) = E_0 + \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2) \tag{3.31}$$

where E_0 is the ground state zone centre confinement energy, and the electrons are free to move in the x–y plane.

When a strong magnetic field, B, is applied perpendicular to the 2DEG, the electron motion becomes quantised in cyclotron orbits in the 2-D plane. It can be shown that classically the cyclotron frequency, ω_c, depends directly on the applied field B as

$$\omega_c = eB/m^* \tag{3.32}$$

When quantisation effects are taken into account, the allowed orbital energies depend directly on ω_c as $E_n = (n + 1/2)\hbar\omega_c$, where n is an integer and the quantised energy levels are referred to as Landau levels. The energy levels of the 2DEG are then given by

$$E_n = E_0 + (n + 1/2)\hbar\omega_c + g\mu_B \boldsymbol{B} \cdot \boldsymbol{s} \tag{3.33}$$

where the last term describes the interaction between the electron spin $\boldsymbol{s}$ and the applied magnetic field $\boldsymbol{B}$.

The form of the density of states then changes in an applied magnetic field from a constant density of states to a series of discrete allowed energy levels, as illustrated in Fig. 3.6a. The total number of electron states is, however, conserved per unit energy range. The total number of states, N, per unit area between energy E and $E + \mathrm{d}E$ is given for the band dispersion of (3.31) by $N = g_{2\mathrm{D}}(E)\mathrm{d}E = \left(4\pi m^*/h^2\right)\mathrm{d}E$, where $g_{2\mathrm{D}}(E)$ is the 2-D density of states per unit area. All the states lying within an energy range $\mathrm{d}E = \hbar\omega_c$ are gathered into each pair of spin up and spin down Landau levels. The number of states, N, in each individual Landau level is then given by

$$N = \frac{1}{2}\left(\frac{4\pi m^*}{h^2}\right)\hbar\omega_c = \frac{eB}{h} \tag{3.34}$$

When j Landau levels are fully occupied, the areal carrier density $n_s = Nj$, and the Hall resistance are given by

$$R_H = \frac{B}{n_s e} = \frac{h}{je^2} \tag{3.35}$$

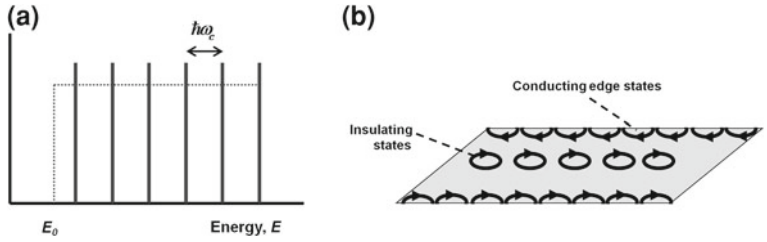


Fig. 3.6 **a** A magnetic field B applied perpendicular to a 2-D structure changes the density of states from a continuous spectrum (*dotted line*) to a series of discrete allowed energy levels (*black lines*), due to quantisation associated with the circular motion of the electrons in the applied magnetic field. For simplicity, the electron spin energy, $g\mu_B \boldsymbol{B} \cdot \boldsymbol{s}$ is ignored in this spectrum; **b** illustration of closed, insulating cyclotron orbits in the body of a 2-D structure, and of conducting states being repeatedly reflected and propagating along the edges of the sample

Because there is no current flow through a filled energy band, it might be expected, when all the Landau levels are filled, that there should be no longitudinal current flow through the Hall sample. This however does not take account of the finite size of the samples, with the carriers being confined overall in a region of width w (see 3.28). An electron near the centre of the sample will have a closed cyclotron orbit, as illustrated in Fig. 3.6b. Such an electron does **not** contribute to the current flow. However, let us now consider an electron at the edge of the layer, and assume for simplicity that it experiences an infinite confining potential to remain within the layer. Such an electron cannot complete a cyclotron orbit, but instead is repeatedly reflected off the wall, giving a conducting edge state. At sufficiently low temperature, there are no final states into which an edge state can be scattered: all edge states on one side of the sample propagate in the same direction, so back scattering would require the electron to be scattered to the opposite side of the sample. In addition, when $\hbar\omega_c \gg k_B T$ (i.e. at high fields and low temperatures) the electrons will not be scattered to other Landau levels. Given n_s carriers per unit area, we then expect that the longitudinal resistance $R_I = 0$ when $n_s = Nj = jeB/h$.

In practice, it is found for many samples that $R_I = 0$, and the Hall resistance is quantised at $R_H = h/je^2$ over a finite range of field in the neighbourhood of $B = hn_s/je$, as illustrated in Fig. 3.7.

The step heights in the quantum Hall effect can be measured to an accuracy of order a few parts in 10^9 and lead to an extremely accurate determination of h/e^2 =25812.807 Ω. A basic semiconductor experiment can, therefore, be used in defining fundamental constants (h or e), and also as a resistance standard, to define the ohm.

The model of the quantum Hall effect here is greatly oversimplified. It does not, for example, account for the width of the plateaux in R_H in Fig. 3.7. The plateaux width can be explained in terms of broadening of the Landau levels, for example, by impurities and the localisation of electron states in the wings of the broadened Landau levels. Conduction occurs through extended states and so, when the Fermi

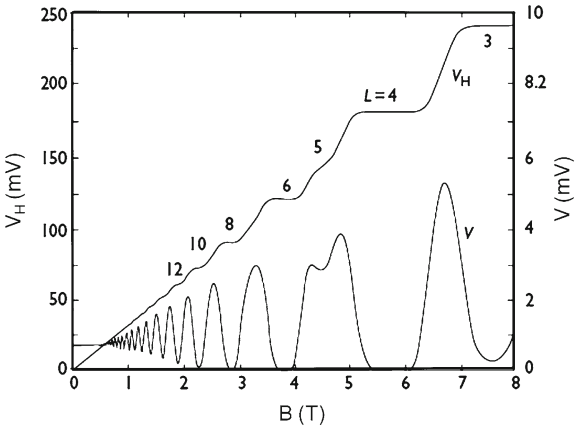


Fig. 3.7 Experimental curves for the longitudinal voltage (V) and the Hall voltage (V_H) of a heterostructure as a function of the magnetic field B for a fixed carrier density in the heterostructure (after [19])

energy lies well within the band of localised states, the conduction electrons again see no states close in energy to which they can scatter.

Once the magnetic field is sufficiently large that all electrons are in the lowest Landau level, there should be no further plateaux in the Hall resistance, R_H or zeros in R_I. It was, therefore, a further big surprise when plateaux and zeros were seen when the lowest level was one-third and two-thirds full, and then, as the material quality improved at further fractions such as 1/5, 2/5, 2/7, 2/9, etc [20]. The theory for these plateaux, to explain the fractional quantum Hall effect, requires many-electron effects which cause energy gaps to open up within the Landau levels: there are bound states containing, for example, three electrons whose excitations have an effective charge of 1/3, and which then account for the plateaux at 1/3 and 2/3. Further discussion of these states is beyond the scope of this book.

3.5 Spin Quantum Hall Effect

The existence and behaviour of edge states in insulating materials has only recently become a subject of interest. It was generally assumed for an insulator—where a gap separates the occupied and empty states—that no current should flow when conducting probes are attached to opposite ends of a sample, and a voltage is applied. This analysis does not however take account of the edge states which may be expected in a bounded insulator. Recent analysis has shown that different insulators can have distinctly different types of edge states. For the vast majority of insulators, the edge states do not provide a viable current path, and so do not need to be considered in discussion of the electronic properties. There are however a set of insulators for which the band structure is topologically distinct from conventional insulators, and where, just as in the quantum Hall effect, the edge states provide an extremely robust conduction path, with states at each edge supporting current flow in a given direction

by carriers with a distinct spin state [21]. The **spin** quantum Hall effect is then of interest both in the quest for spin-based electronic devices, and also in the search for topologically distinct states of matter.

In order to examine the spin quantum Hall effect and the different possible topologies for the band structure of an insulator, it is useful to consider a model $\boldsymbol{k}\cdot\boldsymbol{p}$ Hamiltonian describing the band dispersion due to the interaction between two $\boldsymbol{E}$lectron states (spin up and down) with s-like symmetry and two $\boldsymbol{H}$ole states (also spin up and down) with p-like symmetry in a quantum well (QW) structure [21]. The zero of energy is chosen midway between the electron and hole states, and the Hamiltonian associated with dispersion in the QW x–y plane is given by:

$$H = \begin{pmatrix} h(k) & 0 \\ 0 & h^*(-k) \end{pmatrix} \tag{3.36}$$

where $h(k)$ is a 2×2 matrix describing the interaction between the $\boldsymbol{E}$ and $\boldsymbol{H}$ spin up states, while $h^*(-k)$ describes the interaction between the spin down states, with $h(k)$ for a conventional III–V quantum well being of the form:

$$h(k) = \begin{pmatrix} \varepsilon(k) + M + B(k_x^2 + k_y^2) & A(k_x + ik_y) \\ A(k_x - ik_y) & \varepsilon(k) - M - B(k_x^2 + k_y^2) \end{pmatrix} \tag{3.37}$$

where the coefficients B and A can both be assumed to be positive numbers, while the term $\varepsilon(k) = C - D(k_x^2 + k_y^2)$ can be neglected in the following analysis, as it only describes the variation of the average energy associated with the two bands.

In a conventional QW structure, where the lowest confined electron state is above the highest confined hole state ($M > 0$) both bands show a conventional dispersion, with the band extremum energy at $k = 0$, and with the lowest $\boldsymbol{E}$ and highest $\boldsymbol{H}$ state at energy M and $-M$, respectively.

Consider however a material where the lowest $\boldsymbol{E}$ state is below the highest $\boldsymbol{H}$ state ($M < 0$). Such an arrangement is possible when the QW is formed from a zero-gap semiconductor, such as HgTe [22]. In the case where $M < 0$, the band extrema are no longer at $k = 0$, but instead are found at finite k, as illustrated in Fig. 3.9a, where the solid black lines show the calculated dispersion for the case $k_y = 0$.

In addition to propagating states within the bands of a periodic solid, there are also evanescent states in the energy gap, which join with the bulk band edges. Because the amplitude of the evanescent states grows exponentially, they are not valid solutions to Schrödinger's equation in an unbounded quantum well. The evanescent states can however give rise to edge states in a bounded quantum well, defined for instance in the range $0 < y < L$.

It is possible based on the Hamiltonian of (3.36) and (3.37) to use analytical models to calculate the edge states both for the conventional case of $M > 0$ and also for $M < 0$ [21]. The edge states join with the bulk bands at the band extrema, and have a topologically different behaviour in the two cases, with a non-trivial dispersion for $M < 0$, as shown by the dashed and dotted lines in Fig. 3.9a. The dashed line shows

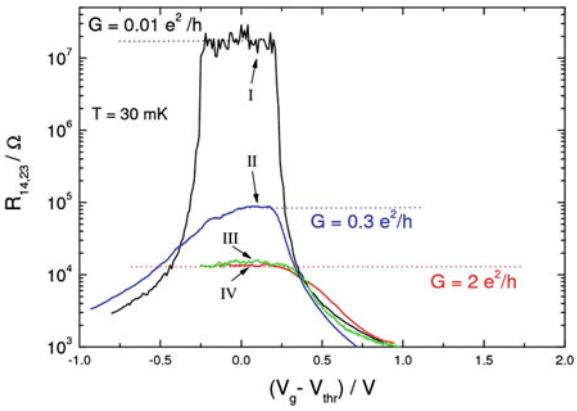


Fig. 3.8 Quantisation of conductance in units of $2e^2/h$ observed in CdTe/HgTe quantum wells (measurements III and IV). This was the first experimental signature of the spin Hall effect due to edge states in a 2-D topological insulator (after [23])

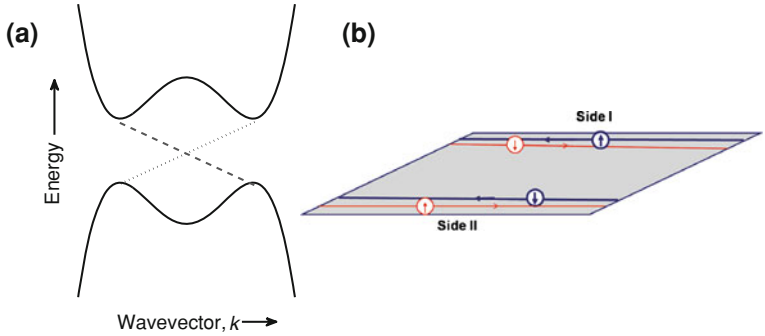


Fig. 3.9 **a** *Black lines* band dispersion for $k_y = 0$ of the Hamiltonian of (3.37); *dashed line* dispersion associated with edge states with spin up on side I and spin down on side II of the sample; *dotted line* dispersion for edge states with spin down on side I and spin up on side II, as illustrated in **b**, where applied voltage is driving net carrier flow to right

the dispersion associated with states with spin up on side I and spin down on side II of the sample, while the dotted line shows the dispersion for states with spin down on side I and spin up on side II. When an electron of a given spin is being accelerated by an applied electric field to a larger k_x value then there are no edge states with the same spin on that side of the sample to which the electron can be backscattered, thereby giving a robust conduction path, as illustrated schematically in Fig. 3.9b. The existence of such a topologically nontrivial phase in HgTe-based systems was first predicted theoretically [22], and subsequently demonstrated experimentally, through a series of elegant experiments on HgTe/CdTe quantum well structures, in which the Fermi level was tuned by a gate voltage to lie in the band gap region. These experiments demonstrated that the conductance σ was quantised, taking the value $\sigma = 2e^2/h$ predicted for topologically protected edge states in such a structure, as shown in Fig. 3.8 [23].

This experimental demonstration of the quantum spin Hall effect has placed the concept of topological insulating states on a firm footing, with the further extension of the concept to 3-D materials and with a growing number of predictions of novel effects related to the carrier spin polarisation at the interfaces between topological insulators and other materials, such as superconductors and ferromagnets.

3.6 Quantised Conduction Through Wires and Dots

As the size of current carrying structures shrinks down, quantum confinement effects can begin to play a key role in determining the conductance, in particular when the confining energies approach or exceed thermal energies in the device. When the lateral dimensions of a current carrying wire become small enough, the wire must be treated as a quantum wire, with states confined in the lateral direction, but with wavenumber k_z remaining a valid quantum number along the wire axis (taken here as the z direction). In addition to current flow along a quantum wire, it is also possible to establish current flow through a quantum dot (QD) structure: a region where carriers are effectively confined in all three dimensions, with the QD only weakly coupled to the surrounding current carrying regions. Current flow through such nanostructures has been of considerable interest both from a fundamental perspective, and also because of its potential consequences for future electronic devices. We provide a brief overview here of transport through wires and dots, outlining how the conductance can become quantised when confinement and thermal energies become comparable to each other. Further details and a more general discussion of transport through such nanostructures can be found in a wide range of texts, including for instance [24] or [25].

Figure 3.10 shows schematically a quantum dot channel region sandwiched between metallic source and drain regions. When a voltage V is applied across the structure, most of the voltage drop will occur across the channel region. The carriers in each of the metallic regions are expected to be in thermal equilibrium, with chemical potential μ_1 in the source, and μ_2 in the drain region. The probability of a given state in region i ($i = 1, 2$) being occupied is then given by:

$$f_i(E) = f_0(E - \mu_i) = \left[1 + \exp\{(E - \mu_i)/k_B T\}\right] \tag{3.38}$$

where $f_i(E)$ is the Fermi distribution function in region i. Each contact tries to equilibrate with the channel: the source therefore tries to pump electrons into the channel, while the drain tries to pull them out, leading to a net current flow through the channel. For n-type conduction, an electron will flow from a filled state in the source through a level in the channel that is *empty* at equilibrium, and on into an empty state in the drain. At very low temperature, when all states below μ_i are filled and all states above it empty, current can flow only through the finite (small) number of

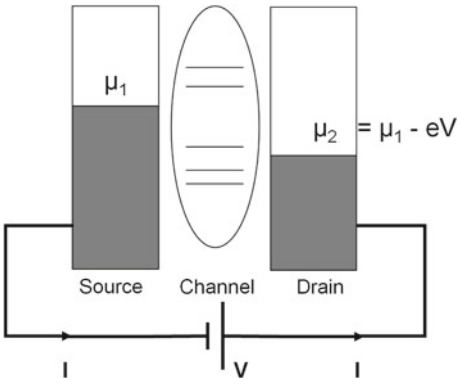


Fig. 3.10 Schematic of band line-up when a bias is applied between a source and drain across a quantum-dot channel region with quantised energy states

levels in the channel with energies between μ_2 and μ_1, and the conduction therefore becomes **quantised**. We now wish to estimate the current flow through a single such channel, using a formalism due to Landauer [26–28] to do so.

We assume for simplicity that the source and drain in Fig. 3.10 are ideal 1-D conductors. leading to reservoirs on the left and right with quasi-Fermi energies μ_1 and μ_2, respectively. The current injected from the left-hand side, I_L is found by integrating for all k states in the wire which propagate to the right (i.e. with $k > 0$) the probability of injection into that k state, $f_1(k)$, times $v(k)$, the velocity with which a particle moves in state k, times $T(E)$, the probability of a particle in that state being transmitted through the channel. A similar expression can be written down for the current injected into the channel from the right, I_R, and the net current $I = I_L - I_R$ is then given by:

$$I = \frac{2e}{2\pi}\left[\int_0^\infty \mathrm{d}k v(k) f_1(k) T(E) - \int_0^\infty \mathrm{d}k' v(k') f_2(k') T(E')\right] \tag{3.39}$$

where the constant is the 1-D density of states in k-space. We can use (3.4) to replace $v(k)$ by $\hbar^{-1}\mathrm{d}E/\mathrm{d}k$, thereby allowing us to transform (3.39) into an integral over energy. The upper limits of integration at low temperature are given by μ_1 and μ_2 for I_L and I_R, respectively, with I then given by:

$$I = \frac{2e}{2\pi\hbar}\int_{\mu_2}^{\mu_1} \mathrm{d}E T(E) \tag{3.40}$$

This is the Landauer formula for low-temperature quantised current flow.

If we assume that we are in the linear response regime, so that $\mu_1 - \mu_2 = eV$, where V is the applied voltage, and further assume that $T = 1$ for current flow through the given channel, so that the integral then equals $\mu_1 - \mu_2$, this gives that the conductance $G = I/V$ for a single current channel takes the value

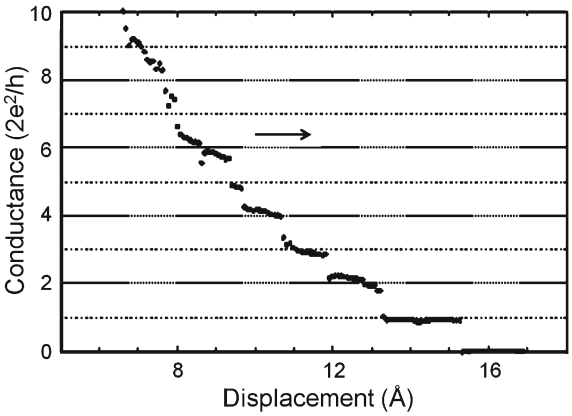


Fig. 3.11 Variation in conductance (in units of $2e^2/h$) of a metal point contact as the contact is pulled apart (after [29])

$$G = 2e^2/h \tag{3.41}$$

where $2e^2/h$ is referred to as the conductance quantum, with its inverse $h/2e^2$ referred to as the resistance quantum.

The derivation in (3.41) assumed that there was only one current path through the channel. It can be seen from Fig. 3.10 that the number of current channels can increase with increasing voltage, due to an increasing number of quantum dot levels becoming available as current paths. At low temperature, the total conductance then increases by $2e^2/h$ each time a new current path becomes available. In addition, the current flow through a channel can vary with channel area: the number of current paths N through a channel increases with the cross-sectional area of the channel, but in discrete steps. This is illustrated in Fig. 3.11, showing how the conductance of a metal point contact changes in units of $2e^2/h$ as the contact is pulled apart.

Several major simplifying assumptions were made in the derivation of the quantised conductance here, including in particular that we can set $T = 1$ in (3.40), and also that the full voltage V which is applied is dropped across the channel, with no voltage drop in the connecting wires. Both these assumptions need to be modified in a more detailed analysis [24], but the overall conclusion from the more detailed analysis remains unchanged concerning the value of the conductance quantum, consistent with its measured value across a wide range of experiments, including the spin quantum Hall measurements discussed in the previous section and point contact measurements illustrated here.

3.7 Graphene

The overall interest in nanoscale transport is driven both by the novel fundamental phenomena which emerge at the nanoscale and also by the demands of future nanoscale devices. Both these interests converge strongly in the case of graphene. It

is very well-known that graphite consists of hexagonal carbon sheets stacked on top of each other. Although the band structure [30] and some of the other properties of a single layer have been understood theoretically for a long time, it was believed until recently that it would not be possible to create and investigate experimentally high quality, single graphite layers. This situation changed dramatically in 2004 with the discovery by Andre Geim and Konstantin Novoselov of a surprisingly simple technique to fabricate such individual layers, referred to as graphene sheets [31]. Their fabrication method is surprisingly simple and, enabled by the details they provided, could be repeated by others in a very short time. In essence, it involves putting sticky tape onto a graphite surface, peeling off an individual layer and placing that layer for measurement and analysis on an inert substrate.

Graphene is a zero-gap semiconductor, It is the first truly 2-D crystalline material, being only one atomic layer thick, and has many remarkable electronic properties. We begin here by first describing the band structure of graphene and then overview how the band structure impacts on the electronic and optical properties.

The band structure of graphene is best understood, using the tight binding method, with one s and three p states on each C atom. The s state and the two p states which lie in the graphene plane form sp^2 hybrids on each C atom. These interact with hybrids on the neighbouring atoms to give bonding and anti-bonding states, separated by a wide energy gap. The band structure in the vicinity of the energy gap is then determined entirely by the interactions between the remaining p_z orbital on each of the C atoms.

Figure 3.12 shows the honeycomb lattice of graphene. There are two atoms per unit cell, with atom A at $(0,0)$ and atom B at $(0,d)$ for the unit cell that includes the origin in Fig. 3.12. In addition, atom A has two neighbours of type B at $\left(\pm\frac{\sqrt{3}}{2}d, -\frac{1}{2}d\right)$, and atom B also has two further neighbours of type A. We can write the Bloch states for such a lattice in the form

$$\psi_{\boldsymbol{k}\pm} = \sum_m \exp(i\boldsymbol{k}\cdot\boldsymbol{R}_m)\left[a_{\boldsymbol{k}\pm}\phi_{am} + b_{\boldsymbol{k}\pm}\phi_{bm}\right] \tag{3.42}$$

where ϕ_{am} (ϕ_{bm}) is the p_z orbital on atom A (B) in unit cell m, and $\boldsymbol{R}_m$ is the lattice vector linking unit cell 0 and m. We assume that each of the p_z states has self-energy E_p, and that there is an interaction U between the p_z states on nearest neighbour atoms. In order to solve the Schrödinger equation, we first multiply $H\psi_{\boldsymbol{k}\pm} = E\psi_{\boldsymbol{k}\pm}$ from the left by ϕ^*_{a0} and then integrate over all space. The only nonzero terms on the left-hand side are the self-interaction and the interaction with each of the three nearest neighbours, which gives the relation:

$$a_{\boldsymbol{k}\pm}E_p + b_{\boldsymbol{k}\pm}U\left[1 + \exp(i\boldsymbol{k}\cdot\boldsymbol{a}_1 + \exp(i\boldsymbol{k}\cdot\boldsymbol{a}_2)\right] = a_{\boldsymbol{k}\pm}E \tag{3.43}$$

where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are two lattice vectors, given by $\boldsymbol{a}_1 = \left(\frac{\sqrt{3}}{2}d, -\frac{3}{2}d\right)$ and $\boldsymbol{a}_2 = \left(-\frac{\sqrt{3}}{2}d, -\frac{3}{2}d\right)$. We find a similar expression to (3.43) when we pre-multiply the

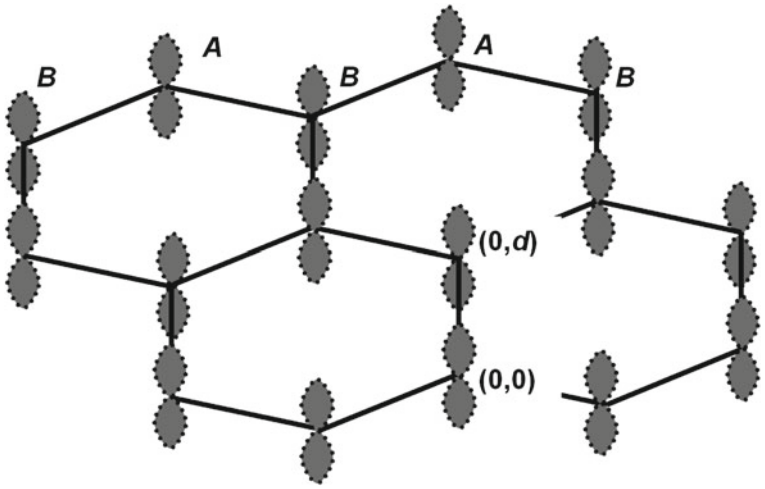


Fig. 3.12 Hexagonal lattice of graphene. There are two atoms per unit cell, with atom A at $(0,0)$ and B at $(0,d)$ in the unit cell at the origin here. The band structure of graphene close to the Fermi energy arises due to interactions between the p_z orbital on each C atom and its neighbours

Schrödinger equation by ϕ^*_{b0} and again integrate over all space. Simplifying the exponential terms in (3.43), the band dispersion near the energy gap in graphene can then be found by solving the 2×2 determinant:

$$\begin{pmatrix} E_p - E & U\left[1 + 2\exp\left(\frac{-i3k_y d}{2}\right)\cos\left(\frac{\sqrt{3}k_x d}{2}\right)\right] \\ U\left[1 + 2\exp\left(\frac{i3k_y d}{2}\right)\cos\left(\frac{\sqrt{3}k_x d}{2}\right)\right] & E_p - E \end{pmatrix} = 0 \tag{3.44}$$

Figure 3.13 shows the band structure of graphene calculated from (3.44). The separation between the bonding and anti-bonding states is greatest at $\boldsymbol{k} = 0$, where $E_{0\pm} = E_p \pm 3U$, but it can be seen that the band gap between the filled and empty states goes to zero at the Brillouin zone K point [$\boldsymbol{k}_K = \left(\frac{4\pi}{3\sqrt{3}d}, 0\right)$ and equivalent points]. In addition, the dispersion has a linear variation with $\boldsymbol{k}$ close to the K point, with $E = \pm\hbar v_F|\boldsymbol{k} - \boldsymbol{k}_K|$.

This linear E–k relation gives graphene many of its special properties. The Fermi level in intrinsic (undoped) graphene lies at $E_F = E_p$. Because the density of states is zero at this point, the electrical conductivity of intrinsic graphene is then very low, being of the order of the conductance quantum, $\sigma \sim e^2/h$, with the exact prefactor still being debated. The linear E–k relation is similar to the dispersion relation for photons. The electron density can be changed with a gate potential and, by shifting the Fermi level, one can have electrons or holes with very similar properties. The mobility of these charge carriers is extremely high ($\sim 10^5$ cm^2/(Vs) at room temperature in

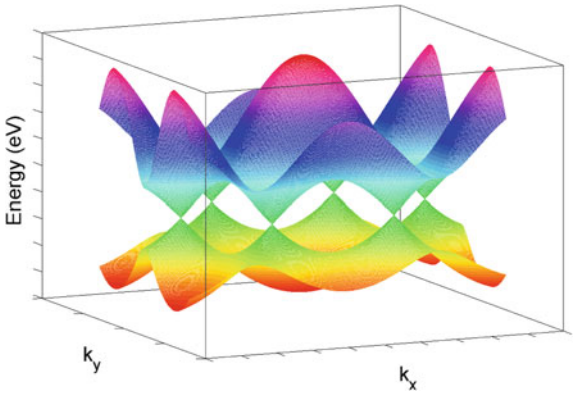


Fig. 3.13 The band structure of graphene calculated from (3.44). The band gap between the filled and empty states goes to zero at the Brillouin zone K point, and the dispersion has a linear variation with $\boldsymbol{k}$ close to K

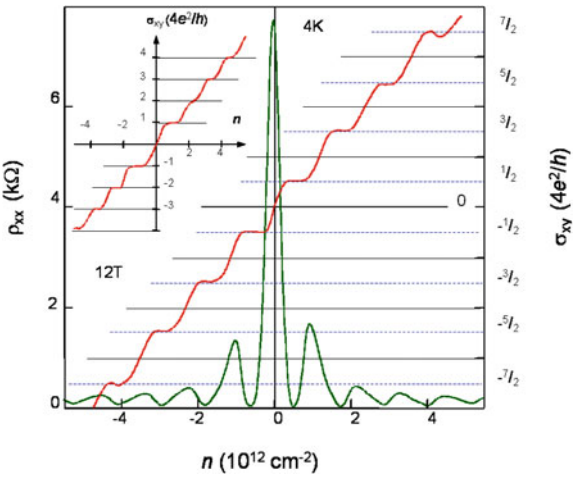


Fig. 3.14 Quantum Hall effect in graphene, with Hall plateaux at values of $\left(n+\frac{1}{2}\right)4e^2/h$, where $n>0$ for electron conduction and $n<0$ for hole conduction (after [32])

the best samples). With a linear dispersion, the carriers behave as massless relativistic fermions, and are best described using the Dirac equation. One result of the special dispersion relation is that the quantum Hall effect becomes unusual in graphene (Fig. 3.14), with the spacing between Hall plateaux equal to $4e^2/h$, twice as large as that in conventional 2-D structures, and with the first plateaux for electrons and holes occurring at $\pm 2e^2/h$, respectively.

With its fascinating properties and ease of fabrication for experimental analysis, there has been an explosion of interest in graphene. The behaviour of graphene bilayers has also been widely studied. It can be shown by extending the tight binding model of (3.42)–(3.44) that there are significant differences between the band structure of graphene bilayers and of single-sheet graphene. Because it is only one atomic layer thick, graphene is also practically transparent—the absorption coefficient per layer of graphene is about 2.3%. Graphene is very interesting for a diverse range of applications. The very high mobility could be very beneficial for high frequency

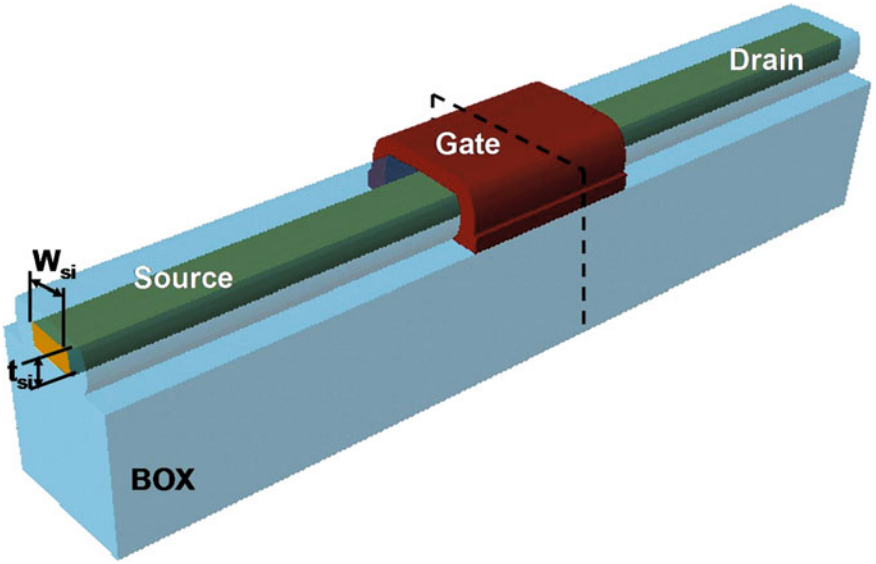


Fig. 3.15 Schematic of an *n*-channel nanowire transistor. The underlying insulator layer (buried oxide) is not shown. The silicon nanowire is uniformly doped *n*-type and the gate material is *p*-type to enable pinch-off. Opposite dopant polarities can be used for *p*-channel devices

electronic applications. As an ultimately thin and mechanically very strong material, it also has high potential for use as a transparent conductor, in applications such as touch screens, light panels and solar cells. Many challenges remain to develop techniques to produce graphene on an industrial scale, but it is clear that the material could bring enormous benefits across a wide range of electronic and other applications.

3.8 Junctionless Transistor

The major technological driver behind the investigation of electron transport at the nanoscale is the continued shrinkage in the size of the transistors in integrated circuits, referred to as Moore's Law. This has resulted in the number of transistors on a single microchip increasing from a few hundred in the early 1970s to over several billion today. Conventional devices are now approaching their limits, and a wide range of approaches are being pursued to allow the continued improvement in device density. Some of these approaches, such as the use of high-*k* dielectrics to reduce the thickness of capacitative layers, can benefit existing devices [33]. Other approaches are targeting exotic new materials, such as the use of carbon nanotube field-effect transistors (CNFETs), that can operate at room temperature and are capable of digital switching using a single electron [34].

All current transistors are based on the use of *p–n* junctions to control the device switching on and off. Because a *p–n* junction relies on the sharp transition between a region doped with donor atoms and a region doped with acceptor atoms, it becomes increasingly difficult to achieve the controlled formation of a junction with decreasing device dimension. Interestingly, an Austro-Hungarian physicist, Julius Edgar Lilienfield proposed in 1925 that it could be possible to achieve transistor action in a piece of semiconductor material with only one type of carrier—a conventional resistor—if the semiconductor layer could be made thin enough to allow for a gate to control the density of electrons, and thus the current flow through the piece of semiconductor. This could not be demonstrated in the macroscale structures investigated at the time, but has recently been shown by our colleague at Tyndall, Jean-Pierre Colinge, to be a viable and very promising approach on the nanoscale.

He and his co-workers fabricated heavily *n*-doped silicon nanowires with cross-sectional dimensions of $\sim 10 \times 10$ nm, and with a section of the wire of length ~ 50 nm covered on three sides by a p-doped gate, as illustrated schematically in Fig. 3.15 [3]. The heavy *n* doping ensures that the nanowires are highly conducting when the gate is off. Application of a gate voltage pinches off current flow through the gated section, thereby giving transistor action.

Proof-of-concept simulations of junctionless gated Si nanowire transistors, based on a first-principles approach, predict that Si-based transistors are physically possible without major changes in design philosophy down to scales of ~ 1 nm wire diameter and ~ 3 nm gate length, with the junctionless transistor avoiding potentially serious difficulties affecting junctioned channels at these length scales [35]. The junctionless structure is relatively simple to build, even at the nanoscale, compared to conventional junction fabrication technologies, which are becoming increasingly complex, and therefore looks to be one of the more promising approaches to develop and extend silicon technology beyond its current limits.

3.9 Summary and Conclusions

This is the first of three chapters to address electronic transport in semiconductor materials and heterostructures. In order to set up the framework for these three chapters, we began by presenting an overview of some of the key factors relevant to electron transport on a macroscopic scale. We then discussed how quantum effects come into play as structure size is scaled down. Many classical phenomena display measurable quantum character as the dimensions in which current flow become restricted. These include not just the well-known quantum Hall effect, and quantised conductance through nanoscale current channels, but also the recent realisation of quantised conductance along the edges of a "topological insulator". As the name suggests, the band structure of a topological insulator is topologically different from that of a conventional semiconductor, with distinctly different gap states at the edges or surfaces of the material. We showed that this has surprising consequences, includ-

ing the possibility to achieve a spin quantum Hall effect in zero magnetic field in suitably chosen samples.

The overall interest in nanoscale transport is driven both by the novel fundamental phenomena which emerge at the nanoscale and also by the demands of future nanoscale devices, some of which may exploit these phenomena. We saw how both these interests converge in the case of graphene, and then finally considered the junctionless transistor, as one of the most promising future device concepts currently being investigated. Overall, the results and topics considered here emphasise the ongoing interest in nanoscale transport, and also provide the background relevant to the more detailed discussion of high-field transport and Monte Carlo techniques in Chaps. 4 and 5.

Acknowledgments We thank several colleagues for their input and support both to the preparation of this chapter and also for the Training School lectures on which it was based, including Chris Broderick, Jean-Pierre Colinge, Conor Coughlan Conor Coughlan, Giorgos Fagas, Stephen Fahy and Jim Greer. We also thank the Science Foundation Ireland for financial support.

References

1. J. Bardeen, W.H. Brattain, Phys. Rev. **74**(2), 230–231 (1948)
2. G.E. Moore, Cramming more components onto integrated circuits. Electron. Mag. **38**(8), 4 (1965)
3. J.P. Colinge, C.W. Lee, A. Afzalian, N.D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White et al., Nat. Nanotechnol. **5**(3), 225–229 (2010)
4. F. Murphy-Armando, S. Fahy, Phys. Rev. Lett. **97**(9), 96606 (2006)
5. S. Hiyamizu, J. Saito, K. Nanbu, T. Ishikawa, Jpn. J. Appl. Phys. **22**(10), L609–L611 (1983)
6. T. Kitatani, M. Kondow, T. Kikawa, Y. Yazawa, M. Okai, K. Uomi, Jpn. J. Appl. Phys. (Part 1) **38**, 5003 (1999)
7. O.F. Sankey, J.D. Dow, K. Hess, Appl. Phys. Lett. **41**(7), 664–666 (1982)
8. M.A. Fisher, A.R. Adams, E.P. O'Reilly, J.J. Harris, Phys. Rev. Lett. **59**(20), 2341–2344 (1987)
9. J. Yu, Influence of N incorporation on the electronic properties of dilute nitride (In)GaAsN alloys. Ph.D. Thesis, University of Michigan, 2010
10. S. Fahy, A. Lindsay, H. Ouerdane, E.P. O'Reilly, Phys. Rev. B **74**, 035203 (2006)
11. W. Shan, W. Walukiewicz, J.W. Ager, E.E. Haller, J.F. Geisz, D.J. Friedman, J.M. Olson, S.R. Kurtz, Phys. Rev. Lett. **82**(6), 1221–1224 (1999)
12. S. Fahy, E.P. O'Reilly, Appl. Phys. Lett. **83**, 3731 (2003)
13. K. Volz, J. Koch, B. Kunert, W. Stolz, J. Cryst. Growth **248**, 451–456 (2003)
14. M. Reason, Y. Jin, H.A. McKay, N. Mangan, D. Mao, R.S. Goldman, X. Bai, C. Kurdak, J. Appl. Phys. **102**, 103710 (2007)
15. P.R.C. Kent, A. Zunger, Phys. Rev. B **64**(11), 115208 (2001)
16. A. Lindsay, E.P. O'Reilly, Phys. Rev. Lett. **93**(19), 196402 (2004)
17. J.R. Hook, H.E. Hall, *Solid State Physics*, 2nd edn. (Wiley, Chichester, 1991)
18. K. von Klitzing, Rev. Mod. Phys. **58**, 519 (1988)
19. M.E. Cage, R.F. Dziuba, B.F. Field, IEEE Trans Instrum Meas **34**(2), 301–303 (1985)
20. R. Willett, J.P. Eisenstein, H.L. Störmer, D.C. Tsui, A.C. Gossard, J.H. English, Phys. Rev. Lett. **59**(15), 1776–1779 (1987)
21. M. König, H. Buhmann, L.W. Molenkamp, T.L. Hughes, C.X. Liu, X.L. Qi, S.C. Zhang, J. Phys. Soc. Jpn. **77**, 031007 (2008)
22. B.A. Bernevig, T.L. Hughes, S.C. Zhang, Science, **314**(5806), 1757 (2006)

23. M. König, S. Wiedmann, C. Brüne, A. Roth, H. Buhmann, L.W. Molenkamp, X.L. Qi, S.C. Zhang, Science, **318**(5851), 766 (2007)
24. D.K. Ferry, S.M. Goodnick, *Transport in Nanostructures*, vol. 6 (Cambridge University Press, Cambridge, 1997)
25. T. Ouisse, *Electron Transport in Nanostructures and Mesoscopic Devices* (Wiley Online Library, London, 2008)
26. R. Landauer, IBM J. Res. Dev. **1**(3), 223–231 (1957)
27. R. Landauer, Philos. Mag. **21**, 863–867 (1970)
28. R. Landauer, Phys. Lett. A **85**(2), 91–93 (1981)
29. M. Brandbyge, J. Schiøtz, M.R. Sørensen, P. Stoltze, K.W. Jacobsen, J.K. Nørskov, L. Olesen, E. Lægsgaard, I. Stensgaard, F. Besenbacher, Phys. Rev. B **52**, 8499 (1995)
30. P.R. Wallace, Phys. Rev. **71**(9), 622 (1947)
31. A.K. Geim, K.S. Novoselov, Nat. Mater. **6**(3), 183–191 (2007)
32. K.S. Novoselov, A.K. Geim, S.V. Morozov, D. Jiang, M.I. Katsnelson, I.V. Grigorieva, S.V. Dubonos, A.A. Firsov, Nature **438**, 197 (2005)
33. R. Chau, J. Brask, S. Datta, G. Dewey, M. Doczy, B. Doyle, J. Kavalieros, B. Jin, M. Metz, A. Majumdar et al., Microelectr. Eng. **80**, 1–6 (2005)
34. H.W.C. Postma, T. Teepen, Z. Yao, M. Grifoni, and C. Dekker. Science **293**(5527), 76 (2001)
35. L. Ansari, B. Feldman, G. Fagas, J.P. Colinge, J.C. Greer, Appl. Phys. Lett. **97**, 062105 (2010)

Chapter 4
Hot Electron Transport

Martin P. Vaughan

Abstract In a high electric field, a population of electrons may be driven out of thermal equilibrium with the crystal lattice, hence becoming 'hot'. In this chapter, the basic concepts of hot electron transport in semiconductors are introduced following a semiclassical approach. Scattering mechanisms pertinent to hot electron transport are described, including phonon, electron–electron and alloy scattering. The high-field phenomena of avalanche breakdown and negative differential resistance are discussed qualitatively in terms of the underlying physics and as a motivation for device applications. Techniques to solve the Boltzmann transport equation are then introduced. A low-field solution, including an introduction to the ladder method for dealing with polar optical phonon scattering, is first discussed as a foundation for the subsequent high-field solution.

4.1 Introduction

Hot electron transport in semiconductors pertains to electrical conductivity in high electric fields. Interest in the high-field electronic properties of semiconductors was originally motivated by the desire to understand electrical breakdown in insulators. Since then, hot electron dynamics have been actively exploited in high-field devices. Impact ionisation, the smoking gun primarily responsible for electrical breakdown, is used as a photo-current gain mechanism in the avalanche photodiode (APD) [1], whilst the phenomenon of negative differential resistance (NDR) is exploited in Gunn diodes [2] to produce microwave oscillations. More recently, novel hot electron lasers have been developed, such as the hot electron light emitting and lasing semiconductor heterostructure (HELLISH) [3] and Gunn [4] lasers.

M. P. Vaughan
Tyndall National Institute Cork Ireland
e-mail: martin.vaughan@tyndall.ie

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_4,

In this chapter, we introduce the reader to some of the basic concepts of hot electron transport following a semiclassical approach. At the heart of this approach is the determination of the distribution function for a population of non-equilibrium electrons. After laying down the basic groundwork, we tackle this problem via solution of the Boltzmann transport equation. Perhaps a more popular approach to solving this problem, especially in high fields, is that of Monte Carlo simulation. This is the subject of the chapter by Vogiatzis and Rorison in this book and we leave it to the interested reader to decide his or her preferred method.

Due to limitations of space, the introduction to the subject given here is necessarily limited. In particular, there is no specialization to low dimensional devices and our discussion of scattering mechanisms pertinent to high-field transport is not claimed to be comprehensive. For those requiring a specific understanding of high-field transport in low-dimensions, we recommend Ridley's review on the subject [5]. It is hoped that this work may provide a fairly gentle introduction to the more detailed literature given in the references.

In the remainder of this introductory section, we introduce the fundamental concepts of lattice temperature and non-equilibrium, or 'hot' electrons, before specifying more exactly the scope of this chapter and giving a general overview.

4.1.1 The Lattice Temperature T_0

In physics, the concepts of temperature and thermal equilibrium are quite fundamental, being enshrined in the 'zeroth' law of thermodynamics. In this law, bodies in thermal equilibrium are defined to have the same temperature. Perhaps a more intuitive picture emerges when we consider physical systems at the microscopic level, where temperature becomes a measure of the average energy of a quantum of a system. In a crystal lattice, energy is stored in the mechanical vibrations of the ions and these are quantised as phonons. Specifically, it is on the basis of the average energy of the *acoustic* phonons (see Sect. 4.3.4) that we may define the lattice temperature T_0.

Since these are bosons, the statistical distribution of phonons over energy is governed by the Bose–Einstein factor

$$n_{\mathbf{q}} = \frac{1}{e^{\hbar\omega_{\mathbf{q}}/k_B T_0} - 1}, \tag{4.1}$$

where $n_{\mathbf{q}}$ is the number of phonons in the mode with wavevector $\mathbf{q}$, $\omega_{\mathbf{q}}$ is the energy of the phonon and k_B is Boltzmann's constant.

4.1.2 Electrons in Thermal Equilibrium

In a population of electrons in thermal equilibrium with the crystal lattice, the average electronic energy will again be given in terms of T_0. In this case, however, electrons are fermions, so their energetic distribution is given by the Fermi-Dirac factor

$$f_0(\epsilon_{\mathbf{k}}) = \frac{1}{1 + \exp\left(\left[\epsilon_{\mathbf{k}} - \epsilon_F\right]/k_B T_0\right)}. \tag{4.2}$$

Here, $f_0(\epsilon_{\mathbf{k}})$ is to be interpreted as the probability that an electronic state with energy $\epsilon_{\mathbf{k}}$ (and labelled by wavevector $\mathbf{k}$) will be occupied. At the absolute zero of temperature, the electronic states fill up completely from the lowest energy to those states with the Fermi energy ϵ_F. Strictly speaking, this is actually fixed for a given system but it is more common in practice to think of the Fermi level as being able to move due to environmental conditions. This should then be referred to as the chemical potential but we shall retain what has become the conventional notation and refer to ϵ_F.

It will be useful in later sections to note that

$$\frac{df_0}{d\epsilon_{\mathbf{k}}} = -\frac{f_0(\epsilon_{\mathbf{k}})\left(1 - f_0(\epsilon_{\mathbf{k}})\right)}{k_B T_0}. \tag{4.3}$$

4.1.3 Hot Electrons

Under non-equilibrium conditions, such as the application of a high electric field over a material sample, the electrons of the system may be driven to higher energy states, thus becoming 'hot'. In such circumstances, we may be able to characterise the electronic population by an electron temperature T_e, such that $T_e > T_0$. Whilst this is a slight abuse of the thermodynamical definition of temperature and lacks a precise formulation, it remains a useful intuitive description relating to the (definable) average electronic energy.

4.1.4 Scope and Overview

It may be argued that a proper treatment of transport in solids should be purely quantum mechanical, incumbent with all the interference effects that wave-particle duality entails. Certainly, as the size of semiconductor devices gets ever smaller, a rigorous quantum mechanical treatment seems increasingly justifiable. However, in practice we often find that quantum effects become washed out by the many interactions that the charge carriers in the system undergo with their environment. In the language of quantum mechanics, we can refer to this as decoherence, since it is

the loss of phase information that leads to the disappearance of interference effects. Equivalently, since the time-dependence of the phase (i.e. the angular frequency ω of a particle) is directly proportional to the particle's energy ϵ, we can describe this in terms of the inelastic interactions that the carriers undergo, principally with the crystal lattice via the electron–phonon interaction.

The many interactions a charge carrier undergoes may be dealt with via scattering theory, based on quantum mechanical perturbation theory (see Sect. 4.3). Whilst this is an inherent feature of a purely quantum mechanical approach, an accurate description of the transport properties of a system can very often be obtained by combining scattering theory with statistical physics. This is known as the semiclassical approach and at its heart is the idea of distribution function, giving the probability of an electron occupying a particular state. The Fermi-Dirac function of (4.2) is a special case of this for thermal equilibrium. In transport theory, what is of interest is the non-equilibrium distribution (discussed in greater detail in Sect. 4.2.5).

After establishing some basic concepts in Sect. 4.2, some of the scattering mechanisms pertinent to hot electron transport are introduced in Sect. 4.3. This section is not claimed to be comprehensive but rather representative, covering inelastic scattering via the electron–phonon interaction and some of the more common elastic scattering mechanisms. A more qualitatively discussion of high-field phenomena is then given in Sect. 4.4. Here, we focus on avalanche breakdown and negative differential resistance. The discussion is principally from the point of view of the underlying physics, although this section is also intended as a motivation for device applications. Finally, in Sect. 4.5 we take on the solution of the Boltzmann transport equation. In the first part of this section, we start with a low-field solution. This introduces some of the concepts we will need for the high-field solution as well as giving some insight into the problem of the relaxation time for polar optical phonon scattering. We deal with this via the ladder method, which represents a physically more realistic approach to the problem than assuming a well-defined relaxation time.

4.2 Basic Concepts

4.2.1 Ballistic Transport

Before considering the more general problem of electronic transport with scattering, we first consider ballistic transport, in which the electron travels only under the influence of an applied electric field. The situation is illustrated schematically in Fig. 4.1a, where the application of an electric field $\mathbf{E}$ gives rise to a spatially varying potential $V(x)$. An electron with a total energy ϵ travelling ballistically in the conduction band for a distance Δx gains a kinetic energy $eE\Delta x$ above the conduction band edge, where e is the magnitude of the electronic charge. If $eE\Delta x$ is significantly greater than the thermal energy $k_B T_0$, then we may describe the electron as being 'hot'.

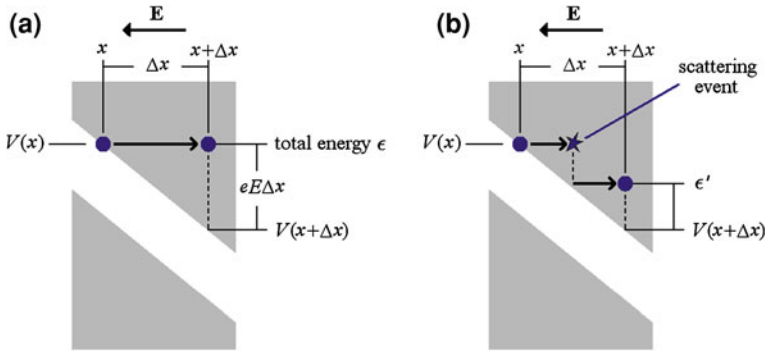


Fig. 4.1 **a** The application of an electric field **E** gives rise to a spatially varying potential $V(x)$. If the electron travels ballistically for a distance Δx, it gains a kinetic energy $e\Delta Ex$ above the conduction band edge. **b** The electron undergoes an energy relaxing (inelastic) scattering event, changing to a state with energy ϵ'

4.2.2 Energy and Momentum Relaxation

Relaxation Times

An electron will not continue to travel ballistically indefinitely. At some point it is likely to scatter, with a consequent change in its momentum and, in inelastic interactions, its energy. Such an inelastic scattering event is illustrated in Fig. 4.1b. Here, the electron undergoes an energy relaxing event, changing to a state with energy ϵ'.

To describe such events statistically, we may define an energy relaxation time $\tau_\epsilon(\epsilon)$ as the average time an electron of energy ϵ will travel before undergoing an inelastic scattering event. Note that since $\epsilon = \hbar\omega$, where $\hbar$ is Dirac's constant, and is therefore related to the phase of the electronic state, $\tau_\epsilon(\epsilon)$ may also be thought of as a coherence time.

In the same way, we may also define a momentum relaxation time $\tau_{\mathbf{k}}(\epsilon_{\mathbf{k}})$ as the average time an electron with wavevector $\mathbf{k}$ will travel before undergoing a scattering event, changing its momentum from $\hbar\mathbf{k}$ to $\hbar\mathbf{k}'$ (without necessarily changing its energy). Note that, particularly at low T_0, we usually have $\tau_\epsilon(\epsilon) \gg \tau_{\mathbf{k}}(\epsilon_{\mathbf{k}})$, meaning an electron may change momentum many times before losing coherence.

Intrinsic Scattering Rates

Before discussing any particular scattering processes, we introduce the concept of the intrinsic scattering rate $s(\mathbf{k}', \mathbf{k})$, by which we mean the probability per unit time that a state $|\mathbf{k}'\rangle$ makes a transition to another state $|\mathbf{k}\rangle$ due to some perturbing potential $V(\mathbf{r})$ (we take the $\mathbf{r}$ dependence of this potential to be tacit and abbreviate to V).

This rate may be derived from time-dependent perturbation theory and is given by

$$s(\mathbf{k}', \mathbf{k}) = \frac{2\pi}{\hbar} \left| \langle \mathbf{k} | V | \mathbf{k}' \rangle \right|^2 \delta \left(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}} + \Delta\epsilon \right). \tag{4.4}$$

This form allows the interaction between states to be inelastic, with $\Delta\epsilon$ representing the energy difference between initial and final energies. Since, overall, energy must be conserved, the $\Delta\epsilon$ must be taken up elsewhere in the system. In this chapter, the only inelastic processes that we will consider will be those due to the electron–phonon interaction (due to the time scales involved, in transport theory the electron–photon interaction, for instance, is considered a rare event).

To obtain the total scattering rates for a given $\mathbf{k}$-vector, will need to perform a summation over all $\mathbf{k}'$-states. Although in reality the states are discrete, the physical size of a system is usually large enough that we can take $\mathbf{k}'$ to vary continuously. This means that whenever we have to sum a particular quantity over the states of the system to determine a macroscopic transport property, we may transform to an integration via the rule

$$\sum_{\mathbf{k}} \rightarrow \frac{V_C}{(2\pi)^3} \int d^3\mathbf{k}, \tag{4.5}$$

where V_C is the crystal volume and $(2\pi)^3 / V_C$ is the (3- D) volume of reciprocal space occupied by a $\mathbf{k}$-state. For lower dimensions this result will need some modification. For a cubic volume of side L, we would have $V_C/(2\pi)^3 \rightarrow (L/(2\pi))^m, d^3\mathbf{k} \rightarrow d^m\mathbf{k}$, where m is the dimension, and the integral would need further summation over the discrete states due to quantum confinement.

Using this notation we are now in a position to define a momentum relaxation time for a purely elastic scattering event (for which $\Delta\epsilon = 0$)

$$\frac{1}{\tau_{\mathbf{k}}(\epsilon_{\mathbf{k}})} = \int s(\mathbf{k}', \mathbf{k}) \left(1 - \cos\alpha'\right) \frac{V_C}{(2\pi)^3} d^3\mathbf{k}' = w(\epsilon_{\mathbf{k}}), \tag{4.6}$$

where α' is the angle between $\mathbf{k}$ and $\mathbf{k}'$ and $w(\epsilon_{\mathbf{k}})$ is the energy-dependent scattering rate. As we shall see in Sect. 4.5.3, if the squared matrix element contains no α' dependence, this expression is equal to Fermi's Golden rule for the energy dependent scattering rate.

Energy and Momentum Relaxation Rates

In an inelastic process, an electron may gain or lose energy. In the context of phonon scattering, we shall speak of rates for absorption, when the electron gains the energy of a phonon, and emission, where the electron loses energy to the lattice. Denoting these processes by the subscripts A and E respectively, the rate of change of energy for an electron with initial energy $\epsilon_{\mathbf{k}}$ may be defined as

$$\left(\frac{\mathrm{d}\epsilon_{\mathbf{k}}}{\mathrm{d}t}\right)_s = \int |\Delta\epsilon| \left\{s_A(\mathbf{k}', \mathbf{k}) - s_E(\mathbf{k}', \mathbf{k})\right\} \frac{V_C}{(2\pi)^3} d^3\mathbf{k}', \tag{4.7}$$

where the s subscript indicates that this is the rate for scattering processes. Whether or not this rate is negative will depend on the electron's energy relative to the thermal energy of the lattice. For hot electrons, the rate must be negative in accordance with the second law of thermodynamics. The energy relaxation time for hot electrons may then be defined over this high-energy range by

$$\left(\frac{\mathrm{d}\epsilon_{\mathbf{k}}}{\mathrm{d}t}\right)_s = -\frac{\epsilon_{\mathbf{k}} - \epsilon_0}{\tau_e\left(\epsilon_{\mathbf{k}}\right)}, \tag{4.8}$$

where ϵ_0 is the energy at which $(\mathrm{d}\epsilon_{\mathbf{k}}/\mathrm{d}t)_s = 0$.

We may also define the rate of change of momentum for an electron with initial wavevector $\mathbf{k}$ along similar lines as

$$\left(\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t}\right)_s = \int \hbar\mathbf{q}\{s_A(\mathbf{k}', \mathbf{k}) + s_E(\mathbf{k}', \mathbf{k})\} \frac{V_C}{(2\pi)^3} d^3\mathbf{k}', \tag{4.9}$$

where $\mathbf{q} = \mathbf{k}' - \mathbf{k}$. Here, the rates are added, as the momentum change for emission will be $-\mathbf{q}$. As the components of $\mathbf{q}$ perpendicular to $\mathbf{k}$ will cancel on integration, we may put $\mathbf{q} \rightarrow (k' \cos\alpha' - k)\hat{\mathbf{k}}$, where $\hat{\mathbf{k}}$ is the unit vector in the direction of $\mathbf{k}$. Furthermore, $s_A(\mathbf{k}', \mathbf{k}) + s_E(\mathbf{k}', \mathbf{k})$ is the total intrinsic scattering rate $s(\mathbf{k}', \mathbf{k})$ for a given process, so we may now put

$$\left(\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t}\right)_s = -\hbar\mathbf{k} \int s(\mathbf{k}', \mathbf{k}) \left(1 - \frac{k'}{k}\cos\alpha'\right) \frac{V_C}{(2\pi)^3} d^3\mathbf{k}'. \tag{4.10}$$

For elastic processes, $k' = k$ and the integral just becomes $1/\tau_{\mathbf{k}}(\epsilon_{\mathbf{k}})$ as in (4.6). More generally, we may define

$$\left(\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t}\right)_s = -\frac{\hbar\mathbf{k}}{\tau_{\mathbf{k}}\left(\epsilon_{\mathbf{k}}\right)}. \tag{4.11}$$

4.2.3 Describing Energy Bands

In this chapter, we shall assume that the dispersion relations, i.e. the variation of electronic energy with wavevector, are well defined, so that we may label electronic states unambiguously by $\mathbf{k}$. Moreover, we shall assume a periodic crystal lattice, so that energy levels $\epsilon = \epsilon(\mathbf{k}+\mathbf{G})$, where $\mathbf{G}$ is a reciprocal lattice vector, are folded back to the point $\mathbf{k}$ in the reduced Brillouin zone (i.e. the periodically repeating primitive cell in $\mathbf{k}$-space—see, for instance, Ref. [6]). This gives rise to different energy bands associated with a given wavevector.

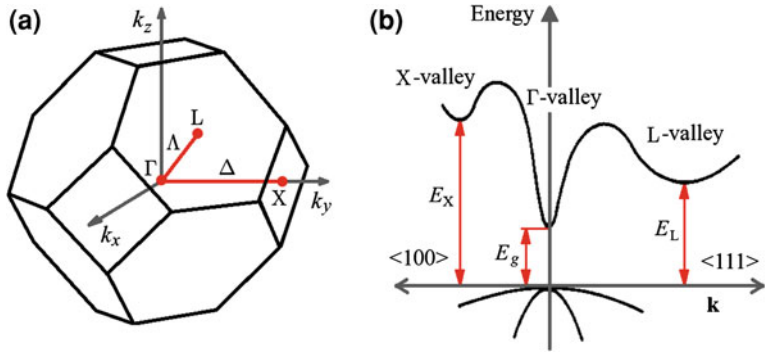


Fig. 4.2 **a** The Brillouin zone of a face-centred cubic crystal, showing the Γ, X and L symmetry points; **b** schematic band-structure of a direct band-gap semiconductor showing the Γ, X and L valleys

As we shall be primarily concerned with electron transport, it is the conduction bands that we shall focus on here. Since these bands have minima at some point in **k**-space, we often refer to these as 'valleys' and are associated with special points of high symmetry in the Brillouin zone. In a face-centred-cubic (FCC) crystal, the valleys of particular interest are the Γ valley, with a minimum at $\mathbf{k} = 0$, the X or Δ valley, with a minimum somewhere along the Δ line and the L valley with a minimum along the Λ line in the $\langle 111 \rangle$ direction (see Fig. 4.2).

Spherical and Spheroidal Valleys

Treating the conduction valleys exactly according to their dispersion relations over the entire Brillouin zone is an onerous task, so certain approximations may be assumed to make transport calculations more tractable. In particular, it usually proves to be a good approximation if we only try to model a valley close to its minimum at some wavevector $\mathbf{k}_0$ and consider how isotropic the valley looks from this point. If, locally, the states of equal energy lie on a sphere centred on the valley minimum, then we describe the band as being 'spherical'. This often proves to be the case for the Γ valley. More generally, the valleys along the Δ and Λ lines tend to be spheroidal, so that states of equal energy lie on a spheroid that has the symmetry line (Δ or Λ) as its major axis. Obviously, as we move towards the Brillouin zone boundaries, the isotropy of the bands no longer holds.

Next, we look at the way the energy varies with the magnitude of **k** away from the valley minima. Close to a minimum, the dispersion relations are qualitatively similar to those for free electrons, i.e. they are parabolic. Specifically, we would have for a spherical energy band

$$\epsilon_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m^*}, \tag{4.12}$$

where m^* is the effective mass at the band-edge. For spheroidal bands, we decompose k^2 into the components parallel to and perpendicular to the symmetry line, which are referred to as the longitudinal and transverse directions respectively. These directions will also be associated with different effective masses m_l^* and m_t^*, so that, relative to the **k**-vector of the band minimum, we have

$$\epsilon_{\mathbf{k}} = \frac{\hbar^2 k_l^2}{2m_l^*} + \frac{\hbar^2 k_t^2}{2m_t^*}. \tag{4.13}$$

Very often, spheroidal valleys can be dealt with by making a transformation to a coordinate system in which they are spherical. We shall not pursue this analysis explicitly in this chapter but assume instead that this can be done and just use the simpler formulae for spherical valleys.

Non-Parabolicity

At higher energies, the conduction bands generally become non-parabolic. This is dealt with formally by defining a function of energy

$$\gamma(\epsilon_{\mathbf{k}}) = \frac{\hbar^2 k^2}{2m^*}, \tag{4.14}$$

where the particular form of $\gamma(\epsilon_{\mathbf{k}})$ must be fitted to actual band-structure calculations. Since we are concerned with high-field transport, we will retain this more general expression and use formulae derived from it throughout this chapter.

As a particular example of non-parabolicity, consider the dispersion relations in a direct band-gap semiconductor, for which the band-gap, ϵ_g, is much larger than the spin-orbit splitting energy. In this case, it can be shown via $\mathbf{k} \cdot \mathbf{p}$ theory that $\gamma(\epsilon_{\mathbf{k}})$ may be approximated by [7, 8]

$$\gamma(\epsilon_{\mathbf{k}}) = \epsilon_{\mathbf{k}} \left(1 + \alpha\epsilon_{\mathbf{k}} + \beta\epsilon_{\mathbf{k}}^2\right), \tag{4.15}$$

where

$$\alpha = \frac{1}{\epsilon_g}\left(1 - \frac{m^*}{m_0}\right)^2, \tag{4.16}$$

$$\beta = -\frac{2}{\epsilon_g^2}\frac{m^*}{m_0}\left(1 - \frac{m^*}{m_0}\right)^3 \tag{4.17}$$

and m_0 is the free electron mass.

IntraValley and InterValley Scattering

Scattering processes that take an electron from a state with wavevector $\mathbf{k}$ to a state with wavevector $\mathbf{k}'$ may be broadly categorized into two types. In intravalley scattering the initial and final wavevectors all lie within the same valley. For low-field transport in direct band-gap materials (i.e. those in which the Γ valley band-edge is the lowest conduction band energy), it is usually sufficient to treat only this kind of scattering. For high-field transport, on the other hand, it is often important to know the distribution of electrons in different valleys and the rate of intervalley scattering between them. This is crucial, for instance, to the understanding of negative differential resistance due to transferred electrons, as we shall see in Sect. 4.4.2.

Bloch Functions

Generally we shall assume that the wavefunctions of the system are Bloch functions. In a periodic structure, such wavefunctions may be given according to Bloch's theorem by

$$\psi_{n,\mathbf{k}} = V_C^{-1/2} \sum_{\mathbf{G}} C_{n,\mathbf{k}+\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \tag{4.18}$$

where n labels the band index, $\mathbf{k}$ the wavevector and the $\mathbf{G}$ are reciprocal lattice vectors of the primitive cell. The factor multiplying the plane-wave $e^{i\mathbf{k}\cdot\mathbf{r}}$,

$$u_{n,\mathbf{k}} = V_C^{-1/2} \sum_{\mathbf{G}} C_{n,\mathbf{k}+\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}}, \tag{4.19}$$

thus has the periodicity of the lattice, with the normalisation condition on the coefficients $C_{n,\mathbf{k}+\mathbf{G}}$

$$\sum_{\mathbf{G}} \left| C_{n,\mathbf{k}+\mathbf{G}} \right|^2 = 1. \tag{4.20}$$

4.2.4 Group Velocity and the Density of States

The dispersion relations defined in (4.12), (4.13) or (4.14) determine the electronic group velocity $\mathbf{v}(\mathbf{k}) = \nabla_{\mathbf{k}} \epsilon_{\mathbf{k}} / \hbar$. Now since

$$\nabla_{\mathbf{k}} \gamma(\epsilon_{\mathbf{k}}) = \frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}} \nabla_{\mathbf{k}} \epsilon_{\mathbf{k}} = \frac{\hbar^2 \mathbf{k}}{m^*}, \tag{4.21}$$

we have

$$\mathbf{v}(\mathbf{k}) = \frac{\hbar \mathbf{k}}{m^*} \left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}} \right)^{-1} \tag{4.22}$$

and the energy dependent magnitude of $\mathbf{v}(\mathbf{k})$ is

$$v(\epsilon_{\mathbf{k}}) = \left(\frac{2\gamma(\epsilon_{\mathbf{k}})}{m^*}\right)^{1/2} \left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}}\right)^{-1}. \tag{4.23}$$

One further essential ingredient required is the density of states (DOS), which is the number of states per unit energy (although this is sometimes given in the literature as the number of states per unit energy per unit volume). In a non-parabolic energy band for which the dispersion relations are well-defined, this is given by

$$D(\epsilon_{\mathbf{k}}) = V_C \frac{(2m^*)^{3/2}}{4\pi^2\hbar^3} \gamma(\epsilon_{\mathbf{k}})^{1/2} \frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}} \tag{4.24}$$

(note that V_C is often omitted from this expression in the literature). The DOS in a spheroidal valley can be rendered in exactly the same way by substituting the density of states effective mass m_d^*, defined by

$$m_d^* = \left(m_l^* m_t^{*2}\right)^{1/3}, \tag{4.25}$$

for m^*. Where a function of wavevector, $g(\epsilon_{\mathbf{k}})$, actually only depends on the energy, we may now transform integrals over wavevector to integrals over energy via the following rule:

$$\frac{V_C}{(2\pi)^3} \int g(\epsilon_{\mathbf{k}})\ d^3\mathbf{k} \rightarrow \sum_n \int_{\epsilon_n}^{\infty} g(\epsilon_{\mathbf{k}}) D_n(\epsilon_{\mathbf{k}})\ d\epsilon_{\mathbf{k}}, \tag{4.26}$$

where the summation is over bands, so ϵ_n and $D_n(\epsilon_{\mathbf{k}})$ are the band-edge energy and DOS respectively of the nth band.

4.2.5 The Non-Equilibrium Distribution Function

Dealing with the plethora of interactions a particle experiences is generally more tractable using a semiclassical approach rather than a full quantum mechanical treatment. In the semiclassical approach, the wavefunctions and energies of the system are still obtained quantum mechanically but the occupancies of the single particle states are assumed to be given by some semiclassical distribution in which interference effects are neglected. Thus, we assume the existence of an electron distribution function $f(\mathbf{k})$ that gives the probability of an electron being in a region of $\mathbf{k}$-space close to wavevector $\mathbf{k}$. Since electrons are fermions, for each $\mathbf{k}$ point we may only have a maximum occupancy of two electrons, each having opposite spin. Generally we do not need to label the spin explicitly.

One caveat that does need to be applied here is that assuming a distribution function $f(\mathbf{k})$ implies assuming $\mathbf{k}$ to be a good quantum number. This means that, at the very least, we can label states unambiguously in terms of their momenta. This will not generally be the case in disordered materials where the electronic states become highly localised in real space (and hence, extended in $\mathbf{k}$-space). In such cases, charge transport may occur via 'hopping' conduction where an electron jumps discontinuously from one localised state to another.

The essence of transport theory calculations is that we are trying to find $f(\mathbf{k})$ under non-equilibrium conditions, i.e. under the influence of some applied force and / or temperature gradient in the material. Electronic conduction may therefore be of any physical quantity the carriers can transport, e.g. charge, spin or energy. In this chapter, we shall assume no temperature gradient or magnetic field and limit our attentions to charge transport in the presence of an electric field $\mathbf{E}$.

The dynamics of $f(\mathbf{k})$ are then governed by two processes. First, there is the change in energy and momentum of the particles under the influence of the electric field. Second, there will be the scattering of the particles due to varied interactions that generally act to relax the energy and randomise the momentum. The states that a carrier can occupy are usually those solved for the system under equilibrium conditions, whilst the scattering processes are found from perturbation theory. The strength of the scattering then determines an average time, τ, between scattering events as discussed in Sect. 4.2.2. On average, a carrier will pick up a wavevector shift of $\delta\mathbf{k} = -e\mathbf{E}\tau/\hbar$, giving an overall displacement to $f(\mathbf{k})$. We use this explicitly in the linearized distribution function for the low-field solution of the Boltzmann equation in Sect. 4.5.2.

4.2.6 Transport Properties

Under the semiclassical approach, the starting point for the determination of the macroscopic transport properties of a material is the expression for the current density $\mathbf{j}$

$$\mathbf{j} = -\frac{2e}{(2\pi)^3}\int \mathbf{v}(\mathbf{k})\, f(\mathbf{k})\, d^3\mathbf{k}. \tag{4.27}$$

Here, the factor of 2 accounts for spin and the negative sign has been inserted for consistency since, by convention, $\mathbf{j}$ is in the opposite direction to the electron flow. Note that the omission of the factor of the crystal volume, V_C, introduces dimensions of reciprocal volume.

Equation (4.27) may be compared to the phenomenological expression for the current density

$$\mathbf{j} = \sigma(\mathbf{E})\mathbf{E}, \tag{4.28}$$

where $\sigma(\mathbf{E})$ is the conductivity tensor. The electric field dependence of $\mathbf{j}$ must enter through the distribution function, which to first order will be linear in $\mathbf{E}$. Hence, for

low-field solutions of $f(\mathbf{k})$, $\sigma(\mathbf{E})$ is constant. As we move to higher electric fields, the higher order terms in $f(\mathbf{k})$ become increasingly significant and $\sigma(\mathbf{E})$ becomes field dependent.

Another common transport property is the mobility, defined by

$$\mu(\mathbf{E}) = \frac{\sigma(\mathbf{E})}{en}, \tag{4.29}$$

where n is the free carrier density, i.e. the density of electrons excited into the conduction bands,

$$n = \frac{2}{(2\pi)^3} \int_{CB} f(\mathbf{k})\, d^3\mathbf{k}, \tag{4.30}$$

where the integration is just over those states in the conduction bands. Another expression for the current density is then

$$\mathbf{j} = ne\mu(\mathbf{E})\mathbf{E} = ne\mathbf{v}_D(\mathbf{E}). \tag{4.31}$$

Here, $\mathbf{v}_D(\mathbf{E})$ is the drift velocity. Hence, when μ is constant, it is the rate of change of $\mathbf{v}_D(\mathbf{E})$ with respect to $\mathbf{E}$.

4.2.7 The Conservation Equations

The Balance of Energy and Momentum

The dynamics of an electron in a material sample over which an electric field $\mathbf{E}$ is applied may be described by equations expressing the conservation of energy and momentum. The balance of energy for a given electron is

$$\frac{\mathrm{d}\epsilon_{\mathbf{k}}}{\mathrm{d}t} = -e\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}} + \left(\frac{\mathrm{d}\epsilon_{\mathbf{k}}}{\mathrm{d}t}\right)_s, \tag{4.32}$$

where $\mathbf{v}_{\mathbf{k}}$ is a shortened notation for the group velocity. Here, the first term on the right hand side is the energy the electron gains from the field and the second term is the energy it loses to the lattice. Similarly, for the balance of momentum we have

$$\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t} = -e\mathbf{E} + \left(\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t}\right)_s. \tag{4.33}$$

Note that since

$$\frac{\mathrm{d}\epsilon_{\mathbf{k}}}{\mathrm{d}t} = \frac{\hbar\mathbf{k}}{m^*} \cdot \frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t} \left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}}\right)^{-1}, \tag{4.34}$$

where we have used (4.14), these coupled equations are clearly nonlinear, even for parabolic bands.

To describe the macroscopic transport properties, (4.32) and (4.33) must be averaged over the electron distribution $f(\mathbf{k})$. We define this averaging process for a general quantity $Q(\mathbf{k})$ by

$$\langle Q(\mathbf{k}) \rangle = A \int_{CB} f(\mathbf{k}) Q(\mathbf{k}) \, d^3\mathbf{k}, \tag{4.35}$$

where

$$A^{-1} = \int_{CB} f(\mathbf{k}) \, d^3\mathbf{k}. \tag{4.36}$$

Using (4.35), the balance equations averaged over the free electron population then become

$$\frac{\mathrm{d} \langle \epsilon_{\mathbf{k}} \rangle}{\mathrm{d}t} = -e\mathbf{E} \cdot \mathbf{v}_D + \left\langle \frac{\mathrm{d}\epsilon_{\mathbf{k}}}{\mathrm{d}t} \right\rangle_s \tag{4.37}$$

and

$$\frac{\mathrm{d} \langle \hbar\mathbf{k} \rangle}{\mathrm{d}t} = -e\mathbf{E} + \left\langle \frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t} \right\rangle_s . \tag{4.38}$$

Note that, using the Boltzmann factor $f(\mathbf{k}) = \exp(-\epsilon_{\mathbf{k}}/k_B T_e)$ in (4.35) and (4.36) for a spherical, parabolic band, we would obtain the thermal energy

$$\langle \epsilon_{\mathbf{k}} \rangle = \frac{3}{2} k_B T_e, \tag{4.39}$$

from classical kinetic theory. Equation (4.39) therefore gives us a rule-of-thumb relation between average electron energy and electron temperature.

Relaxation Time Approximations

We may gain some insight into the hot electron dynamics by returning to (4.32) and (4.33) and substituting in the definitions of the energy and momentum relaxation times given by (4.8) and (4.11), yielding

$$\frac{\mathrm{d}\epsilon_{\mathbf{k}}}{\mathrm{d}t} = -e\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}} - \frac{\epsilon_{\mathbf{k}} - \epsilon_0}{\tau_e} \tag{4.40}$$

and

$$\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t} = -e\mathbf{E} - \frac{\hbar\mathbf{k}}{\tau_{\mathbf{k}}}. \tag{4.41}$$

Before proceeding, it should be noted that these equations are now no longer strictly accurate. In the first place, as commented in Sect. 4.2.2, the energy relaxation

time used in (4.40) is only really appropriate for hot electrons. More generally, a relaxation time cannot always be well-defined, as we shall see in Sect. 4.5.3 in the case of polar optical phonon scattering.

Using the definition of group velocity given by (4.22) together with (4.41), we find

$$\mathbf{v}_{\mathbf{k}} = -\frac{\tau_{\mathbf{k}}}{m^*}\left(\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t} + e\mathbf{E}\right)\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}}\right)^{-1}. \tag{4.42}$$

Combining this with (4.40) and making use of (4.34), we obtain

$$\frac{\mathrm{d}\hbar\mathbf{k}}{\mathrm{d}t}\cdot\left(\frac{\hbar\mathbf{k}}{m^*} - \frac{e\tau_{\mathbf{k}}}{m^*}\mathbf{E}\right) = \frac{(eE)^2\,\tau_{\mathbf{k}}}{m^*} - \frac{(\epsilon_{\mathbf{k}} - \epsilon_0)}{\tau_e}\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}}. \tag{4.43}$$

In the steadystate, (4.42) and (4.43) reduce to

$$\mathbf{v}_{\mathbf{k}} = -\frac{e\tau_{\mathbf{k}}}{m^*}\mathbf{E}\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}}\right)^{-1} \tag{4.44}$$

and

$$\epsilon_{\mathbf{k}} = \epsilon_0 + \frac{(eE)^2\,\tau_{\mathbf{k}}\tau_e}{m^*}\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon_{\mathbf{k}}}\right)^{-1}. \tag{4.45}$$

Field Dependencies of $\epsilon_{\mathbf{k}}$ and $\mathbf{v}_{\mathbf{k}}$

As Ridley points out in Ref. [9], the field dependencies of $\mathbf{v}_{\mathbf{k}}$ and $\epsilon_{\mathbf{k}}$ may be neatly illustrated by assuming energy dependencies for the relaxation times of the form $\tau_{\mathbf{k}}(\epsilon) = A\epsilon^p$ and $\tau_e(\epsilon) = B\epsilon^q$. For simplicity, we shall take the variation of $(d\gamma/d\epsilon_{\mathbf{k}})^{-1}$ with energy to be small so that, taking the derivative of (4.45) with respect to $\epsilon_{\mathbf{k}}$ and re-arranging, we have

$$\epsilon_{\mathbf{k}} = \left(\frac{(eE)^2\,(p+q)\,AB}{m^*\mathrm{d}\gamma/\mathrm{d}\epsilon_{\mathbf{k}}}\right)^{1/(1-p-q)}. \tag{4.46}$$

Hence, the energy varies with the field as

$$\epsilon_{\mathbf{k}} \propto E^{2/(1-p-q)}. \tag{4.47}$$

Note that, for $p+q > 1$, we would have the physically invalid situation of the energy going to infinity at $E = 0$ and approaching zero asymptotically thereafter, whilst $\epsilon_{\mathbf{k}}$ is undefined for $p+q = 1$. Under these conditions, then, there can be no steady-state solution and high-field electron dynamics may be unstable.

The field variation of the magnitude of $\mathbf{v}_{\mathbf{k}}$ is now given by

$$v_{\mathbf{k}} = -\frac{eAE}{m^* \mathrm{d}\gamma/\mathrm{d}\epsilon_{\mathbf{k}}} \left(\frac{(eE)^2 (p+q) AB}{m^* \mathrm{d}\gamma/\mathrm{d}\epsilon_{\mathbf{k}}} \right)^{p/(1-p-q)}, \tag{4.48}$$

so

$$v_{\mathbf{k}} \propto \left(\frac{E^{(1+p-q)}}{(\mathrm{d}\gamma/\mathrm{d}\epsilon_{\mathbf{k}})^{(1-q)}} \right)^{1/(1-p-q)}. \tag{4.49}$$

Neglecting the variation of $(\mathrm{d}\gamma/\mathrm{d}\epsilon_{\mathbf{k}})^{-1}$, we see that for $q = p + 1$, the field dependency of $v_{\mathbf{k}}$ disappears. Such cases are referred to as velocity saturation or, when this happens in the averaged electron population, drift velocity saturation.

For $q < p + 1$ in the region of $p - q$ space where steady-state solutions exist, it can be shown that the exponent $(1 + p - q)/(1 - p - q)$ is always positive, meaning that $v_{\mathbf{k}}$ increases with increasing E. In the particular case when $v_{\mathbf{k}}$ varies linearly with E, we must have $p = 0$ and hence a constant momentum relaxation time.

Where the energy dependencies are such that $q > p + 1$, we find that $v_{\mathbf{k}}$ decreases with increasing E. This is not a tenable situation over all energy ranges since, physically, we must have $v_{\mathbf{k}} = 0$ at $E = 0$. However, it may be that as different scattering processes become predominant at higher energies, the energy dependence of the relaxation times will change. Thus we can envisage, in principle at least, situations in which $v_{\mathbf{k}}$ turns over and starts to decrease at higher fields due scattering processes alone. This would be an example of a negative differential resistance, which we discuss in greater detail in Sect. 4.4.2.

A more common mechanism for NDR would be due to the non-parabolicity of the band. Using the form for $\gamma(\epsilon)$ given in (4.15), we see from (4.49) that $v_{\mathbf{k}}$ becomes multiplied by a factor

$$\left(\frac{d\gamma}{d\epsilon_{\mathbf{k}}} \right)^{-1/(1-p-q)} = \left(\frac{1}{1 + 2\alpha\epsilon_{\mathbf{k}} + 3\beta\epsilon_{\mathbf{k}}^2} \right)^{1/(1-p-q)}. \tag{4.50}$$

For $p + q < 1$, for which steady-state solutions exist, this factor decreases monotonically with increasing energy, which, by (4.47) increases monotonically with E. Hence, this factor will act in opposition to processes that would otherwise lead to an increase in $v_{\mathbf{k}}$, possibly leading to velocity saturation or a decrease in $v_{\mathbf{k}}$ with increasing field.

4.3 Scattering Mechanisms

4.3.1 General Comments

So far we have only discussed scattering processes in the abstract, doing little more than differentiating between elastic and inelastic mechanisms. In this section, we

go a little way towards rectifying this deficiency. Whilst not a comprehensive or in-depth discussion of scattering processes, we highlight some of the more important phenomena and flesh out a little more of the physics.

Scattering processes may be divided into two classes: impurity scattering, in which an electron interacts with a localised perturbation of the potential, and phonon scattering, in which the electron interacts with the extended oscillations of the crystal lattice. The former class of processes are usually taken to be elastic, whilst phonon scattering always involves some exchange of the electronic energy with the lattice. However, in the case of acoustic phonons, this energy exchange is usually small enough to consider this type of scattering to be elastic.

4.3.2 Electron–Electron Scattering

Coulomb Scattering

The interaction of charged particles is governed by the Coulomb potential. However, in a solid material, an electron will rarely see a bare Coulomb potential due to the redistribution of free charge, thus screening the charge centre. In the most simple model of this, the Coulombic potential energy seen by an electron is then weighted by a decaying exponential

$$V(\mathbf{r}) = -\frac{eQ}{4\pi\varepsilon\,|\mathbf{r}-\mathbf{R}|}e^{-q_0|\mathbf{r}-\mathbf{R}|}, \tag{4.51}$$

where ε is the permittivity of the medium and q_0 is a reciprocal screening length. This may be defined by [9]

$$q_0^2 = \frac{e^2 n}{\epsilon k_B T}, \tag{4.52}$$

where n is the free electron density. Note that T is the temperature emerging from the Fermi-Dirac factor, so is the electron temperature, although no consideration of hot electron effects is assumed here. The scattering matrix due to (4.51) is then (to good approximation) given by

$$\left\langle \psi_{n'}, \mathbf{k}' \,|V|\, \psi_{n,\mathbf{k}} \right\rangle = -J^{n'n}_{\mathbf{k}'\mathbf{k}} \frac{eQ/(\varepsilon V_C)}{|\mathbf{k}-\mathbf{k}'|^2 + q_0^2}, \tag{4.53}$$

where $J^{n'n}_{\mathbf{k}'\mathbf{k}}$ is the overlap of the periodic parts of the Bloch functions integrated over the primitive cell.

Using (4.53) for a fixed charge centre, one may derive the Brooks-Herring result for ionized impurity scattering [10]. However, the $\left|\mathbf{k}-\mathbf{k}'\right|^2$ term in the denominator, has the consequence that the scattering rate (apart from the depen-

dence of a screening factor involving q_0) varies with energy as $\gamma(\epsilon_{\mathbf{k}})^{-3/2}$. Hence, ionized impurity scattering becomes weak at high energies and the process is not significant for hot electron transport.

On the other hand, for electron–electron scattering the matrix element involves the product of two electron states, for which it is the total energy and momentum that must be conserved. In this case, the matrix element contains a term like $\left|\mathbf{k}_{12} - \mathbf{k}'_{12}\right|^2$ in the denominator, where $\hbar\mathbf{k}_{12}$ is the relative momentum between electrons. This means that for electrons with similar momenta, the absolute wavevector, and hence energy, of either particle will not be a limiting factor. Note that electron–electron scattering cannot have any net relaxation on the momentum of a population of electrons but will act to relax the relative momentum of pairs of electrons, thus randomizing the overall distribution in **k**-space. The reader may find explicit expressions and derivations for this class of scattering in Ref. [9].

Impact Ionisation

A particular case of electron–electron scattering of great importance in hot electron transport is that of impact ionisation, in which high-energy carriers create electron–hole pairs in collision events. Suppose an electron has been accelerated by an applied electric field and has acquired an energy $\Delta\epsilon$ above the conduction band edge, such that $\Delta\epsilon$ is greater than the band-gap energy ϵ_g. Under these conditions, this electron may collide with another electron in the valence band with sufficient energy to ionize the latter, producing an electron–hole pair in addition to the original electron.

An analogous situation pertains for a hole travelling in the valence band. In this case, we imagine a hole travelling in the opposite direction to an electron, eventually obtaining an energy $\Delta\epsilon$ below the valence band edge. Although the term ‘cold’ hole might seem more appropriate, this is usually termed a ‘hot’ hole. An electron close to the valence band edge may then drop into the hole, giving up the energy it has lost to another valence band electron and ionizing it into the conduction band. Thus, we now have two holes and an electron.

A rate for impact ionisation may be derived from constructing the matrix element in terms of the carriers (now in different bands) and the Coulomb potential. However, it turns out that more significant than this is the probability that a given carrier will obtain the necessary threshold energy to create an electron–hole pair. This is covered in Sect. 4.4.1 along with a discussion of the phenomena of avalanche multiplication.

4.3.3 Alloy Scattering

In an alloy, the periodicity of the crystal lattice is broken up by the random positions of the substitutional components. This is usually treated on the basis of the virtual crystal approximation (VCA) due to Nordheim [11]. In this model, the potential

seen by the carriers is divided into two parts: a periodic part, taken to be a linear interpolation between the alloying species (i.e. the 'virtual crystal') and a random part seen as a perturbation that gives rise to scattering. Hence, two fundamental assumptions are that there exist Bloch solutions for the virtual crystal and that the random part of the potential is small enough to be dealt with via perturbation theory. This is not likely to be the case in materials that exhibit extreme disorder, where typically the wavefunctions may become highly localised. However, within the remit of this current text, according to which the dispersion relations are assumed to be well-defined, we shall consider only 'well-matched' alloys, for which the substituting atomic species have similar covalent radii and electronegativities, and which can be grown without significant defects.

The formalism of the VCA was developed in the context of metallic alloys by Flinn [12], Ash and Hall [13, 14] to deal with perturbing potentials extending beyond the primitive cell of the crystal incorporating short-range order or clustering. In the generally accepted model of alloy scattering in semiconductors, Harrison and Hauser [15] assume completely random alloys and model the perturbation as a potential step associated with some characteristic energy difference, $\Delta\epsilon$, between the atomic species. Various candidates have been suggested for $\Delta\epsilon$, including electronegativity difference, band-offset and electron affinity difference. However, more often in practice, the alloy scattering potential is fitted to the experimentally observed mobilities.

More recently, Murphy-Armando and Fahy [16] have pursued a first-principles approach to alloy scattering, in which the scattering matrix is obtained from band-structure calculations based on density functional theory (DFT). At present, this approach is limited to group IV alloys, for which DFT calculations yield realistic dispersion relations. In this case, the well-known problem that DFT does not accurately predict the bandgap is not an issue since it is only the difference in the eigenvalues that is of importance in the method (see Ref [17] for more details on DFT).

For an alloy of the form A_xB_{1-x}, where x is the molar fraction of component A, the intrinsic alloy scattering rate, including intervalley scattering, may be written as

$$s(n', \mathbf{k}'; n, \mathbf{k}) = \frac{2\pi}{\hbar} \frac{x\,(1-x)}{N_C} \left| \left\langle V^{n'n}_{\mathbf{k}'\mathbf{k}} \right\rangle \right|^2 \delta\left(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}}\right), \tag{4.54}$$

where N_C is the number of primitive cells in the crystal (or supercell in the case of the first-principles approach). The matrix element giving the transition probability between states is

$$\left\langle V^{n'n}_{\mathbf{k}'\mathbf{k}} \right\rangle = N_C \left\langle \psi_{n',\mathbf{k}'} \right| \Delta V_A - \Delta V_B \left| \psi_{n,\mathbf{k}} \right\rangle, \tag{4.55}$$

where the $\psi_{n,\mathbf{k}}$ are the Bloch functions of the virtual crystal and ΔV_A and ΔV_A are the perturbing potentials arising from substitution of atoms of type A and B respectively. Assuming that the energy dependence of the scattering matrix is weak and that it has no angular dependence, (4.54) may be substituted into (4.6) to obtain

a total scattering rate

$$w(\epsilon_{\mathbf{k}}) = \frac{2\pi}{\hbar}\frac{x\,(1-x)}{N_C}\sum_{n'}\left|\left\langle V^{n'n}_{\mathbf{k}'\mathbf{k}}\right\rangle\right|^2 D_{n'}(\epsilon_{\mathbf{k}}), \tag{4.56}$$

where the summation is over final valleys.

Now the energy dependence of (4.56) arises solely from the density of states, which varies as $\gamma^{1/2}(\epsilon_{\mathbf{k}})\mathrm{d}\gamma/\mathrm{d}\epsilon_{\mathbf{k}}$. Hence, at least up to very high energies, this is a monotonically increasing function of $\epsilon_{\mathbf{k}}$ and so will be of importance in hot electron transport. Moreover, alloy scattering provides a mechanism for intervalley scattering, which may have a significant effect on high-field electron dynamics. On the other hand, being an elastic process, alloy scattering cannot provide a means of energy relaxation.

4.3.4 Phonons

The Electron–Phonon Interaction

The electron–phonon interaction gives rise to inelastic scattering, in which the electron can lose or gain energy to or from the crystal lattice. With this type of interaction the states involved in the scattering matrix are products of electron and phonon states. The phonon states are described in the mode occupation representation, so that $\left|n_{\mathbf{q}}\right\rangle$ is the state containing $n_{\mathbf{q}}$ phonons of wavevector $\mathbf{q}$.

The action of the electron–phonon interaction potential is such that a single phonon is either added or subtracted to a mode, described as the electron emitting or absorbing a phonon respectively. Overall, both energy and momentum are conserved. Hence, the energy of the electron either decreases or increases by the phonon energy $\hbar\omega_{\mathbf{q}}$, whilst the electronic wavevector changes by the phonon wavevector $\mathbf{q}$.

Denoting the electron–phonon interaction matrix element for scattering from a combined state $\left|n'_{\mathbf{q}}, \mathbf{k}'\right\rangle$ to $\left|n_{\mathbf{q}}, \mathbf{k}\right\rangle$ by $M^{n'n}_{\mathbf{k}'\mathbf{k}}$, the intrinsic scattering rate is given by

$$s(\mathbf{k}', \mathbf{k}) = s_A(\mathbf{k}', \mathbf{k})\delta\left(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}} + \hbar\omega_{\mathbf{q}}\right) + s_E(\mathbf{k}', \mathbf{k})\delta\left(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}} - \hbar\omega_{\mathbf{q}}\right), \tag{4.57}$$

where

$$s_A(\mathbf{k}', \mathbf{k}) = \frac{2\pi}{\hbar}\left|M^{n'n'-1}_{\mathbf{k}'\mathbf{k}}\right|^2 \tag{4.58}$$

and

$$s_E(\mathbf{k}', \mathbf{k}) = \frac{2\pi}{\hbar}\left|M^{n'n'+1}_{\mathbf{k}'\mathbf{k}}\right|^2. \tag{4.59}$$

Note that $s(\mathbf{k}', \mathbf{k})$ is the scattering from $\mathbf{k}'$ to $\mathbf{k}$, so since $\delta\left(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}} + \hbar\omega_{\mathbf{q}}\right)$ implies $\epsilon_{\mathbf{k}} = \epsilon_{\mathbf{k}'} + \hbar\omega_{\mathbf{q}}$, $s_A(\mathbf{k}', \mathbf{k})$ is the intrinsic scattering rate for the absorption of a phonon

by the new $\mathbf{k}$-state. Hence the phonon mode loses a phonon and the new occupation number is $n_{\mathbf{q}} = n'_{\mathbf{q}} - 1$. Similarly, $s_E(\mathbf{k}', \mathbf{k})$ is the rate for phonon emission, meaning the electron has lost energy and the phonon mode increases by unity.

The matrix element $M^{n'n}_{\mathbf{k}'\mathbf{k}}$ is given by [18]

$$M^{n'n'-1}_{\mathbf{k}'\mathbf{k}} = i\, I_{\mathbf{k}'\mathbf{k}} n_{\mathbf{q}}^{1/2} \left(\frac{\hbar C_{\mathbf{q}}^2}{2 N_C M \omega_{\mathbf{q}}} \right)^{1/2} \delta_{\mathbf{k}'+\mathbf{q},\mathbf{k}} \tag{4.60}$$

for absorption and

$$M^{n'n'+1}_{\mathbf{k}'\mathbf{k}} = -i\, I_{\mathbf{k}'\mathbf{k}} \left(n_{\mathbf{q}} + 1 \right)^{1/2} \left(\frac{\hbar C_{\mathbf{q}}^2}{2 N_C M \omega_{\mathbf{q}}} \right)^{1/2} \delta_{\mathbf{k}'-\mathbf{q},\mathbf{k}} \tag{4.61}$$

for emission. $C_{\mathbf{q}}$ is the electron–phonon coupling coefficient for the particular process and M is a characteristic mass for the oscillator. The Kronecker delta serves to enforce the conservation of total momentum due to the electron gaining or losing the phonon momentum.

The overlap factor $I_{\mathbf{k}'\mathbf{k}}$ is defined in terms of the Bloch functions, $\psi_{\mathbf{k}}(\mathbf{r})$,

$$I_{\mathbf{k}'\mathbf{k}} = \int \psi^*_{\mathbf{k}'}(\mathbf{r})\, \psi_{\mathbf{k}}(\mathbf{r})\; d^3\mathbf{r} \tag{4.62}$$

where the integral is over the primitive cell and normalised such that $I_{\mathbf{k}\mathbf{k}} = 1$ for all $\mathbf{k}$. For large $\mathbf{q}$-vector, $I_{\mathbf{k}'\mathbf{k}}$ will generally be less than unity, becoming more markedly so as the band structure becomes more non-parabolic. As a first approximation, however, we may take $I_{\mathbf{k}'\mathbf{k}} = 1$.

Substituting (4.60) and (4.61) into (4.58) and (4.59) respectively and taking the Kronecker delta to be tacit, we have

$$s_A(\mathbf{k}', \mathbf{k}) = |I_{\mathbf{k}'\mathbf{k}}|^2\, n_{\mathbf{q}} \frac{\pi C_{\mathbf{q}}^2}{N_C M \omega_{\mathbf{q}}} \tag{4.63}$$

and

$$s_E(\mathbf{k}', \mathbf{k}) = |I_{\mathbf{k}'\mathbf{k}}|^2 \left(n_{\mathbf{q}} + 1 \right) \frac{\pi C_{\mathbf{q}}^2}{N_C M \omega_{\mathbf{q}}}. \tag{4.64}$$

Acoustic Phonons

Acoustic phonons are lattice vibrations associated with the displacement of the primitive cells of a crystal. Close to $\mathbf{q} = 0$, the angular frequency of these modes varies linearly with wavevector and it proves a good approximation to put $\omega_{\mathbf{q}} = v_q q$, where v_q is the magnitude of the velocity of a mode averaged over direction. The low

frequencies (and hence energies) of these modes is then reflected in the name 'acoustic', since these are the phonon modes that carry audible sound in the material.

It can be shown via considerations of energy and momentum conservation [9] that intravalley scattering via acoustic phonons is limited to long-wavelength modes, for which the process may be approximated as elastic. Moreover, for temperatures much above 1 K, $\hbar\omega_{\mathbf{q}} \ll k_B T_0$ for this range of phonon energies, so that the Bose–Einstein factor for the mode occupation, given by (4.1) may be approximated by

$$n_{\mathbf{q}} + 1 \approx n_{\mathbf{q}} \approx \frac{k_B T_0}{\hbar\omega_{\mathbf{q}}}. \tag{4.65}$$

Note that T_0 is the lattice temperature. With these considerations, $s_A(\mathbf{k}', \mathbf{k}) = s_E(\mathbf{k}', \mathbf{k})$ and (4.57) reduces to

$$s(\mathbf{k}', \mathbf{k}) = \frac{2\pi |I_{\mathbf{k}'\mathbf{k}}|^2}{V_C} \frac{k_B T_0}{\hbar\rho v_q^2 q^2} C_{\mathbf{q}}^2 \delta(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}}) \tag{4.66}$$

where $\rho = N_C M / V_C$ is the mass density.

In this chapter, we shall only discuss deformation potential acoustic phonon scattering (see Ref. [9] for details on piezoelectric phonon scattering). The deformation potential tensor Ξ_{ij} gives the change in the band-edge due to applied strain. The elements of Ξ_{ij} can be reduced by symmetry to two components Ξ_d, for cubical dilation, and Ξ_u for shear strain. The coupling coefficient $C_{\mathbf{q}}$ is then given by [9]

$$C_{\mathbf{q}}^2 = \Xi^2(\theta_{\mathbf{q}}) q^2, \tag{4.67}$$

where $\Xi^2(\theta_{\mathbf{q}}) = \Xi_d$ for the Γ valley. For spheroidal valleys, $\theta_{\mathbf{q}}$ is the angle between $\mathbf{q}$ and the principal axis of the valley. $\Xi^2(\theta_{\mathbf{q}})$ is then decomposed into longitudinal and transverse components, denoted by the subscripts L and T respectively

$$\Xi_L(\theta_{\mathbf{q}}) = \Xi_d + \Xi_u \cos^2\theta_{\mathbf{q}} \tag{4.68}$$

and

$$\Xi_T(\theta_{\mathbf{q}}) = \Xi_u \sin\theta_{\mathbf{q}} \cos\theta_{\mathbf{q}}. \tag{4.69}$$

Substituting (4.67) into (4.66), we have

$$s(\mathbf{k}', \mathbf{k}) = \frac{2\pi |I_{\mathbf{k}'\mathbf{k}}|^2}{V_C} \frac{k_B T_0 \Xi^2(\theta_{\mathbf{q}})}{\hbar\rho v_q^2} \delta(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}}). \tag{4.70}$$

On insertion of (4.70) into (4.6), we find that the scattering rate for this process will be proportional to the density of states and, as such, a relevant process for momentum relaxation in high-field transport.

Optical Phonons

Whereas acoustic phonons are due to the displacements of the primitive cells in a crystal, high-frequency optical phonons are associated with the relative displacements of the basis atoms within the cells. The scattering strength of optical phonons is much greater in polar materials, where an oscillating electric dipole is set up between the basis atoms. A great simplification is obtained by making the standard assumption that the optical phonon energy $\hbar\omega_{\mathbf{q}}$ is a constant.

For non-polar optical phonons, the electron–phonon interaction energy has the general form $\mathbf{d}_o \cdot \mathbf{u}$, where $\mathbf{d}_o$ is some optical deformation potential and $\mathbf{u}$ is the relative displacement of the basis atoms. The coupling coefficient may then be equated as $C_{\mathbf{q}}^2 = d_o^2$, where d_o has been suitably scaled. The intrinsic rates for absorption and emission are then just given by (4.63) and (4.64) with this substitution. As in the case of deformation potential acoustic phonon scattering, this ultimately yields a scattering rate proportional to the density of states.

Non-polar optical phonon scattering is of interest in the group IV semiconductors, particularly in SiGe material systems. In this case, the acoustic and optical deformation potentials have been obtained from a combination of DFT calculations and fitting to experimental data [19], as well as from density functional perturbation theory (DFTP) based on a 'frozen phonon' approach [20, 21].

Polar Optical Phonons

In polar materials, an electric dipole is set up between the basis atoms. This causes strong electron scattering via the Fröhlich interaction [22]

$$V(\mathbf{r}) = -\frac{1}{\epsilon_0} \int \mathbf{D}(\mathbf{r}, \mathbf{R}) \cdot \mathbf{P}(\mathbf{R})\, d^3\mathbf{R}, \tag{4.71}$$

where $\mathbf{D}(\mathbf{r}, \mathbf{R})$ is the electric displacement at ionic position $\mathbf{R}$ due to the electron at $\mathbf{r}$ and $\mathbf{P}(\mathbf{R})$ is the polarization, given by

$$\mathbf{P}(\mathbf{R}) = \frac{e^* \mathbf{u}(\mathbf{R})}{\Omega}. \tag{4.72}$$

Here, e^* is the effective charge on the basis atoms, $\mathbf{u}(\mathbf{R})$ is the optical displacement and Ω is the primitive cell volume.

It is instructive to expand $\mathbf{u}(\mathbf{R})$ as a series of plane-waves with phonon wavevector $\mathbf{q}$

$$\mathbf{u}(\mathbf{R}) = N_C^{-1/2} \sum_{\mathbf{q}} u_{\mathbf{q}} \mathbf{e}_p e^{i\mathbf{q}\cdot\mathbf{R}}, \tag{4.73}$$

where $\mathbf{e}_p$ is the unit polarization vector for the phonon (strictly, this expansion should also include the complex conjugate of each term). In general, $\mathbf{e}_p$ may have longitu-

dinal (i.e. parallel to $\mathbf{q}$) and transverse components, which we can express by putting $\mathbf{e}_p = \alpha_L \mathbf{e}_L + \alpha_T \mathbf{e}_T$. Taking the curl of $\mathbf{u}(\mathbf{R})$, we find for each term in the expansion

$$\nabla \times \left(u_{\mathbf{q}} \mathbf{e}_p e^{i\mathbf{q}\cdot\mathbf{R}} \right) = i u_{\mathbf{q}} \left(\mathbf{q} \times \mathbf{e}_p \right) e^{i\mathbf{q}\cdot\mathbf{R}} = i u_{\mathbf{q}} q \alpha_T \mathbf{e}_T e^{i\mathbf{q}\cdot\mathbf{R}}. \tag{4.74}$$

Hence we may rewrite (4.73) as

$$\mathbf{u}(\mathbf{R}) = N_C^{-1/2} \sum_{\mathbf{q}} u_{\mathbf{q}} \left[\alpha_L \mathbf{e}_L e^{i\mathbf{q}\cdot\mathbf{R}} - \frac{i}{q} \nabla \times \left(\mathbf{e}_p e^{i\mathbf{q}\cdot\mathbf{R}} \right) \right]. \tag{4.75}$$

Now the divergence of $\mathbf{P}(\mathbf{R})$ gives the polarization charge density. However, from (4.72) we see that this is proportional to the divergence of $\mathbf{u}(\mathbf{R})$. Since for any vector field $\mathbf{u}$, $\nabla \cdot (\nabla \times \mathbf{u})$ is identically zero, taking the divergence of (4.75) gives non-zero contributions only for the longitudinal terms. Hence, only longitudinal optical (LO) phonons couple with electrons in this type of scattering.

The electric displacement, including a simple model of screening, is given by

$$\mathbf{D}(\mathbf{r}, \mathbf{R}) = \nabla \left(\frac{e}{4\pi \left| \mathbf{r} - \mathbf{R} \right|} e^{-q_0 |\mathbf{r} - \mathbf{R}|} \right). \tag{4.76}$$

Using this together with (4.71)–(4.73) enables us to derive an expression for the coupling coefficient $C_{\mathbf{q}}$. After analysis of the effective charge e^* via consideration of the equation of motion for an LO phonon mode, $C_{\mathbf{q}}$ is found to be given by [9]

$$C_{\mathbf{q}}^2 = \frac{e^2 M \omega_{\mathbf{q}}^2}{\Omega \varepsilon_p} \frac{q^2}{\left(q^2 + q_0^2 \right)^2}. \tag{4.77}$$

Here, M is the reduced mass and we have defined

$$\frac{1}{\varepsilon_p} = \frac{1}{\varepsilon_0} \left(\frac{1}{\kappa_\infty} - \frac{1}{\kappa_0} \right) \tag{4.78}$$

in terms of the permittivity of free space ε_0 and the high and low frequency dielectric constants κ_∞ and κ_0 respectively.

Substituting (4.77) into (4.63) and (4.64) gives

$$s(\mathbf{k}', \mathbf{k}) = \frac{(2\pi)^2 \left| I_{\mathbf{k}'\mathbf{k}} \right|^2}{V_C} \frac{\hbar W_0}{q^{*2}} \left(\frac{\hbar \omega_{\mathbf{q}}}{2m^*} \right)^{1/2} \left(n_{\mathbf{q}} + 1/2 \mp 1/2 \right) \delta \left(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}} \pm \hbar \omega_{\mathbf{q}} \right), \tag{4.79}$$

where $q^* = q \left(1 + (q_0/q)^2 \right)$ and we have defined a characteristic rate for polar optical phonon scattering

$$W_0 = \frac{e^2}{4\pi\hbar\varepsilon_p}\left(\frac{2m^*\omega_{\mathbf{q}}}{\hbar}\right)^{1/2}. \tag{4.80}$$

Note that the $n_{\mathbf{q}}$ appearing in (4.79) is the mode occupation for polar optical phonons and that this is a function of temperature. Very often, this temperature is taken to be the lattice temperature, although it is possible to have a population of optical phonons that are out of thermal equilibrium with the lattice. The topic of 'hot phonons' is beyond the scope of this chapter but the interested reader may pursue the subject in Refs. [23, 24].

As we shall discover in Sect. 1.5.3, no unique relaxation time may be found for polar optical phonon scattering. However, reasonable approximations have been found for the energy and momentum relaxation rates and have been given by Conwell and Vassell [8] (see also Ref. [9]). For momentum relaxation, a composite relaxation time can be found accurately using the ladder method, which we discuss in Sect. 1.5.3.

4.4 High-Field Phenomena

4.4.1 Impact Ionisation and Avalanche Breakdown

Impact Ionisation Coefficients

One of the earliest motivations for studying high-field transport was to gain an understanding of electrical breakdown in materials. An important process contributing to this phenomenon is that of impact ionisation and the consequent effect, under certain conditions, of avalanche breakdown. Less disastrously, impact ionisation may be exploited to provide gain in the avalanche photodiode (APD) via avalanche multiplication of charge carriers.

We define the probability that a given carrier will ionize an electron-hole pair in distance dx by αdx (for electrons) or βdx (for holes), where α and β are the ionisation coefficients for electrons and holes respectively. These quantities may be interpreted as spatial rates (i.e. they have dimensions of reciprocal distance). Alternatively, we can think of the quantities $1/\alpha$ and $1/\beta$ as the average distance traveled between ionizing collisions for the respective carrier type.

Clearly, the ionisation coefficients will depend on the electric field, although the exact dependence is difficult to analyze. In an early model of the field dependence, Wolff [25] derived an expression of the form $\alpha \sim \exp(-a/E^2)$ based on a simple band structure and a population of equilibrium electrons. For hot electrons, we might think of this in terms of thermalization, in which electrons exchange energy and momentum with each other to form an equilibrium-like distribution.

An alternative form was proposed by Shockley [26], who argued that only those 'lucky' carriers that had managed to avoid collisions and gain the required threshold energy ϵ_I could impact ionize. Hence the ionisation coefficients should be proportional to a factor $\exp(-\epsilon_I/eE\lambda)$, where λ is the mean-free path of the carriers.

In Shockley's model, the time between collisions corresponds to a momentum relaxation time, so that at times less than this, the electron is travelling ballistically. It was pointed out by Ridley [27] that an intermediate state existed between this state of ballistic motion and Wolff's regime of complete thermalization, due to the disparity between the energy and momentum relaxation times, τ_ϵ and $\tau_{\mathbf{k}}$ respectively. If $\tau_\epsilon \gg \tau_{\mathbf{k}}$, a carrier may spend significant time in a state of drift, undergoing momentum relaxing events without energy relaxation. Some of these carriers may then reach the threshold energy despite having undergone elastic scattering. Ridley termed this condition 'lucky-drift'. Using similar terminology, Shockley's model may be termed 'lucky-ballistic' or 'lucky-flight'.

Incorporating the different possible mechanisms by which an electron may reach the threshold for impact ionisation, Ridley's expression for the ionisation coefficient is then

$$\alpha = \frac{eE}{\epsilon_I}\left[P_F^0 + P_D^0 + P_T\left(P_F^T + P_D^T\right)\right], \tag{4.81}$$

where ϵ_I/eE is the path length for an electron to reach ϵ_I and the P terms are the probabilities for the various processes. The F and D subscripts are for lucky-flight and lucky-drift respectively, whilst the 0 and T superscripts denote acceleration from zero energy and acceleration from the average thermalized energy respectively. P_T is then the probability that the electron will thermalize to the hot electron distribution.

P_F^0 is essentially Shockley's result

$$P_F^0\left(\epsilon_I\right) = \exp\left(-\int_0^{\epsilon_I} \frac{d\epsilon_{\mathbf{k}}}{eE\lambda(\epsilon_{\mathbf{k}})}\right), \tag{4.82}$$

with the mean free path allowed to vary with energy. The probability for lucky drift from zero energy is

$$P_D^0\left(\epsilon_I\right) = \int_0^{\epsilon_I} P_F^0\left(\epsilon_{\mathbf{k}}\right) P_D^0\left(\epsilon_{\mathbf{k}}, \epsilon_I\right) \frac{d\epsilon_{\mathbf{k}}}{eE\lambda(\epsilon_{\mathbf{k}})}, \tag{4.83}$$

where

$$P_D^0\left(\epsilon_{\mathbf{k}}, \epsilon_I\right) = \exp\left(-\int_{\epsilon_{\mathbf{k}}}^{\epsilon_I} \frac{m^*\left(d\gamma/d\epsilon_{\mathbf{k}}\right) d\epsilon_{\mathbf{k}}}{e^2E^2\tau_{\mathbf{k}}(\epsilon_{\mathbf{k}})\tau_\epsilon(\epsilon_{\mathbf{k}})}\right). \tag{4.84}$$

The probabilities P_F^T and P_D^T are obtained from (4.82) and (4.83) by replacing the lower limit of the integrals with ϵ_T, the energy of the thermalized electron. Finally, for the probability of thermalization, Ridley proposed

$$P_T = 1 - \exp\left(-\frac{(\epsilon/eE) - 3\overline{\lambda}}{\overline{(v_D\tau_\epsilon)}}\right), \tag{4.85}$$

where the bar notation denotes averaging.

Avalanche Breakdown

Having discussed the field dependence of the ionisation coefficients, we turn our attention to the dynamics of a population of carriers undergoing impact ionisation. Consider a region in a material of width δx over which an electric field $\mathbf{E}$ is applied, pointed in the negative x direction, so that electrons travel to the right and holes to the left. Now the total current in the system is the sum of electron and hole currents

$$j = -e(n_e v_e + n_h v_h), \tag{4.86}$$

where n_e and n_h are the densities of electrons and holes respectively and v_e and v_h are the magnitudes of the drift velocities for each type of carrier. The minus sign just indicates that the current, in this case, is in the negative x direction.

Now the number of electrons leaving this region at $x + \delta x$ per unit time will be equal to the electron flux into it at x plus the flow of electrons generated within it via impact ionisation for each type of carrier. For a sufficiently small area, we will then have

$$\begin{aligned} n_e(x + \delta x)v_e = n_e(x)v_e \\ + (\alpha n_e(x)v_e + \beta n_h(x + \delta x)v_h)\,\delta x, \end{aligned} \tag{4.87}$$

with a similar expression for holes

$$\begin{aligned} n_h(x)v_h = n_h(x + \delta x)v_h \\ + (\alpha n_e(x)v_e + \beta n_h(x + \delta x)v_h)\,\delta x. \end{aligned} \tag{4.88}$$

Multiplying (4.87) and (4.88) by the magnitude of the electronic charge e, rearranging and taking the limit $\delta x \to 0$, we obtain the simultaneous differential equations

$$\begin{aligned} \frac{dj_e}{dx} &= \alpha j_e + \beta j_h, \\ \frac{dj_h}{dx} &= -\alpha j_e - \beta j_h. \end{aligned} \tag{4.89}$$

Solving these equations with boundary conditions

$$\begin{aligned} j_e(0) &= J_0, \\ j_e(\Delta x) &= J, \\ j_h(\Delta x) &= 0, \end{aligned} \tag{4.90}$$

gives us the particular solution

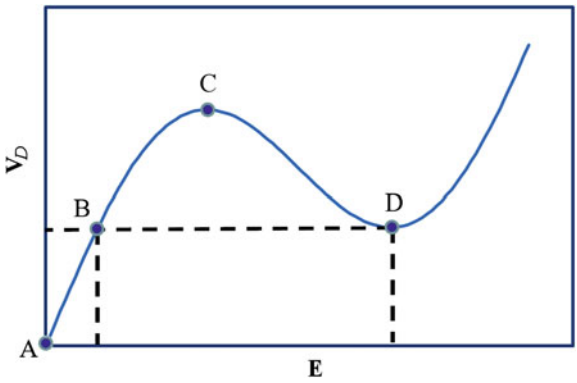


Fig. 4.3 Schematic example of negative differential resistance ($\mathrm{d}V/\mathrm{d}I$), showing drift velocity (proportional to current I) against the magnitude of electric field (proportional to voltage V). Between points A and C, $\mathrm{d}V/\mathrm{d}I$ is positive, negative between points C and D, then positive again thereafter

$$\frac{J}{J_0} = \frac{(\alpha - \beta)\, e^{(\alpha-\beta)\Delta x}}{\alpha - \beta e^{(\alpha-\beta)\Delta x}}. \tag{4.91}$$

Now, when $\alpha = \beta e^{(\alpha-\beta)\Delta x}$, the current J becomes infinite, implying some runaway condition. This situation is known as avalanche breakdown.

Avalanche Photodiode Gain

In an APD, the photocurrent generated from detected light may be amplified via the process of impact ionisation. The boundary conditions imposed earlier on (4.89) describe the situation of electron injection into a multiplication region. Looking back at the gain figure in (4.91), we see that if $\beta = 0$, then the gain is just a simple exponential $J/J_0 = \exp(\alpha \Delta x)$. Clearly this is a more stable situation and provides one reason why it is desirable to have a large disparity between α and β in the materials used to fabricate APDs. Other reasons include keeping the response time of the device low [28] and reducing the excess noise factor [28, 29].

4.4.2 Negative Differential Resistance

One of the more interesting (and useful) nonlinear phenomena in high-field transport is that of negative differential resistance. At low field, conduction generally tends to be ohmic, following the relation $V = IR$ for the applied voltage, V, current I and resistance R. Ohm's law is, of course, not an actual law of nature but rather a general rule of thumb. More generally, we may take the derivative $\mathrm{d}V/\mathrm{d}I = R$, so that R may now be referred to as the differential resistance. Somewhat counter-intuitively, we find that under certain circumstances R may become negative.

A particular example is shown in Fig. 4.3. The graph is labelled as drift velocity, $\mathbf{v}_D$ against (the magnitude) of the applied field $\mathbf{E}$. However, with only a change of scale (unnecessary here as the scales are in arbitrary units), we could have labelled the axes with current I along the ordinate and voltage V along the abscissa. Between point A (zero voltage, zero current) and C, we have the familiar situation of a positive differential resistance. However, between C and D, R becomes negative (this would have been more obvious if we had plotted V along the ordinate and I along the abscissa). Such a situation may sometimes give rise to oscillations in the current, as we shall see. However, before discussion of that, we shall consider how an NDR may arise in the first place.

Transferred Electrons

Several mechanisms for the occurrence of an NDR are known. One of the best known is that due to transferred electrons [30, 31], in which hot electrons in an initial valley are scattered into a higher lying valley with a larger effective mass. At low fields most of the electrons will be in the smaller effective mass valley. As the electron temperature is increased with the application of a higher electric field, an increasing number of electrons become resonant with higher lying valleys and intervalley scattering into these will start to become significant. The electrons in these higher valleys see a larger effective mass and so have a lower mobility than those in the lower valley. As the population of these slower moving electrons builds up, the overall mobility of the entire electron population may then begin to decrease, leading to a reduction of current and, hence, a negative differential resistance.

Other Mechanisms for NDR

Intervalley scattering into valleys with a larger effective mass is not a necessary condition for NDR. In some cases, the non-parabolicity within a single band may be sufficient to bring about a reduction in mobility as the electrons are driven to higher energies, as discussed briefly in Sect. 4.2.7. NDR due to non-parabolicity is often considered phenomenologically in terms of the larger effective mass seen by an electron above the conduction band edge but it is more accurate to describe the phenomenon in terms of the closely related group velocity.

Consider the dispersion relations of a band as one moves out from the wavevector at the band-edge towards the Brillouin zone boundary. Initially, at the band-edge, the group velocity $\mathbf{v}(\mathbf{k}) = 0$, then begins to increase with increasing wavevector. However, near the zone boundary, $\epsilon_{\mathbf{k}}$ will begin to flatten off again, as demanded by the periodicity of the Brillouin zone, and the group velocity will again become zero. Hence, at some point in between, $\epsilon_{\mathbf{k}}$ must go through an inflexion point, after which the group velocity decreases with increasing energy. If a sufficient proportion of electrons lies in this range, it is possible to see the emergence of an NDR.

A similar situation appears to arise in the case of the dilute nitrides, in which dilute concentrations of nitrogen are substituted into arsenic sites in GaInAs. The nitrogen atoms form localised states resonant with the conduction band that is believed to split the conduction band via an anti-crossing [32]. According to this band anti-crossing (BAC) model, the lower subvalley becomes highly non-parabolic and flattens off even at relatively low energies quite far from the zone boundaries. An NDR has in fact been observed in dilute GaAs:N by Patanè et al. [33], who interpret their results according to the BAC model.

Yet another example of NDR occurs in resonant tunnelling. However, in this case, the phenomenon arises due to quantum mechanical tunnelling, which is not easily dealt with using the semiclassical approach of this chapter.

Charge Fluctuations

It was noted by Ridley and Watkins [30] that the presence of an NDR could lead to instabilities due to charge fluctuations and the emergence of travelling electrical domains. Any localised fluctuation in the charge density will cause a space-charge potential to be imposed onto the potential due to the electric field and will be moving with the drift velocity of the carriers. At the trailing edge of the space-charge profile the field will be reduced, whilst at the leading edge, the field is increased. Under the normal conditions of a positive differential resistance, the carriers at the trailing edge will be slowed down whilst those at the leading edge will be accelerated. Hence, the fluctuation will tend to be pulled apart and smoothed out.

On the other hand, if a negative differential resistance pertains, the opposite will happen. Carriers at the trailing edge will be sped up, whilst those at the leading edge will slow down with the net effect of causing the carriers to bunch together. This will continue until the carriers at either edge of the fluctuation obtain the same drift velocity. It is not immediately obvious under what conditions this will occur as there may be many pairs of field strengths for which the group velocity will be equal. However, it has been argued by Ridley [34] on the basis of thermodynamical considerations that for an NDR of the form shown in Fig. 4.3, this will occur at the points B and D on the graph.

Current Oscillations

Consider the existence of such travelling electric domains in a device of length L. If the device is short enough, it is likely that there will only be one such domain in the device at a time. As this domain leaves the device at the positive terminal, charge conservation within the device will lead to another being nucleated at the negative terminal. Hence, we will have a regular series of current pulses travelling through the device with an approximate frequency of v_D/L. The process by which this occurs, given the transferred electron effect as the cause of the NDR is known as the Ridley–Watkins–Hilsum mechanism. The oscillations were subsequently observed by Gunn

[2] and the phenomenon has become better known as the Gunn or Gunn-Hilsum effect, with the domains often being referred to as Gunn domains. This is, of course, the physics at the heart of the Gunn diode for generating microwave oscillations.

4.5 The Boltzmann Transport Equation

4.5.1 General Form of the BTE

The Boltzmann transport equation (BTE) governs the statistical distribution of particles under non-equilibrium conditions. For our purposes, it therefore determines the dynamics of the distribution function $f(\mathbf{k})$. In its most general form, the BTE may be written as [35]

$$\frac{\mathrm{d}f(\mathbf{k})}{\mathrm{d}t} = \left(\frac{\partial f(\mathbf{k})}{\partial t}\right)_s - \frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}) - \mathbf{v}(\mathbf{k}) \cdot \nabla f(\mathbf{k}). \tag{4.92}$$

The first term on the right-hand side of (4.92) is the temporal rate of change of $f(\mathbf{k})$ due to scattering and can be written

$$\left(\frac{\partial f(\mathbf{k})}{\partial t}\right)_s = \int s(\mathbf{k}', \mathbf{k}) f(\mathbf{k}') [1 - f(\mathbf{k})] - s(\mathbf{k}, \mathbf{k}') f(\mathbf{k}) \left[1 - f(\mathbf{k}')\right] \frac{V_C}{(2\pi)^3} d^3\mathbf{k}'. \tag{4.93}$$

Note that the rates (per unit $\mathbf{k}$-space) inside the integral are multiplied not only by the probability that the initial state is occupied but also, since electrons are fermions, the probability that the final state is *not* occupied.

The last term on the right-hand-side of (4.92) involves the spatial variation of $f(\mathbf{k})$, which may be due to a temperature or carrier density gradient. In this current work, we set this to zero.

Writing the acceleration of a state in terms of a driving force $\mathbf{F}$ as $d\mathbf{k}/dt = \mathbf{F}/\hbar$, in the steadystate we have

$$\frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}) = \left(\frac{\partial f(\mathbf{k})}{\partial t}\right)_s. \tag{4.94}$$

In what follows, we shall assume no magnetic field so that the force is just that due to an applied electric field $\mathbf{F} = -e\mathbf{E}$.

Before embarking on an attempt to solve (4.94) under high-field conditions, it will be useful to underpin our understanding by working through a low-field solution based on the linearization of $f(\mathbf{k})$. One reason for this is that it will caution us to the limitations of using relaxation times in our formulations, since as we shall see, for polar optical phonon scattering no unique time exists.

Since many of the expressions we shall encounter in this chapter are quite long, as a notational short-hand we shall drop the **k** and **q** subscripts on $\epsilon_{\mathbf{k}}$ and $\omega_{\mathbf{q}}$, as this should not result in any ambiguity.

4.5.2 The Linearized Distribution Function

In equilibrium, the distribution function is just the Fermi factor, given by (4.2). If the distribution function is now displaced in **k**-space by an average drift wavevector $\delta\mathbf{k}$ due to **E**, then to first order we can expand $f(\mathbf{k})$ as

$$f(\mathbf{k}) = f_0(\mathbf{k} - \delta\mathbf{k}) = f_0(\epsilon) - \nabla_{\mathbf{k}} f_0(\epsilon) \cdot \delta\mathbf{k}. \tag{4.95}$$

Since the group velocity is given by $\mathbf{v}(\mathbf{k}) = \nabla_{\mathbf{k}}\epsilon/\hbar$, application of the chain rule gives

$$f(\mathbf{k}) = f_0(\epsilon) - \hbar\frac{\mathrm{d}f_0(\epsilon)}{\mathrm{d}\epsilon}\mathbf{v}(\mathbf{k}) \cdot \delta\mathbf{k}. \tag{4.96}$$

For an electron accelerating in an electric field in the presence of scattering processes, we can put

$$\delta\mathbf{k} = -\frac{e\mathbf{E}\tau(\epsilon)}{\hbar}, \tag{4.97}$$

where $\tau(\epsilon)$ is the energy dependent relaxation time. Hence, we have

$$f(\mathbf{k}) = f_0(\epsilon) + e\tau(\epsilon)\frac{\mathrm{d}f_0(\epsilon)}{\mathrm{d}\epsilon}\mathbf{v}(\mathbf{k}) \cdot \mathbf{E}. \tag{4.98}$$

It will be convenient to denote the asymmetric contribution to $f(\mathbf{k})$ by

$$f_1(\epsilon) = e\tau(\epsilon)\frac{\mathrm{d}f_0(\epsilon)}{\mathrm{d}\epsilon}Ev(\epsilon)x, \tag{4.99}$$

where the magnitude of the group velocity $v(\epsilon)$ only depends on energy in a spherical band and x is the cosine of the angle θ between $\mathbf{v}(\mathbf{k})$ and **E**. Note that we shall assume throughout that $\mathbf{v}(\mathbf{k})$ and **k** are parallel.

Low-Field Transport Properties

Limiting ourselves to a single, spherical energy band, using (4.95) and (4.99) with (4.27) and (4.28), the conductivity tensor becomes the scalar

$$\sigma = -\frac{2e^2}{3\,(2\pi)^3}\int v^2(\epsilon_{\mathbf{k}})\tau(\epsilon_{\mathbf{k}})\frac{\mathrm{d}f_0(\epsilon_{\mathbf{k}})}{\mathrm{d}\epsilon}\,d^3\mathbf{k}. \tag{4.100}$$

Here, the factor involving the symmetric part of $f(\mathbf{k})$, $f_0(\epsilon)$ disappears due to the odd parity of $\mathbf{v}(\mathbf{k})$, as do the cross-products involving different Cartesian coordinates of $\mathbf{v}(\mathbf{k})$. The factor of $1/3$ emerges due to the average value $v_i^2 = v^2/3$ of the squared components of $\mathbf{v}(\mathbf{k})$ that remain.

Converting to an integral over energy using (4.26), we have

$$\sigma = -\frac{2e^2}{3V_C}\int v^2(\epsilon)\tau(\epsilon)\frac{\mathrm{d}f_0(\epsilon)}{\mathrm{d}\epsilon}D(\epsilon)\,d\epsilon. \tag{4.101}$$

Now, the group velocity may be written in terms of $\gamma(\epsilon)$ according to (4.23), so only $\tau(\epsilon)$ is unknown at this stage. This is the quantity that we must compute from the BTE in the low-field solution.

4.5.3 Low Field Solution and the Ladder Method

Since $f_1(\epsilon)$ is proportional to $\mathbf{E}$, taking the dot product of this contribution with $\mathbf{E}$ on the left-hand-side of (4.94) gives a term proportional to E^2, which in the present low-field solution we take to be negligible. Hence, the first order corrections are neglected entirely from this side of (4.94) and we are left with

$$-\frac{e}{\hbar}\mathbf{E}\cdot\nabla_{\mathbf{k}}f(\mathbf{k}) = -e\mathbf{E}\cdot\mathbf{v}(\mathbf{k})\frac{\mathrm{d}f_0(\epsilon)}{\mathrm{d}\epsilon} = -eEv(\epsilon)x\frac{\mathrm{d}f_0(\epsilon)}{\mathrm{d}\epsilon}. \tag{4.102}$$

Turning our attention to the other side of (4.94), we now substitute $f(\mathbf{k}) = f_0(\epsilon) + f_1(\epsilon)$ into the scattering integral (4.93). Neglecting products of the first order components of $f(\mathbf{k})$, the scattering integral may be formally decomposed into

$$\left(\frac{\partial f(\mathbf{k})}{\partial t}\right)_s = \left(\frac{\partial f_0(\epsilon)}{\partial t}\right)_s + \left(\frac{\partial f_1(\epsilon)}{\partial t}\right)_s, \tag{4.103}$$

where

$$\left(\frac{\partial f_0(\epsilon)}{\partial t}\right)_s = \int s(\mathbf{k}',\mathbf{k})f_0(\epsilon')\left[1-f_0(\epsilon)\right] - s(\mathbf{k},\mathbf{k}')f_0(\epsilon)\left[1-f_0(\epsilon')\right]\frac{V_C}{(2\pi)^3}d^3\mathbf{k}' \tag{4.104}$$

and

$$\begin{aligned}\left(\frac{\partial f_1(\epsilon)}{\partial t}\right)_s = &\int s(\mathbf{k}',\mathbf{k})\left\{f_1(\epsilon')\left[1-f_0(\epsilon)\right] - f_1(\epsilon)f_0(\epsilon')\right\}\\ &- s(\mathbf{k},\mathbf{k}')\left\{f_1(\epsilon)\left[1-f_0(\epsilon')\right] - f_1(\epsilon')f_0(\epsilon)\right\}\frac{V_C}{(2\pi)^3}d^3\mathbf{k}'.\end{aligned} \tag{4.105}$$

Using the equilibrium condition that for the Fermi factor, $f_0(\epsilon)$, the zero order contribution given by (4.104) disappears, the scattering integral reduces to

$$\left(\frac{\partial f(\mathbf{k})}{\partial t}\right)_s = \left(\frac{\partial f_1(\epsilon)}{\partial t}\right)_s = \int s(\mathbf{k}', \mathbf{k}) \left\{ f_1(\epsilon') \frac{1 - f_0(\epsilon)}{1 - f_0(\epsilon')} - f_1(\epsilon) \frac{f_0(\epsilon')}{f_0(\epsilon)} \right\} \frac{V_C}{(2\pi)^3} d^3\mathbf{k}'. \quad (4.106)$$

Substituting for $f_1(\epsilon)$ from (4.99) and using (4.3) for the derivative of the Fermi factor to eliminate the factor of $1/\left(1 - f_0(\epsilon')\right)$ then gives

$$\left(\frac{\partial f(\mathbf{k})}{\partial t}\right)_s = -e \frac{df_0(\epsilon)}{d\epsilon} E v(\epsilon) x \times \int s(\mathbf{k}', \mathbf{k}) \frac{f_0(\epsilon')}{f_0(\epsilon)} \left\{ \tau(\epsilon) - \tau(\epsilon') \frac{v(\epsilon') x' d}{v(\epsilon) x} \right\} \frac{V_C}{(2\pi)^3} d^3\mathbf{k}'. \quad (4.107)$$

The ratio of cosines x'/x may be dealt with by choosing coordinates such that, say, the $\mathbf{k}$ vector lies along the z' axis and $\mathbf{E}$ lies in the $x - z$ plane. The dot product of the unit vectors $\hat{\mathbf{E}}$ and $\hat{\mathbf{k}}'$ is then

$$\hat{\mathbf{E}} \cdot \hat{\mathbf{k}}' = \cos\theta' = \cos\theta \cos\alpha' + \sin\theta \sin\alpha' \cos\phi, \quad (4.108)$$

where α' is the angle between $\hat{\mathbf{k}}'$ and $\hat{\mathbf{k}}$ and ϕ is the azimuthal angle. Since any angular dependence can only come from $s(\mathbf{k}', \mathbf{k})$, which depends, at most, on α', integrating $\cos\phi$ over 2π gives zero. So, effectively, $x'/x = \cos\alpha'$. Inserting this result into (4.107) and equating with the expression for the force term given by (4.102), we arrive at

$$\int s(\mathbf{k}', \mathbf{k}) \frac{f_0(\epsilon')}{f_0(\epsilon)} \left\{ \tau(\epsilon) - \tau(\epsilon') \frac{v(\epsilon')}{v(\epsilon)} \cos\alpha' \right\} \frac{V_C}{(2\pi)^3} d^3\mathbf{k}' = 1. \quad (4.109)$$

As we shall shortly see, for elastic processes at least, this expression enables us to derive a well-defined result for the relaxation time $\tau(\epsilon)$. Before that, we pause to note that, substituting this result back into (4.107), from (4.106) we have

$$\left(\frac{\partial f_1(\epsilon)}{\partial t}\right)_s = -e \frac{df_0(\epsilon)}{d\epsilon} E v(\epsilon) x = -\frac{f_1(\epsilon)}{\tau(\epsilon)}. \quad (4.110)$$

Although this is a low-field result that is only strictly true for elastic scattering mechanisms, it will prove a useful approximation in our later treatment of high-field solutions.

Elastic Scattering Processes

Let us consider purely elastic processes characterized by $s_e(\mathbf{k}', \mathbf{k})$, so that

$$s(\mathbf{k}', \mathbf{k}) = s_e(\mathbf{k}', \mathbf{k})\delta\left(\epsilon' - \epsilon\right). \tag{4.111}$$

On insertion of (4.111) into (4.109), the action of the delta function means that $\tau(\epsilon)$ takes the same value of ϵ whilst the $f_0(\epsilon)$ and $v(\epsilon)$ cancel out, leaving

$$\tau(\epsilon)\int s_e(\mathbf{k}', \mathbf{k})\left\{1 - \cos\alpha'\right\}\delta\left(\epsilon' - \epsilon\right)\frac{V_C}{(2\pi)^3}d^3\mathbf{k}' = 1. \tag{4.112}$$

Since α' is the angle between $\mathbf{k}'$ and $\mathbf{k}$, the integral in (4.112) gives the definition of an elastic momentum relaxation time $\tau_{\mathbf{k}}(\epsilon)$, defined earlier in (4.6), as the reciprocal of the elastic scattering rate $w_e(\epsilon)$ for a given process:

$$\int s_e(\mathbf{k}', \mathbf{k})\left\{1 - \cos\alpha'\right\}\delta\left(\epsilon' - \epsilon\right)\frac{V_C}{(2\pi)^3}d^3\mathbf{k}' = w_e\left(\epsilon\right) = \frac{1}{\tau_{\mathbf{k}}\left(\epsilon\right)}. \tag{4.113}$$

Note that since $s(\mathbf{k}', \mathbf{k})$ is given by (4.4), for isotropic processes the elastic scattering rate $w_e\left(\epsilon\right)$ may be expressed as

$$\begin{aligned} w_e\left(\epsilon\right) &= \frac{2\pi}{\hbar}\left|\langle\mathbf{k}|V|\mathbf{k}'\rangle\right|^2\int\delta\left(\epsilon' - \epsilon\right)\frac{V_C}{(2\pi)^3}d^3\mathbf{k}', \\ &= \frac{2\pi}{\hbar}\left|\langle\mathbf{k}|V|\mathbf{k}'\rangle\right|^2 D(\epsilon), \end{aligned} \tag{4.114}$$

which is the most familiar form of Fermi's Golden Rule.

Inelastic Scattering

The principal mechanism of energy relaxation for an electron is via phonon scattering. Strictly speaking, all such interactions are inelastic, however in low-field calculations acoustic phonon scattering is often approximated as being energy conserving. The same is not true for optical phonon scattering and the situation becomes more complicated. In fact, we shall find that for polar optical phonon scattering no unique relaxation time can be found.

Consider the case of scattering due to an optical mode of energy $\hbar\omega$. The intrinsic rate is given formally by (4.57). When this is inserted into (4.109), the effect of the delta functions is to introduce values of the $\tau(\epsilon)$ at energies $\epsilon \pm \hbar\omega$, which can then be taken outside the integrals. The result is an expression of the form

$$A(\epsilon)\tau(\epsilon - \hbar\omega) + B(\epsilon)\tau(\epsilon) + C(\epsilon)\tau(\epsilon + \hbar\omega) = 1, \tag{4.115}$$

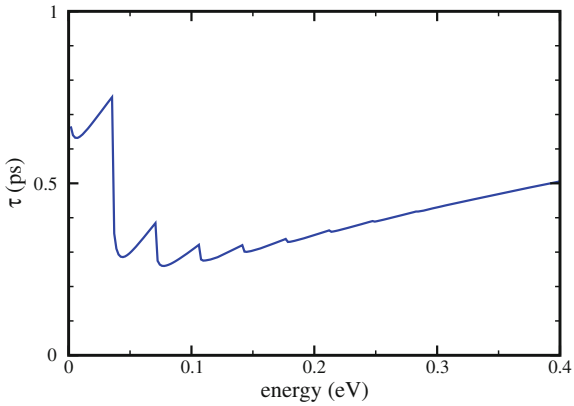


Fig. 4.4 Calculation of the room-temperature relaxation times for polar optical scattering in GaAs using the ladder method (see Ref. [37] for details). Note the characteristic saw tooth energy dependence, particularly at low energies, with discontinuities at intervals of the phonon energy ($\hbar\omega_{\mathbf{q}} = 35\,\mathrm{meV}$)

where

$$
\begin{aligned}
A(\epsilon) &= -\Theta(\epsilon - \hbar\omega) \int s_A(\mathbf{k}', \mathbf{k}) \frac{f_0(\epsilon')}{f_0(\epsilon)} \frac{v'}{v} \cos\alpha' \, \delta(\epsilon' - \epsilon + \hbar\omega) \, \frac{V_C}{(2\pi)^3} d^3\mathbf{k}', \\
B(\epsilon) &= \Theta(\epsilon - \hbar\omega) \int s_A(\mathbf{k}', \mathbf{k}) \frac{f_0(\epsilon')}{f_0(\epsilon)} \delta(\epsilon' - \epsilon + \hbar\omega) \, \frac{V_C}{(2\pi)^3} d^3\mathbf{k}' \\
&\quad + \int s_E(\mathbf{k}', \mathbf{k}) \frac{f_o(\epsilon')}{f_0(\epsilon)} \delta(\epsilon' - \epsilon - \hbar\omega) \, \frac{V_C}{(2\pi)^3} d^3\mathbf{k}', \\
C(\epsilon) &= -\int s_E(\mathbf{k}', \mathbf{k}) \frac{f_o(\epsilon')}{f_0(\epsilon)} \frac{v'}{v} \cos\alpha' \, \delta(\epsilon' - \epsilon - \hbar\omega) \, \frac{V_C}{(2\pi)^3} d^3\mathbf{k}'.
\end{aligned}
\tag{4.116}
$$

Here, we have introduced the step function

$$
\Theta(\epsilon) = \begin{cases} 0, & \epsilon \leq 0, \\ 1, & \epsilon > 0, \end{cases} \tag{4.117}
$$

since an electron with a final energy $\epsilon < \hbar\omega$ could not have absorbed a phonon. Hence the scattering rate for this process must be zero.

It is clear from (4.115) that, unless the coefficients $A(\epsilon)$ and $C(\epsilon)$ disappear, the relaxation time at ϵ will be coupled with the times at $\epsilon \pm \hbar\omega$. This suggests the picture of a phonon energy 'ladder' with rungs $\hbar\omega$ apart: the scattering rate on any particular rung being related to the rates on adjacent rungs.

The notation can be made a little more concise by writing the energy as $\varepsilon + j\hbar\omega$, where $0 \leq \varepsilon < \hbar\omega$ and $j \in \{0, 1, \ldots\}$, and denoting all functions of energy $G(\varepsilon + j\hbar\omega)$ by G_j. Equation (4.115) can then be written out in matrix form as

$$\begin{bmatrix} B_0 & C_0 & 0 & 0 & 0 \cdots \\ A_1 & B_1 & C_1 & 0 & 0 \cdots \\ 0 & A_2 & B_2 & C_2 & 0 \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \ddots \end{bmatrix} \begin{bmatrix} \tau_0 \\ \tau_1 \\ \tau_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix}. \tag{4.118}$$

Equation (4.118) makes it clear that if we require any τ_j at a particular level ϵ, we need to solve for all the values of τ_j separated by integral multiples of $\hbar\omega$. In order to minimize any truncation error incurred from only solving for a finite number of rungs, it is assumed that as $j \to \infty$, $\tau_j \to \tau_{j+1}$. Hence, for the last rung N, we put $B_N \to B_N + C_N$. To include elastic scattering processes, we just make the substitution

$$B(\epsilon) \to B(\epsilon) + \sum_i w_{e,i}(\epsilon). \tag{4.119}$$

It turns out that for non-polar optical scattering, the $A(\epsilon)$ and $C(\epsilon)$ coefficients do disappear but not for polar optical phonons. The procedure for determining the relaxation times described above is known as the ladder method and its development involves the determination of the ladder coefficients $A(\epsilon)$, $B(\epsilon)$ and $C(\epsilon)$. Fletcher and Butcher [36] give a good introduction to the method for bulk material in the presence of a magnetic field, assuming parabolic energy bands. A more detailed account, covering the generalization to non-parabolic bands as well as scattering in 2D structures may be found in Ref. [37]. Here we just quote the results for bulk:

$$A(\epsilon) = -\Theta(\epsilon_-) I^2_{\mathbf{k}'\mathbf{k}} W_0 \frac{\pi^2\hbar^3}{m^*} \left(\frac{\hbar\omega}{2m^*}\right)^{1/2} n_{\mathbf{q}} \frac{f_0(\epsilon_-)}{f_0(\epsilon)} \frac{v(\epsilon_-)}{v(\epsilon)} \times \frac{D(\epsilon_-)}{\gamma^{1/2}(\epsilon_-)\gamma^{1/2}(\epsilon)} \left[\frac{\gamma(\epsilon_-) + \gamma(\epsilon)}{\gamma^{1/2}(\epsilon_-)\gamma^{1/2}(\epsilon)} \tanh^{-1} \left(\frac{\gamma(\epsilon_-)}{\gamma(\epsilon)}\right)^{1/2} - 1 \right],$$

$$B(\epsilon) = I^2_{\mathbf{k}'\mathbf{k}} W_0 \frac{2\pi^2\hbar^3}{m^*} \left(\frac{\hbar\omega}{2m^*}\right)^{1/2} \times \left[\Theta(\epsilon_-) n_{\mathbf{q}} \frac{f_0(\epsilon_-)}{f_0(\epsilon)} \frac{D(\epsilon_-)}{\gamma^{1/2}(\epsilon_-)\gamma^{1/2}(\epsilon)} \tanh^{-1} \left(\frac{\gamma(\epsilon_-)}{\gamma(\epsilon)}\right)^{1/2} + (n_{\mathbf{q}} + 1) \frac{f_0(\epsilon_+)}{f_0(\epsilon)} \frac{D(\epsilon_+)}{\gamma^{1/2}(\epsilon_+)\gamma^{1/2}(\epsilon)} \coth^{-1} \left(\frac{\gamma(\epsilon_+)}{\gamma(\epsilon)}\right)^{1/2} \right]$$

and

$$C(\epsilon) = -I^2_{\mathbf{k}'\mathbf{k}} W_0 \frac{\pi^2\hbar^3}{m^*}\left(\frac{\hbar\omega}{2m^*}\right)^{1/2}\left(n_{\mathbf{q}}+1\right)\frac{f_0(\epsilon_+)}{f_0(\epsilon)}\frac{v(\epsilon_+)}{v(\epsilon)} \times \frac{D(\epsilon_+)}{\gamma^{1/2}(\epsilon_+)\gamma^{1/2}(\epsilon)}\left[\frac{\gamma(\epsilon_+)+\gamma(\epsilon)}{\gamma^{1/2}(\epsilon_+)\gamma^{1/2}(\epsilon)}\coth^{-1}\left(\frac{\gamma(\epsilon_+)}{\gamma(\epsilon)}\right)^{1/2} - 1\right], \tag{4.120}$$

where we have used the short-hand $\epsilon_\pm = \epsilon \pm \hbar\omega$ and W_0 was given earlier in (4.80)

Figure 4.4 shows a graph of the room-temperature polar optical phonon relaxation times in GaAs using these ladder coefficients. The first point to note is the characteristic saw-tooth pattern at low energies at intervals of the phonon energy ($\hbar\omega_{\mathbf{q}} = 35\,\text{meV}$). The first discontinuity occurs at the phonon energy when the step function in (4.116) becomes non-zero and the process for absorption of a phonon by a final state becomes viable. The relaxation time goes through a minimum at about 80 meV, implying that the scattering is strongest at this energy. Thereafter, $\tau(\epsilon)$ increases, smoothing out and approaching a roughly $\epsilon^{1/2}$ dependence at very high energy.

Using (4.39) we find that a population of electrons with an average energy around the minimum of $\tau(\epsilon)$ would have an electron temperature of roughly 600 K. Thus, whilst polar optical scattering is the dominant scattering process in GaAs at room temperature, Fig. 4.4 indicates that at high electron temperatures, the scattering rate weakens. Since there is no weakening of acoustic phonon scattering at high energy, this latter process starts to become predominant in a hot electron population.

4.5.4 High-Field Solution

Solution in a Single Valley

For the high-field solution of the BTE it will be useful to resurrect the time dependence so that, for a single valley labelled by n, we have

$$\frac{\mathrm{d}f^n(\mathbf{k})}{\mathrm{d}t} = \left(\frac{\partial f^n(\mathbf{k})}{\partial t}\right)_s + \frac{e\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f^n(\mathbf{k}). \tag{4.121}$$

The distribution function may be expanded as a series of Legendre polynomials [8]

$$f^n(\mathbf{k}) = \sum_j f^n_j(\epsilon) P_j(x), \tag{4.122}$$

where x is, again, the cosine of the angle θ between $\mathbf{E}$ and $\mathbf{k}$. It should be borne in mind that, despite the similarity in notation between this and that used in the last section, $f^n_j(\epsilon) \neq f_j(\epsilon)$. For $j > 0$ this is easily seen since in (4.122) these functions

of energy are multiplied by $P_j(x)$. In the case of $f_0^n(\epsilon)$, however, it should be noted that this is *not* the Fermi-Dirac factor. Rather, it may be approximated as being so in the low-field solution.

The first two Legendre polynomials are $P_0(x) = 1$ and $P_1(x) = x$. The higher order $P_j(x)$ and their derivatives may then be obtained from the relations [38]

$$x P_j(x) = \frac{j+1}{2j+1} P_{j+1}(x) + \frac{j}{2j+1} P_{j-1}(x) \tag{4.123}$$

and

$$\left(1 - x^2\right) \frac{\mathrm{d}P_j}{\mathrm{d}x} = \frac{j\,(j+1)}{2j+1} \left(P_{j-1}(x) - P_{j+1}(x)\right). \tag{4.124}$$

The action of the $\nabla_{\mathbf{k}}$ operator on the terms of the summation in (4.122) is

$$\nabla_{\mathbf{k}} f_j^n(\epsilon) P_j(x) = \hbar \mathbf{v}(\epsilon) \frac{\mathrm{d}f_j^n}{\mathrm{d}\epsilon} P_j(x) + f_j^n(\epsilon) \nabla_{\mathbf{k}} P_j(x) \tag{4.125}$$

where the derivative of $f_j^n(\epsilon)$ has been carried out in the same way as before, bringing out the group velocity $\mathbf{v}(\epsilon)$. For the action on $P_j(x)$ it is more convenient to change to spherical polar coordinates. Since $P_j(x)$ only depends on θ via $x = \cos\theta$, the only relevant component of $\nabla_{\mathbf{k}}$ is

$$\frac{1}{k} \frac{\partial}{\partial \theta} \mathbf{e}_\theta = \frac{1}{k} \frac{\mathrm{d}x}{\mathrm{d}\theta} \frac{\mathrm{d}}{\mathrm{d}x} \mathbf{e}_\theta = -\frac{1}{k} \left(1 - x^2\right)^{1/2} \frac{\mathrm{d}}{\mathrm{d}x} \mathbf{e}_\theta, \tag{4.126}$$

where $\mathbf{e}_\theta$ is the unit vector pointing in the direction of increasing θ. Thus, $\mathbf{E} \cdot \mathbf{e}_\theta = -E \sin\theta = -E(1 - x^2)^{1/2}$ and substitution of (4.122) into the force term of the BTE gives

$$\frac{e}{\hbar} \mathbf{E} \cdot \nabla_{\mathbf{k}} f^n(\mathbf{k}) = eE \sum_j \left\{ v(\epsilon) \frac{\mathrm{d}f_j^n}{\mathrm{d}\epsilon} x P_j(x) + \left(1 - x^2\right) \frac{\mathrm{d}P_j}{dx} \frac{f_j^n(\epsilon)}{\hbar k} \right\}. \tag{4.127}$$

Using (4.123) and (4.124) to eliminate x and $\left(1 - x^2\right)$, we have

$$\begin{aligned} \frac{e}{\hbar} \mathbf{E} \cdot \nabla_{\mathbf{k}} f^n(\mathbf{k}) = eE \sum_j \Bigg\{ & \frac{j+1}{2j+1} \left[v(\epsilon) \frac{\mathrm{d}f_j^n}{\mathrm{d}\epsilon} - j \frac{f_j^n(\epsilon)}{\hbar k} \right] P_{j+1}(x) \\ & + \frac{j}{2j+1} \left[v(\epsilon) \frac{\mathrm{d}f_j^n}{\mathrm{d}\epsilon} + (j+1) \frac{f_j^n(\epsilon)}{\hbar k} \right] P_{j-1}(x) \Bigg\}. \end{aligned} \tag{4.128}$$

In practice, the summation in (4.128) must be truncated at some maximum j, which gives the highest order of the $f_j^n(\epsilon)$. Here, we will consider only terms up to $j = 1$. We then have

$$\frac{e}{\hbar}\mathbf{E}\cdot\nabla_{\mathbf{k}}f^n(\mathbf{k}) = eE\left\{\frac{1}{3}\left[v(\epsilon)\frac{\mathrm{d}f_1^n}{\mathrm{d}\epsilon}+2\frac{f_1^n(\epsilon)}{\hbar k}\right]P_0(x)+v(\epsilon)\frac{\mathrm{d}f_0^n}{\mathrm{d}\epsilon}P_1(x)\right.$$
$$\left.+\frac{2}{3}\left[v(\epsilon)\frac{\mathrm{d}f_1^n}{\mathrm{d}\epsilon}-\frac{f_1^n(\epsilon)}{\hbar k}\right]P_2(x)\right\}. \tag{4.129}$$

The scattering integral, (4.93), may also be expanded as per (4.122). Retaining only terms in $f_0^n(\epsilon)$ and $f_1^n(\epsilon)$, we obtain integrals with the same form for the integrands as found earlier in (4.104) and (4.105), with $f_j^n(\epsilon)P_j(x)$ in the place of the $f_j(\epsilon)$. Recalling that $P_0(x)=1$ and $P_1(x)=x$, we then define the scattering integrals I_0 and I_1 by

$$P_0(x)I_0 = \int s(\mathbf{k}',\mathbf{k})f_0^n(\epsilon')\left[1-f_0^n(\epsilon)\right]-s(\mathbf{k},\mathbf{k}')f_0^n(\epsilon)\left[1-f_0^n(\epsilon')\right]\frac{V_C}{(2\pi)^3}d^3\mathbf{k}' \tag{4.130}$$

and

$$P_1(x)I_1 = x\int s(\mathbf{k}',\mathbf{k})\left\{f_1^n(\epsilon')\cos\alpha'\left[1-f_0^n(\epsilon)\right]-f_1^n(\epsilon)f_0^n(\epsilon')\right\}$$
$$-s(\mathbf{k},\mathbf{k}')\left\{f_1^n(\epsilon)\left[1-f_0^n(\epsilon')\right]-f_1^n(\epsilon')\cos\alpha'\ f_0^n(\epsilon)\right\}\frac{V_C}{(2\pi)^3}d^3\mathbf{k}', \tag{4.131}$$

where we have used the same trick as in Sect. 4.5.3 to transform x'/x to $\cos\alpha'$. Equating the coefficients of the $P_j(x)$, we then have the simultaneous equations

$$\frac{\mathrm{d}f_0^n}{\mathrm{d}t} = I_0+\frac{eE}{3}\left[v(\epsilon)\frac{\mathrm{d}f_1^n}{\mathrm{d}\epsilon}+2\frac{f_1^n(\epsilon)}{\hbar k}\right], \tag{4.132}$$

$$\frac{\mathrm{d}f_1^n}{\mathrm{d}t} = I_1+eEv(\epsilon)\frac{\mathrm{d}f_0^n}{\mathrm{d}\epsilon}. \tag{4.133}$$

We do not include $P_2(x)$ in the above for two reasons. Firstly, we did not expand the scattering integral to high enough order, so there was nothing to equate to, even though these terms would exist in a more exact treatment. Secondly, even in similar treatments when $P_2(x)$ is included (for instance, Ref. [9]), it is usually averaged to make it spherically symmetric. However, in 3D, this means that $\langle P_2(x)\rangle = 0$ anyway. Note that putting $f_0^n(\epsilon)=f_0(\epsilon)$ and $f_1^n(\epsilon)x=f_1(\epsilon)$, I_0 disappears and, in the steady state, we are left with the low-field solution of Sect. 4.5.3.

Equations (4.133) may be re-written using (4.14) and (4.23) to express $v(\epsilon)$ and k in terms of $\gamma(\epsilon)$ and its derivative

$$\frac{\mathrm{d}f_0^n}{\mathrm{d}t} = I_0+\frac{eE}{3}\left(\frac{2}{m^*\gamma(\epsilon)}\right)^{1/2}\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon}\right)^{-1}\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left[\gamma(\epsilon)f_1^n(\epsilon)\right], \tag{4.134}$$

$$\frac{\mathrm{d}f_1^n}{\mathrm{d}t} = I_1 + eE\left(\frac{2\gamma(\epsilon)}{m^*}\right)^{1/2}\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon}\right)^{-1}\frac{\mathrm{d}f_0^n}{c\epsilon}. \tag{4.135}$$

These may be simplified by noting that $I_1 = (df_1^n/dt)_s/x$ and using the low-field result of (4.110) to make the approximation that in the steady state

$$\frac{1}{x}\left(\frac{\mathrm{d}f_1^n(\epsilon)x}{\mathrm{d}t}\right)_s = -\frac{f_1^n(\epsilon)}{\tau(\epsilon)}. \tag{4.136}$$

Substituting this into the second equation of (4.135) with $df_1^n/dt = 0$, we then find

$$f_1^n(\epsilon) = eE\left(\frac{2\gamma(\epsilon)}{m^*}\right)^{1/2}\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\epsilon}\right)^{-1}\frac{df_0^n}{\mathrm{d}\epsilon}\tau(\epsilon), \tag{4.137}$$

which, in turn, is substituted into the first equation of (4.135) to give

$$\frac{df_0^n}{dt} = I_0 + \frac{2e^2E^2}{3m^*}\left(\frac{1}{\gamma(\epsilon)}\right)^{1/2}\left(\frac{\mathrm{d}\gamma(\epsilon)}{\mathrm{d}\epsilon}\right)^{-1}\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left[\frac{\gamma^{3/2}(\epsilon)\tau(\epsilon)}{\mathrm{d}\gamma/\mathrm{d}\epsilon}\frac{\mathrm{d}f_0^n}{\mathrm{d}\epsilon}\right]. \tag{4.138}$$

Equation (4.38) is the differential equation that needs to be solved for transport in a single valley. However, as it stands, I_0 would still require numerical integration. As a first simplification, it is straightforward to show that for purely elastic processes $I_0 = 0$. Thus, we need only consider phonon scattering and put

$$I_0 = \left(\frac{\partial f_0^n}{\partial t}\right)_{ph}, \tag{4.139}$$

with the *ph* subscript standing for 'phonon'. Considering this scattering term to be due to polar optical phonon scattering only, Conwell and Vassell's result [8] may be written as

$$\begin{aligned}\left(\frac{\partial f_0^n}{\partial t}\right)_{PO} =& I_{\mathbf{k}'\mathbf{k}}^2 W_0 \frac{2\pi^2\hbar^3}{m^*}\left(\frac{\hbar\omega}{2m^*}\right)^{1/2}\Big[\Theta(\epsilon_-)\big\{n_{\mathbf{q}}f_0(\epsilon_-) - (n_{\mathbf{q}}+1)f_0(\epsilon)\big\}\\ &\times\frac{D(\epsilon_-)}{\gamma^{1/2}(\epsilon_-)\gamma^{1/2}(\epsilon)}\tanh^{-1}\left(\frac{\gamma(\epsilon_-)}{\gamma(\epsilon)}\right)^{1/2}\\ &+\big\{(n_{\mathbf{q}}+1)f_0(\epsilon_+) - n_{\mathbf{q}}f_0(\epsilon)\big\}\\ &\times\frac{D(\epsilon_+)}{\gamma^{1/2}(\epsilon_+)\gamma^{1/2}(\epsilon)}\coth^{-1}\left(\frac{\gamma(\epsilon_+)}{\gamma(\epsilon)}\right)^{1/2}\Big].\end{aligned}$$

The method of Seifikar et al. [39] may then be used to solve (4.138) via the finite difference method. The calculation is started at zero field with an initial function for f_0^n and allowed to evolve to a steady state. In the next iteration, the previously found

result for f_0^n is used as a starting value with a small increase in the electric field. The process is repeated until the desired field strength is reached.

InterValley Transfer

The effects of transferred electrons via intervalley scattering may be incorporated by a generalization of the single valley solution. There will then be a system of equations like (4.138) for each valley but now additional scattering terms must be included in each equation for the intervalley scattering rates. In addition, these rates must also be included in the solution of the I_1 integrals for each valley, so that the relaxation times obtained will be modified. The paper by Conwell and Vassell [8] remains a very good introduction to this method.

Acknowledgments The author thanks Masoud Seifikar for useful discussions on the high-field solution of the Boltzmann equation. The author's current position at the Tyndall National Institute is funded by the Science Foundation Ireland.

References

1. K.M. Johnson, Digest of Technical Papers, International Solid State Circuits Conference, vol. 7, 64 (1964)
2. J.B. Gunn, IBM J. Res. & Dev. **8**, 141 (1964)
3. A. O'Brien, N. Balkan, J. Roberts, Appl. Phys. Lett. **70**, 366 (1997)
4. S. Chung, N. Balkan, Appl. Phys. Lett. **86**, 211111 (2005)
5. B.K. Ridley, Rep. Prog. Phys. **54**, 169 (1991)
6. N.W. Aschcroft and N.D. Mermin, Solid State Physics, (Saunders, Philadelphia 1976)
7. H. Ehrenreich, Phys. Rev. **120**, 1951 (1960)
8. E.M. Conwell, M.O. Vassell, Phys. Rev. **166**, 797 (1967)
9. B.K. Ridley, Quantum Processes in Semiconductors, 4th ed. (Clanderon Press, Oxford 1999)
10. H. Brooks, Adv. Electron. Electron Phys. **7**, 85 (1955)
11. L. Nordheim, Ann. Phys. **9**, 607 (1931)
12. P.A. Flinn, Phys. Rev. **104**, 350 (1956)
13. G.L. Hall, Phys. Rev. **116**, 604 (1959)
14. A.E. Asch, G.L. Hall, Phys. Rev. **132**, 1047 (1963)
15. J.W. Harrison, J.R. Hauser, Phys. Rev. B **13**, 5347 (1976)
16. F. Murphy-Armando, S. Fahy, Phys. Rev. Lett. **97**, 96606 (2006)
17. R.M. Martin, Electronic structure: basic theory and practical methods (Cambridge University Press, Cambridge 2004)
18. J.M. Ziman, Electrons and Phonons (Oxford University Press, Oxford 1960)
19. M.V. Fischetti, S.E. Laux, J. Appl. Phys. **80**, 2234 (1996)
20. S. Joyce, F. Murphy-Armando, S. Fahy, Phys. Rev. B **75**, 155201 (2007)
21. F. Murphy-Armando, S. Fahy, Phys. Rev. B **78**, 35202 (2008)
22. H. Fröhlich, Adv. Phys. **3**, 325 (1954)
23. B.K. Ridley, Semicond. Sci. Technol. **4**, 1142 (1989)
24. B.K. Ridley, W.J. Schaff, L.F. Eastman, J. Appl. Phys. **96**, 1499 (2004)
25. P.A. Wolff, Phys. Rev. **95**, 1415 (1954)
26. W. Shockley, Solid State Electron. **2**, 35 (1961)

27. B.K. Ridley, J. Phys. C: Solid State Phys. **16**, 3373 (1983)
28. T.P. Lee and T. Li in *Optical Fiber Telecommunications I, Ch. 18*, ed. by S.E. Miller and A.G. Chynoweth (Academic Press, San Diago 1979)
29. R.J. McIntyre, IEEE Trans. Electron. Dev. **13**, 164 (1966)
30. B.K. Ridley, T.B. Watkins, Proc. Phys. Soc. **78**, 293 (1961)
31. C. Hilsum, Proc. IRE **50**, 185 (1962)
32. W. Shan, W. Walukiewicz, J.W. Ager III, E.E. Haller, J.F. Geisz, D.J. Friedman, J.M. Olson, S.R. Kurtz, Phys. Rev. Lett. **82**, 1221 (1999)
33. A. Patanè, A. Ignatov, D. Fowler, O. Makarovsky, L. Eaves, L. Geelhaar, H. Riechert, Phys. Rev. B **72**, 033312 (2005)
34. B.K. Ridley, Proc. Phys. Soc. **82**, 954 (1963)
35. C. Kittel and H. Kroemer, Thermal Physics 2nd ed. (W.H. Freeman and Co., New York 2002), p. 408
36. K. Fletcher, P.N. Butcher, J. Phys. C: Solid State Phys. **5**, 212 (1972)
37. M.P. Vaughan, Alloy and Phonon Scattering-Limited Electron Mobility in Dilute Nitrides (University of Essex, UK 2007)
38. W. Koepf, Hypergeometric Summation: An Algorithmic Approach to Summation and Special Function Identities (Braunschweig, Germany 1998) p. 2
39. M. Seifikar, E.P. O'Reilly and S. Fahy, Phys. Rev. B **84**, 165216 (2011)

Chapter 5
Monte Carlo Techniques for Carrier Transport in Semiconductor Materials

N. Vogiatzis and Judy M. Rorison

Abstract Monte Carlo has become a powerful tool for describing complex systems with many degrees of freedom. It involves simulating a combination of deterministic and stochastic processes. Here, after a basic introduction to the technique, we focus on its application in the analysis of carrier transport in semiconductors. This method is applied to GaAs and to dilute nitride materials.

5.1 Introduction to Monte Carlo

Monte Carlo (MC) is a technique that simulates problems with the help of computer algorithms. These problems are usually very complex in nature with many degrees of freedom which do not have an analytical solution.

Monte Carlo techniques are employed nowadays in engineering, physics, biology, mathematics, finance, telecommunications and many other different areas. Forecast of the dynamical evolution of a tornado, prediction of the expectation value of a stock price or planning of the wireless network coverage to optimise capacity, traffic and quality are some of the many examples of how this technique can be applied. Essentially, MC employs statistical sampling to solve quantitative problems.

Statistical sampling is associated with the generation of random numbers. However, "truly" random numbers are not always practical to use for a number of reasons. This is discussed in the Appendix. From the early stages that this method was proposed, it was obvious that a quick and reliable way of generating random numbers was the computer. Since then MC has been associated with computers and their efficiency in generating random numbers automatically and simulating the whole problem.

N. Vogiatzis · J.M. Rorison (✉)
Department of Electrical and Electronic Engineering,
University of Bristol, Merchant Venturers Building,
Woodland Rd, Bristol, BS8 1UB, UK
e-mail: judy.rorison@bristol.ac.uk

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_5,

5.1.1 Historical Review

The earliest record of Monte Carlo is in 1777 describing a problem known as the Buffon needle problem [1]. This involved dropping a needle to a lined surface in order to estimate π. The credit of inventing the modern Monte Carlo method goes to Ulam who worked with von Neumann on the Manhattan project during the Second World War. Ulam invented the Monte Carlo technique in 1946 while trying to work out the probabilities of winning a card game known as solitaire [2]. He figured out that he could come up with a method that he could repeat many times and at the end count the number of successes of these random operations. He then associated this problem with the neutron diffusion, other problems from mathematical physics and generally other processes which are described by differential equations.

His big contribution was that the statistical sampling that was required for this method could be performed by means of the newly invented computer. He described his idea to von Neumann and Metropolis and they started developing computer algorithms for this problem, as well as for other non-random problems which could however be given by some random forms and then through statistical sampling get the solution. Metropolis named the newly invented method after the casinos in Monte Carlo and he published with Ulam the first paper on this method in 1946 [3].

5.1.2 Simple Examples of Monte Carlo

First example

Monte Carlo is very useful in the calculation of integrals. Traditional numerical integration can be done in a number of ways, such as using the rectangle rule, the trapezoidal rule or Simpson's rule. Monte Carlo can also do numerical integration by using random numbers. Especially for the calculation of higher dimensional integrals this method is very efficient. An example of higher dimensional integration is in statistical physics for the calculation of the average distance between particles within a box of length L.

A characteristic example of Monte Carlo integration is the so-called "hit or miss" method [4] that is used for the calculation of π. The idea is that we want to calculate the integral

$$\int_a^b f(x)dx \tag{5.1}$$

that is the area under the curve $f(x)$ from $x = a$ to $x = b$. For the case of π the problem is stated as follows. Imagine that we have some grains of rice. The aim is to calculate how many grains of rice will be within the area defined by the function $y = \sqrt{1 - x^2}$ where $0 < x < 1$ and $0 < y < 1$ to derive the value of π. Figure 5.1 shows the problem graphically. We can see that the area we want to measure is the quarter circle which can be squared by a square of length $x = y = 1$. Then the ratio

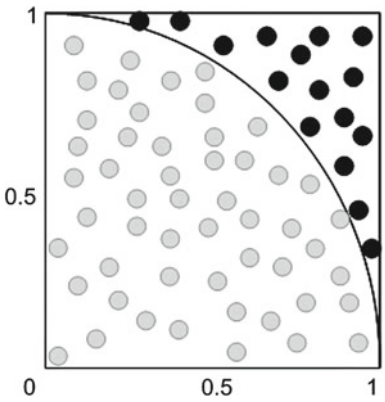


Fig. 5.1 Randomly generated pairs of x, y for the calculation of π

of the area of the quarter circle over the area of the square is

$$A = \frac{\frac{\pi}{4}1^2}{1^2} \tag{5.2}$$

which equals $\pi/4$. This means that if we throw randomly the grains of rice over the entire area, then by counting the grains in the quarter circle (grey circles) and dividing them by the total amount of grains (grey and black circles) we should find a value close to $\pi/4$.

This can be done easily with Monte Carlo. All we need to do is to generate some pairs x, y, where x, y are uniformly distributed random numbers within 0 and 1. Then if $x^2 + y^2 < 1$ the grain is within the circle (grey spots), otherwise it will be outside (black spots). This is shown in Fig. 5.2 where we can see that by increasing the number of random pairs we get a value of $\pi = 3.14570$ which tends to converge towards the real value (3.14159). The standard error will be inversely proportional to the square root of the sample size (number of iterations). We shall see how the standard error is related to the computational time in the section where we will talk about the ensemble Monte Carlo in carrier transport in semiconductors.

It is worth pointing out that the random numbers x and y are taken from the uniform distribution, i.e. all numbers between 0 and 1 have exactly the same probability of occurrence. This is very important, for the validity of the results. For example, if we choose numbers from the bell distribution, then at the corners of the square of Fig. 5.1 the probability of finding a grain will be small as it will correspond to the tail of the distribution. The numbers will not be uniformly distributed anymore which will affect the final outcome. For the same reason if we take vertical slices at $x = x_1, x = x_2, x = x_3, \ldots$ it is easy to understand that the number of grains measured would be less sensitive to the shape of the distribution at $x \to 0$ than at $x = 0.5$. The random numbers (or pseudorandom numbers to be more accurate as discussed in the Appendix) which will be used in the following sections are all generated from the uniform distribution.

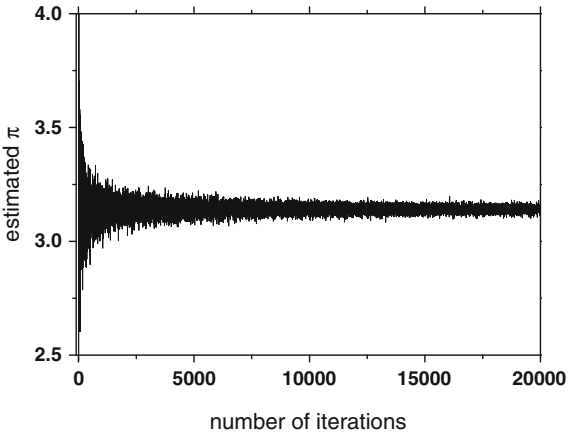


Fig. 5.2 Estimated value of π as a function of the number of generated pairs. The estimate converges towards the real value of π as the random pairs increase

Second example

Another example is a gambling experiment [4]. In this there is a person called gambler who has 100$ and plays against the house who has 2000$. Let us assume that we have a random number r which is generated from a uniform distribution. The rule of the game is the following: if $r < 0.5$ then the player wins 1$ otherwise the house wins 1$. The game finishes when either the player or the house goes bankrupt.

The problem with this experiment is that it is difficult to predict not only who will be the winner (although the odds are clearly in favour of the house), but most importantly how long the game is going to last (how many iterations within a single game), as there is no obvious time limit. However, we can create a small computer algorithm that will simulate this problem and then we can do multiple runs to get an idea of who wins and what would the average duration of the game be. This is shown in Fig. 5.3. In (a) four independent games are shown. We can see that in three of them the gambler loses and only in one he manages to win. Also, we can see that the game duration (number of iterations) is not the same. In (b) the histograms show the number of games (here only 30 are simulated) that will have finished within a specific number of iterations per game either by means of the gambler winning or losing. We can see that most of the games are fairly short and within 50000 iterations (50000 times choosing a random number) there will be a winner, whilst some other games are more "dramatic" and take longer for someone to win. In fact for the record, out of the 30 games the gambler wins only in three of them.

This example shows that by using some random numbers and a deterministic set of rules we can extract a meaningful statistical average of the quantities we are interested in and which would otherwise be difficult to obtain by other analytical or numerical methods.

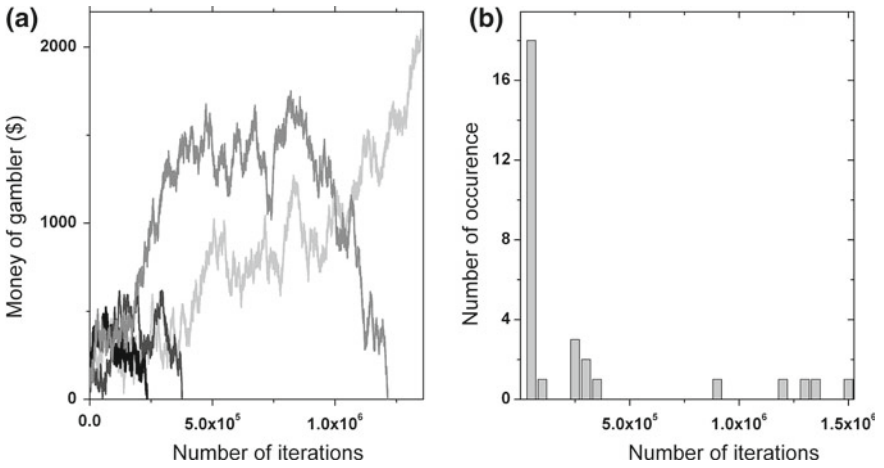


Fig. 5.3 **a** Number of iterations for a game to finish either when the gambler loses (0 ($)) or wins (2100 $). Each curve represents a single game, **b** histograms of the frequency of occurrence describing the game time. 30 games are simulated and most of them have relative short duration

5.2 Carrier Transport in Semiconductors

The transport problem in the semiconductors attempts to solve the Boltzmann Transport Equation (BTE) which gives the distribution function in momentum space **p** of the carriers. By carriers we mean either electrons or holes. In other words it describes the probability of finding a carrier with a momentum **p**, at the location **r** and at time **t**. Let us from now on assume for simplicity that whenever we say carriers we shall assume electrons. If n_i is the number of electrons in the ith valley per unit volume of the crystal then

$$\int_i f_i(\mathbf{p}) d\mathbf{p} = n_i \tag{5.3}$$

The Boltzmann equation for the distribution function $f_i(\mathbf{p})$ of carriers of momentum **p** in the ith valley in the presence of an applied field **F** can be written as [5]

$$\frac{\partial f_i(\mathbf{p})}{\partial t} = [\frac{\partial f_i(\mathbf{p})}{\partial t}]_F + \sum_j [\frac{\partial f_i(\mathbf{p})}{\partial t}]_{ij} \tag{5.4}$$

where the first term on the right hand side is the rate of change of $f_i(\mathbf{p})$ in time due to the field

$$[\frac{\partial f_i(\mathbf{p})}{\partial t}]_F = \frac{q}{\hbar} \mathbf{F} \nabla_k f_i(\mathbf{p}) \tag{5.5}$$

with q being the electronic charge. The second term on the right-hand side is the summation over all possible valleys i of the rate of change of $f_i(\mathbf{k})$ due to the scattering processes. If $j = i$, then $[\partial f_i(\mathbf{p})/\partial t]_{ii}$ is due to intravalley scattering (in valley i) and if $j \neq i$ the change $[\partial f_i(\mathbf{p})/\partial t]_{ij}$ is due to intervalley scattering (between valleys i and j). In 5.4 the external applied field is assumed to be electric. The problem can be generalised to include also magnetic fields. For simplicity in what follows we shall assume that carriers move only under the presence of an electric field.

Equation 5.4 gives a classical description. An analytical solution of this equation is difficult if summation is over more than one scattering rate making this approach only useful at low fields or for a specific form of the distribution function. For moderate electric fields several scattering processes must be included to obtain meaningful results showing the interplay between competing scattering processes and acceleration by the electric field. For high electric fields the so-called hot electron problem may arise. By hot electrons we mean electrons subject to high electric fields resulting in their energy increasing in such a way that their Fermi-Dirac-like distribution is hotter than the lattice. In such a case we say that electrons are not in thermal equilibrium with the lattice. Because of the complexity of the scattering processes 5.4 turns out to be complicated. Other constraints of this method have to do with transport in space and time as well as with the incorporation of the band structure of the semiconductors.

Another way to solve the Boltzmann equation is by a numerical method. It is easier to simulate the trajectories of carriers when they move in a device by using a combination of stochastic and deterministic processes under the presence of an applied field than solving the BTE. Historically, the proposed techniques were the iterative method [6] and the Monte Carlo method [7]. Both of them are numerical methods, but the latter has became more popular and has managed to simulate successfully the transport behaviour of a number of semiconductor material systems. For a comparison between the numerical techniques and the analytical one as well as for an extensive description on the Monte Carlo method in particular, the reader is advised to look at the work of Jacoboni et al. [8, 9], Price [10] and Boardman [11].

In the following pages a detailed description of the Monte Carlo method for carrier transport in semiconductors is given in a step-by-step approach. In each step we shall attempt to highlight the deterministic and stochastic processes involved. First, the single electron Monte Carlo (SMC) will be examined which is appropriate for *steady-state* conditions. Then the Ensemble Monte Carlo (EMC) will be described which is related to the dynamical behaviour (*transient characteristics*) of the carriers under an external field. Examples of the two methods are given within the GaAs semiconductor. Also, there will be a discussion on the same problem in dilute nitride semiconductors. Finally, there will be a reference on full band Monte Carlo simulation where analyticity is not always applicable.

5.3 Single Electron Monte Carlo

Motion of carriers in a semiconductor takes place within a bath of other carriers. Therefore, carrier transport is a many-body problem and normally all interactions between carriers should be taken into account in the description of the system. However, to our benefit the particular type of problem refers to a system which is *ergodic*.

We can understand ergodicity through the following example. Imagine that we are performing a statistical analysis on a group of people at the type of film they choose to watch at a specific moment in time. Some may choose to watch a comedy, others a film with historic content, others an action film, etc. Let us take now a specific individual from that group and follow his preferences over a large period of time. We can understand that we can either make a statistical analysis of an ensemble of people at a specific moment or an analysis of just one person for a longer period. If the system is ergodic this means that both types of statistical analysis should give the same result.

In examining the steady-state behaviour of the motion of a carrier, i.e. a long time after an electric field has been applied and the independent carriers within the ensemble have obtained a specific velocity, energy or distribution, we can monitor the behaviour of just a single carrier, assuming that it is identical with the rest. This is the meaning of the SMC simulation and luckily it does not require any knowledge of the initial distribution function which simplifies things.

The two subsequent sections will serve as an introduction to the concept of scattering and drift of carriers. The algorithm of the SMC will be presented and each step will be analysed.

5.3.1 Scattering Processes

When applying an electric field the carrier drifts. It is common to use the term *free flight* to describe this process which can be described in a classical way as we shall see later on. At some point in time the free flight will terminate because the carrier will be scattered by a process. This scattering that takes place can be due to phonons, i.e. lattice vibrations, impurities or other carriers (electrons or holes). Also, after scattering carriers may retain the energy they had before (elastic scattering) or may have an altered one (inelastic scattering). By scattering we do not necessarily mean that the carriers have to collide physically with the scatterer as this process can take place from a distance, such as in a Coulombic type of interaction.

Some basic assumptions which are done in MC are the following. Whilst the free flight time of the carrier is finite and can be determined easily, the scattering process is assumed to happen instantaneously. The motion of carriers is described classically(analogy of carriers/particles being seen as solid spheres), whilst the scattering is treated quantum mechanically. As shown in Fig. 5.4 the main scattering

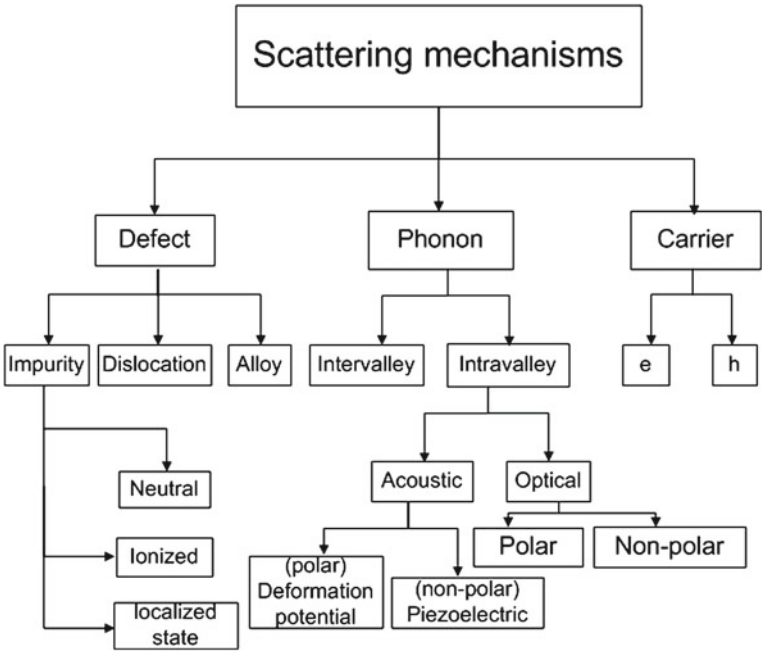


Fig. 5.4 Scattering processes in semiconductors

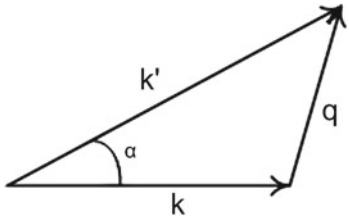


Fig. 5.5 Carrier scattering from a state with wavevector $\mathbf{k}$ to a new $\mathbf{k}'$ The change in the wavevector is $\mathbf{q}$ and α is the polar angle between the incident and the scattered state

processes can be further analysed to other subcategories. Although not all scattering processes are equally important in all types of semiconductors, what remains the same is their general quantum mechanical expression. This means that for MC the scattering rate S_j of a process j is a *deterministic* input. As we shall show the *randomness* is associated with the *frequency of occurrence* of this process.

The deterministic carrier scattering rate (or transition rate) from a state $\mathbf{k}$ to a state $\mathbf{k}'$ (Fig. 5.5) is given by the following formula known as Fermi's golden rule

$$S(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar} |\langle \mathbf{k}' | H' | \mathbf{k} \rangle|^2 \delta(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}} \mp \hbar\omega) \tag{5.6}$$

where $\mathbf{k}$ and $\mathbf{k}'$ is the initial and final wavevector, $\epsilon_{\mathbf{k}}$ and $\epsilon_{\mathbf{k}'}$ is the energy of the carrier before and after scattering and $\hbar\omega$ the energy that has been absorbed $(-)$ or emitted

(+). For $\Delta\epsilon = \hbar\omega$=0 the scattering is elastic. Also, H' is the perturbative potential characteristic of each scattering process. It is beyond the scope of this chapter to give the expressions for the potential H'. For a detailed analysis the reader is advised to see other textbooks [9, 12, 13].

Let us take as an example some scattering mechanisms for electrons in the Γ valley of GaAs [14]. The first is acoustic deformation potential scattering, the second is polar optical phonon and the third is ionised impurity scattering. The electronic band is assumed to be spherical and non-parabolic. Non-parabolicity refers to the shape of the conduction band and it is a very important concept when describing electron motion away from the bottom of the band, which happens for moderate and high fields. Therefore, non-parabolicity describes the deviation of the energy from the simple quadratic dependence for values of $\mathbf{k}$ away from the band minima [14–16] and must be accounted for in the expressions of the scattering rates and later in the post-scattering selection process. The scattering rates are plotted as a function of energy in Fig. 5.6(a). Note that we have explicitly treated polar optical phonon absorption and emission. The total scattering rate is the sum of the scattering rate of each independent process

$$S_{total} = \sum_j S_j \tag{5.7}$$

We can use S_{total} to obtain the fractional contribution (weight factor) of each one of them as

$$\Gamma_j = \frac{S_j}{S_{total}} \tag{5.8}$$

This is shown in Fig. 5.6(b). We shall use this graph later when we will discuss the selection of the scattering process which happens in a random fashion.

5.3.2 Drift Process

Let us assume that we have an electron that is moving in the semiconductor crystal under the presence of an electric field $\mathbf{F}$. We mentioned earlier in Sect. 5.3.1 that this motion can be described in a classical way, meaning that we can use Newton's second law which says that the total force f applied in a body is equal to the time derivative of the momentum of the body.

$$\mathbf{f} = m\mathbf{a} = m\dot{\mathbf{v}} = \dot{\mathbf{p}} \tag{5.9}$$

In our case this can be written after integration as

$$(-q)\mathbf{F} = \dot{\mathbf{p}} = \hbar\dot{\mathbf{k}} \tag{5.10}$$

where $\dot{\mathbf{v}}$, $\dot{\mathbf{p}}$, $\dot{\mathbf{k}}$ is the time derivative d/dt of the velocity, momentum and wavevector respectively. The time dependence of $\mathbf{k}$ can therefore be written as

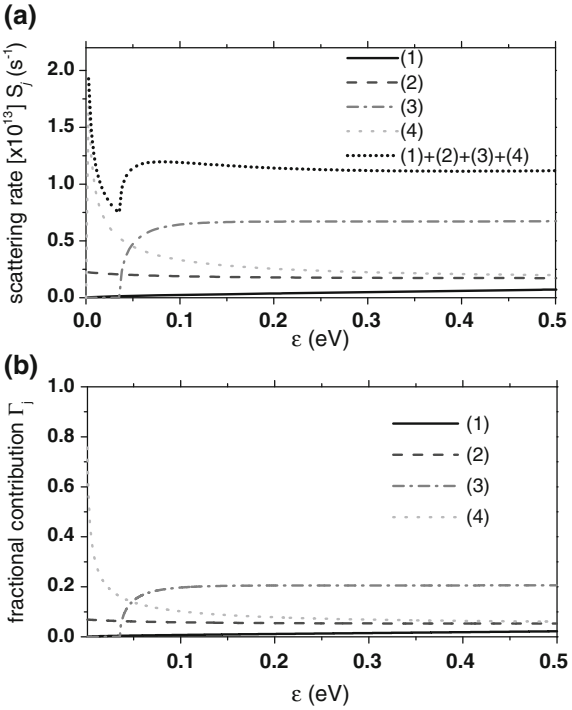


Fig. 5.6 **a** Absolute magnitude of scattering rates in GaAs. (1) acoustic deformation potential (2) polar optical phonon absorption, (3) polar optical phonon emission and (4) ionised impurity scattering with an impurity concentration of n_I=10^{16}cm^{-3} and a free electron concentration n_0=$3n_I$ using the Brooks-Herrings approach [13]. **b** The fractional contribution of each scattering process within the same energy interval. Any intervalley scattering to higher valleys has been omitted for simplicity

$$\hbar\mathbf{k}(t) = \hbar\mathbf{k}_0 - q\mathbf{F}t \tag{5.11}$$

In the above it is assumed that the electron's energy varies slowly as a function of the position and therefore it can be treated as a free particle with an effective mass m_e^*. The electron motion could in principle continue forever if it were not for one of the mechanisms shown in Fig. 5.6 that terminates the drift process by scattering. We have defined the duration between two successive scattering events as *free flight time*. Figure 5.7(a) and (b) show the trajectories of the electrons in the momentum and real space respectively. The applied electric field can be assumed to be parallel to only one direction for simplicity (here it is x). The scattering event causes change of the trajectory in the momentum space. The scattering time is assumed to happen instantaneously which is indicated by the thin arrows. Figure 5.7(c) shows the drift velocity which is being built up with time until it reaches a steady state. Typically, in such simulations electrons are expected to have obtained a steady state within a few ns. One can get an empirical understanding of how good the convergence is by

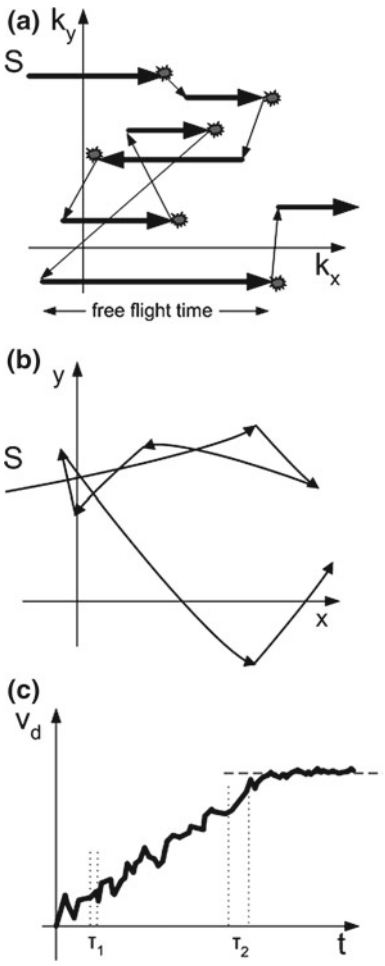


Fig. 5.7 **a** Electron trajectory in the x-y plane. z dimension is omitted for simplicity. "S" stands for start, thick arrows represent the drift process happening within the free flight time and the light arrows correspond to the change in momentum upon a scattering event (*in grey circles*). The electric field F is assumed only along x-direction as indicated by the trajectories. **b**. Electron trajectories in the real space **c** Drift velocity evolution versus time. The outcome of the simulation is sensitive both to the simulation time and to the initial conditions chosen. The free flight times such as τ_1 and τ_2 do not have to be the same and are chosen in a stochastic way. After some time electrons reach a steady-state velocity (*dashed parallel line*)

allowing a few different simulation times in the code and by monitoring the progress of the quantity of interest, or by defining some criteria such as a minimum allowed deviation of velocity values within a frame of time. Later, we shall mention how the computational cost can be reduced.

The position of a carrier varies with time as

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t v(t')dt' \tag{5.12}$$

where $\mathbf{x}(0)$ is the position of the carrier at $t = 0$. Figure 5.7(b) has only half of the trajectories of (a) because it is assumed that the instantaneous scattering process does not change the position of the carriers. In other words since the scattering time $\delta t_{\mathbf{k}\to\mathbf{k}'} = 0$, we have δ_x=0 ($\delta_x = v\delta_t$).

5.3.3 Description of the Algorithm

A description of the algorithm is given below:

1. *Input of "external" physical parameters and simulation characteristics*: Parameters such as electric field strength $\boldsymbol{F}$, lattice temperature T, maximum permitted time of simulation t_{max}[1] and other physical constants are defined. All values are deterministic.
2. *Definition of the physical system*: The band structure of the examined semiconductor is given. Typically, we assume zero energy at the band edge. The notion of the non-parabolicity parameter a_f which is input later in the scattering rates is defined here (makes use of the energy band gap value). All values are deterministic.
3. *Scattering parameters*: The scattering rates of the various processes and their fractional contribution are given in a deterministic manner (Fig. 5.6). It can be seen as the "heart" of the code or as the most sophisticated part of it because the input distribution of the scattering processes will retain the physical characteristics of the system and electrons throughout the simulation time will map on this distribution.
4. *Initial conditions of motion*: The initial energy ϵ, the wavevector $\mathbf{k}$ and its components $\mathbf{k}_x$, $\mathbf{k}_y$ and $\mathbf{k}_z$ are defined. The energy is chosen randomly but we should pay attention to giving a value which will not affect the final result. For example, an unrealistically high value of ϵ for a low electric field may take more time to converge to the steady state (computational cost) and in the case where the simulation time t_{max} is not big enough the final result will be influenced by the initial wrong conditions. The initial wavevector $\mathbf{k}$ is given deterministically by $k = \frac{\sqrt{2m_e^*}\sqrt{\gamma(\epsilon)}}{\hbar}$, where $\gamma(\epsilon) = \epsilon(1 + a_f\epsilon)$ is an expression of carrier energy accounting for non-parabolicity. ϵ is random and can be a few times bigger or smaller than the thermal energy kT by being dependent on a random number in

[1] The simulation time is the time after which the carriers stop drifting. It should not be confused with the time required for computer simulation, which will be referred to as *computational time.*

the form $log(r)$ for example. Also k_x, k_y and k_z use k (which is random because it depends on ϵ) and another set of rules with random numbers. An example will be given later.

5. *Free flight time*: The free flight time is determined in a stochastic way (see for example Fig. 5.7(c)). Details will be given later.
6. *Drift process*: The electron drifts deterministically as described in Sect. 5.3.2 for time equal to the free flight time. For systems that incorporate more than one valley a control of the carrier energy must be made at this point as carriers may reside by the end of the free flight on a different band. The final energy before scattering is obtained.
7. *Collection of data for estimators*: Quantities of interest such as drift velocity or energy are being monitored and stored at this point. They will be recalled before the next scattering event.
8. *Selection of scattering process*: By using the scattering distribution in step 3 the selection of the mechanism that terminates the free flight time is made stochastically. An example will be given later.
9. *Determination of the post scattering state*: The state after scattering is described. It is a combination of a deterministic expression characteristic of a specific mechanism and a use of two more random numbers related to the polar and the azimuthal angle. Details will be also given later.
10. *Logic check of the simulation time*: Monitor the simulation time. If it is smaller than t_{max}, Step 5 is repeated, otherwise the simulation stops and the final results are collected.

These steps are described by the flowchart in Fig. 5.8.

5.3.3.1 Initial Conditions of Motion

Figure 5.9 shows the initial distribution of an electron in the Γ valley of GaAs at two different temperatures.A total of 30000 iterations have been performed and the vertical axis describes the normalised number of coincidences. Subfigure (a) shows the energy distribution, (b) the distribution of the respective $\mathbf{k}_x$ wavevectors and (c) the distribution of $\mathbf{k}$ wavevectors. Also, the average value for each quantity is shown.

5.3.3.2 The Free Flight Time

It is obvious that the duration of the free flight time t_r must be associated with the total scattering rate—the carrier will drift for less time when the scattering is higher. Let us now assume a group of carriers n_c that have not undergone any scattering since t'=0 [21]. The rate of change of the collision-free carriers can be given by

$$\frac{dn_c}{dt'} = -S_{total} n_c \tag{5.13}$$

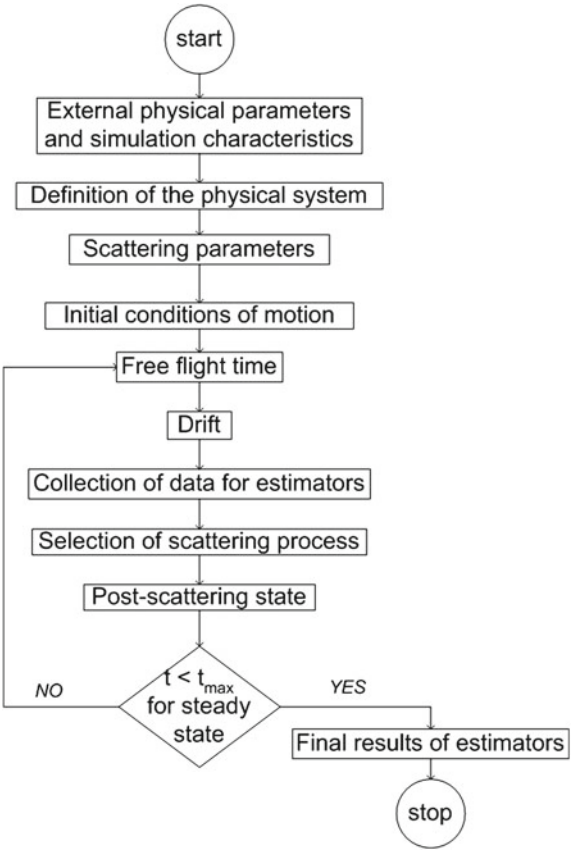


Fig. 5.8 Flowchart for Single Monte Carlo

where S_{total} is the total scattering rate (5.7). Solution of 5.13 gives

$$n_c(t) = n_c(0)e^{-\int_0^t S_{total}dt'} \tag{5.14}$$

Therefore the probability that the carriers that have suffered a collision at time t'=0, have not suffered any other collision at time t is

$$\frac{n_c(t)}{n_c(0)} = e^{-\int_0^t S_{total}dt'} \tag{5.15}$$

The probability that a carrier undergoes its first collision between t and $t + dt$ is the product of the scattering rate times the probability that it has not suffered any other collision by that time t

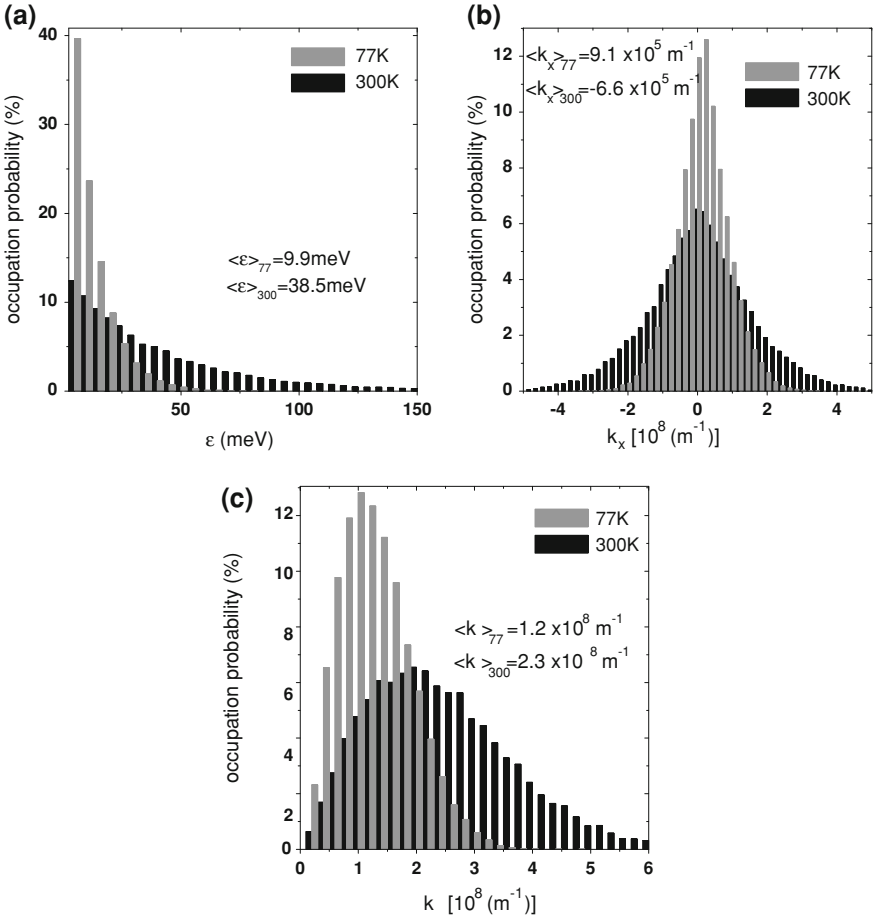


Fig. 5.9 Histograms of initial distribution of 30000 particles (**a**) of energy, (**b**) of $\mathbf{k}_x$ wavevector (parallel to the electric field) and (**c**) of $\mathbf{k} = (\mathbf{k}_x^2 + \mathbf{k}_y^2 + \mathbf{k}_z^2)^{\frac{1}{2}}$ for T=77K and T=300K. The average values are also indicated

$$P(t)dt = S_{total} e^{-\int_0^t S_{total} dt'} dt \tag{5.16}$$

In order to determine the free flight time one can produce a uniformly distributed random number r between 0 and 1 and try to evaluate t of 5.16 through the relationship

$$r = \int_0^t P(t)dt' \tag{5.17}$$

The problem in the evaluation of the integral is associated with the complex form that some of the scattering rates S_j, within S_{total} (5.7) may have. For example in polar scattering there is no analytic solution of 5.17 and the evaluation becomes very difficult. However, Rees [17] came up with a method that greatly simplifies this task by proposing a virtual quantity called "self-scattering" (SS). The SS, S_{self} changes neither the carrier's momentum ($\mathbf{k}' = \mathbf{k}$) nor its energy. All it does is that it adds up to the total scattering rate S_{total} in such a way that the sum of these two will be constant for all energies.

$$\Gamma = S_{total} + S_{self} \tag{5.18}$$

With this assumption 5.16 can be written as

$$P(t) = \Gamma e^{-\Gamma t} \tag{5.19}$$

and the free flight time t_r can be generated by using random numbers as

$$t_r = \frac{-\ln(1-r)}{\Gamma} \tag{5.20}$$

and because r is obtained from the uniform distributed between 0 and 1 we get

$$t_r = \frac{-\ln r}{\Gamma} \tag{5.21}$$

Computational efficiency

Figure 5.10 shows the concept of SS and how this can be implemented in practice to obtain a Γ which is used to derive the free flight time (constant Γ technique) (5.21). The shaded region is the SS rate and here we have taken $\Gamma = \max(S_{total})$. It is worth pointing out that the simplicity in deriving t_r comes at an expense of the computational time. The "wasted" computer time corresponds to the times that a SS event takes place. For example, as shown in Fig. 5.10, whilst at very small energies of few meV the code is efficient as the SS events are few, as the energy increases the SS events appear approximately 40–50% of the time. This means that the computational time has increased 2-fold to obtain the same result. This makes it clear that the choice of Γ is very important. To make things even more evident, imagine the case where ionised impurity scattering which dominates at low energies, is an order of magnitude higher because of the increase the number of the ionised impurities n_{II} [13]. This would result in a Γ being approximately 10 times bigger than S_{total} for almost all energies, meaning that a real scattering event would happen only once out of ten times, which makes the code highly inefficient.

It is possible to improve the efficiency of the code by introducing a variable Γ scheme [8, 9]. This is shown in Fig. 5.11. We take two Γ corresponding to two different energies.

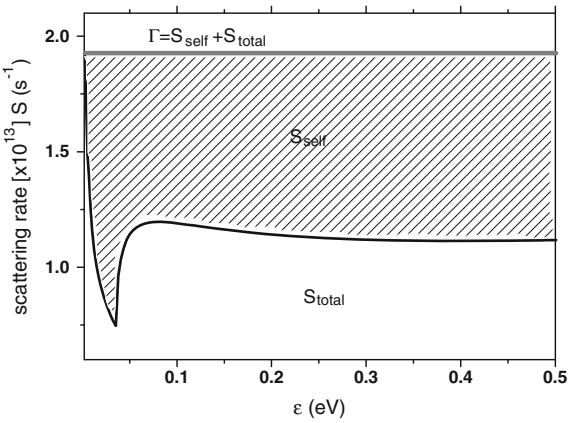


Fig. 5.10 Total scattering rate S_{total} and SS S_{self} to derive a constant scattering rate Γ for obtaining the free flight time

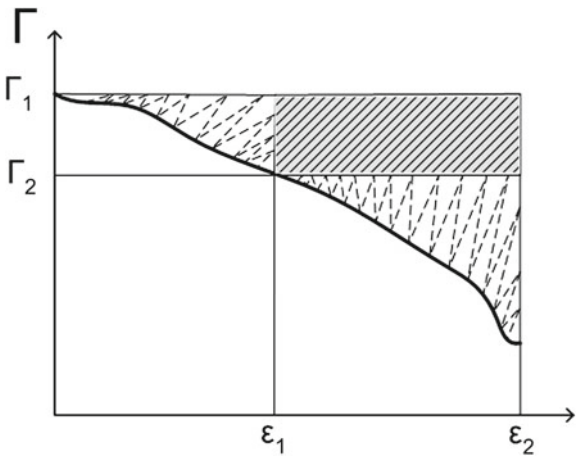


Fig. 5.11 Variable two-Γ scheme with two regions for increase of the computational efficiency. The grey shaded line corresponds to the saved computational cost of having a unique Γ. The same concept can be extended to an N-times Γ scheme

$$\Gamma = \begin{cases} \Gamma_1, \text{ if } & \epsilon < \epsilon_1 \\ \Gamma_2, \text{ if } & \epsilon < \epsilon_2 \end{cases} \tag{5.22}$$

By applying this technique the computation time is reduced by an amount which equals the grey shaded box in Fig. 5.11. It is straightforward when a carrier both at the beginning and at the end of the free flight is located within the same region (indicated with zigzag lines). However, it is more complicated when the carrier at the start of the free flight is below ϵ_1 and just before scattering it has an energy $\epsilon > \epsilon_1$. t_r will be modified from 5.21 as follows [8]:

$$t_r = -\ln r \Gamma_2 + t_d(1 - \frac{\Gamma_1}{\Gamma_2}) \tag{5.23}$$

where t_d is the time required by the carrier to reach energy ϵ_1 and which is easily monitored during the execution of the code. This equation can be further extended for an N-variable Γ technique. A detailed analysis with an evaluation of the "saved" CPU time can be found in Refs. [18, 19].

5.3.3.3 Choice of the Scattering Mechanism

After the free flight time has been defined we need to choose one scattering mechanism randomly that will terminate the electron's drifting [9, 20, 21]. The choice is done by using a set of functions Θ representing the successive summation of the fractional contribution Γ_j of each scattering rate

$$\Theta_n(\epsilon) = \frac{\sum_{j=1}^{n} S_j(\epsilon)}{\Gamma} \quad \text{for} \quad n = 1, 2, \ldots, N \tag{5.24}$$

where S_j is the energy dependent scattering rate of the jth process and N is the total number of processes. The scattering mechanism is chosen by generating a uniformly distributed random number r between 0 and 1 and comparing it to Θ_n. The nth scattering mechanism is chosen if

$$\Theta_{n-1}(\epsilon) < r < \Theta_n(\epsilon) \tag{5.25}$$

is satisfied. Let us take for example the scattering rates shown in Fig. 5.6. Figure 5.12(a) shows the probability of occurrence of each scattering process at ϵ=200meV. We remind that (1) is the acoustic deformation potential (ADP),(2) and (3) the polar optical phonon absorption (POPab) and emission (POPem) respectively, (4) the ionised impurity (II) and (5) the SS. In (b) the sum of the fractional contributions $\Theta_n(\epsilon)$ for the processes in (a) is shown. The shaded region corresponds to ϵ=200meV.

We can now choose a number $0 < r < 1$ and compare it with the $\Theta_n(\epsilon)$ to choose the scattering process.

- if r<0.01138 then acoustic deformation potential is chosen
- if 0.01138<r<0.06798 then POP absorption is chosen
- if 0.06798<r<0.27340 then POP emission is chosen
- if 0.27340<r<0.34961 then ionised impurity is chosen
- if 0.34961<r<1 then SS is chosen

The same thing is repeated for every energy of the carrier. This part of the code is a characteristic example of the core concept of the MC method, i.e. a combination of stochastic a deterministic processes. In other words, the scattering mechanism is chosen randomly, but because it maps on a deterministic input (weight factor of scattering process) it allows the steady-state estimators that will emerge in the

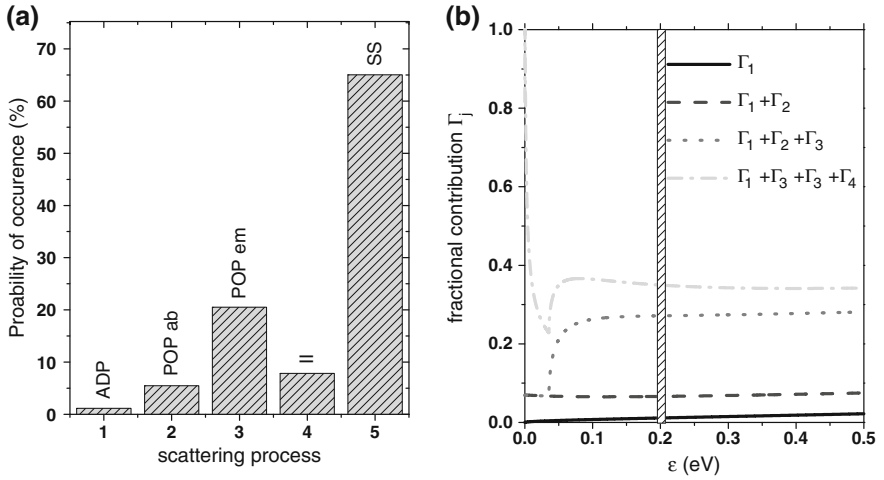


Fig. 5.12 **a** Fractional probability Γ_j for various scattering processes at an energy $\epsilon = 200meV$. **b** The sum of the fractional probabilities Γ_j as a function of energy. The shaded box corresponds to subfigure (a). For example for this energy by adding the histogram values $\Gamma_{1(ADP)} + \Gamma_{2(POPab)} + \Gamma_{3(POPem)} + \Gamma_{4(II)}$ in (**a**) we get a sum of ∼0.35 as shown in (**b**)

output of the MC code, to retain the physical characteristics of the system under investigation.

5.3.3.4 Selection of the Final State After Scattering

After the scattering mechanism has been chosen we must define the new state just before the next free flight starts. This post-scattering state $\mathbf{k}'$ corresponds to the end of the light arrows of Fig. 5.7(a). Unless the SS process has been selected, the new state will have a different magnitude and orientation. The new wavevector $\mathbf{k}'$ will be

$$\mathbf{k}' = \frac{\sqrt{2m_e^*}}{\hbar}\sqrt{\gamma(\epsilon)} \tag{5.26}$$

where $\gamma(\epsilon) = \epsilon'(1 + a_f\epsilon')$ and $\epsilon' = \epsilon + \Delta\epsilon$. For elastic scattering, $\Delta\epsilon = 0$ and for inelastic $\Delta\epsilon \neq 0$. Whether a process is elastic or not will depend on how the energy of the carrier is compared versus the energy of a phonon which is being absorbed or emitted. For example, acoustic deformation potential scattering (emission and absorption) can be taken as elastic at room temperature and inelastic at very low temperatures.

Also, scattering can be categorised as isotropic or non-isotropic. Isotropic means that there is an equal probability for a carrier to be scattered in all directions. For example, ADP scattering and intervalley phonon scattering are assumed to be isotropic.

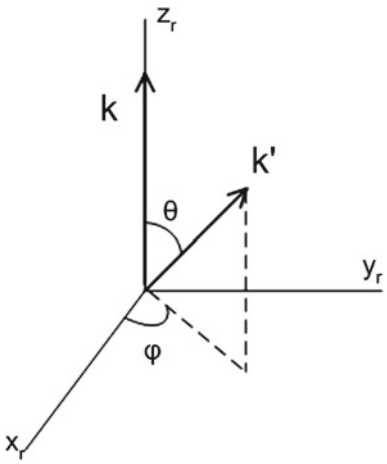


Fig. 5.13 Rotated coordinate system where the initial wavevector **k** is aligned with the vertical axis. θ and ϕ are polar and the azimuthal angles respectively

In such a case no coordinate system is required to obtain $\mathbf{k}'$. For the updated orientation the azimuthal angle ϕ and the polar angle θ are defined by two new random numbers r_1 and r_2 uniformly distributed between 0 and 1

$$\begin{aligned} \phi &= 2\pi r_1 \\ \cos\theta &= 1 - 2r_2 \end{aligned} \tag{5.27}$$

For anisotropic scattering such as Coulomb scattering or polar optical phonon scattering usually the coordinate system k_x, k_y, k_z is rotated by an angle ϕ about the $\hat{z}$ axis and then θ about the $\hat{y}$ axis. Scattering is performed in the rotated coordinate system $k_{x,r}, k_{y,r}, k_{z,r}$ and then transformation back to the original coordinate system is performed. The reason for performing scattering into the rotated coordinates is that the initial wavevector **k** can be set parallel to the $\hat{z}$ axis. This is shown in Fig. 5.13. For more details on how to perform the transformations from the initial coordinates to the rotated and then back to the initial ones can be found in the textbooks of Lundstrom [21] and Tomizawa [20]. The azimuthal and polar angles in anisotropic scattering will be

$$\begin{aligned} \phi &= 2\pi r_3 \\ \cos\theta &= 1 - \frac{2r_4}{1+4k^2(1-r_4)L_D^{-2}} \quad \text{for Coulomb scattering} \\ \cos\theta &= \frac{(1+\xi)-(1+2\xi)^{r4}}{\xi} \quad \text{for POP scattering} \end{aligned} \tag{5.28}$$

where r_3 and r_4 are random numbers again and L_D^{-1} is the screening length (inverse Debye length) and $\xi = \frac{2kk'}{(k-k')^2}$, where k' was given in 5.26.

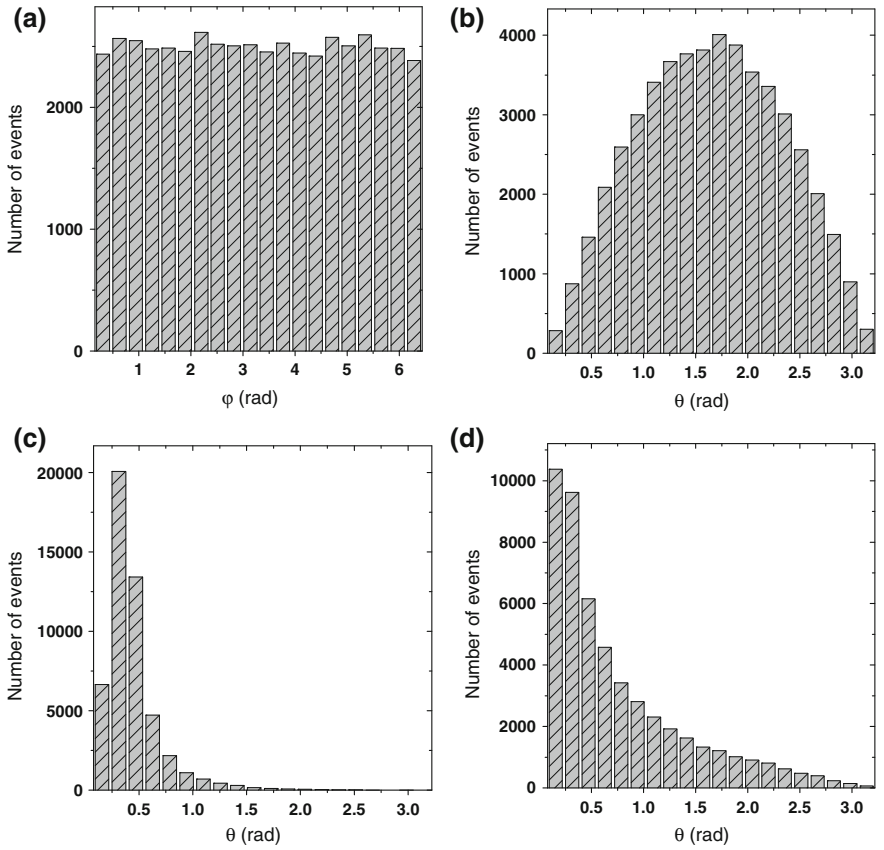


Fig. 5.14 Histograms for the distribution of scattering events in various angles (**a**) polar angle, (**b**) azimuthal angle for isotropic scattering, (**c**) azimuthal angle for anisotropic Coulombic scattering and (**d**) azimuthal angle for anisotropic polar optical phonon emission

Figure 5.14 shows the distribution of the azimuthal and polar angles for 50000 events for isotropic and anisotropic scattering. For polar angles (a) the probability distribution of the carrier within an angle of $\phi = [0, 2\pi]$ is uniform. For isotropic scattering (b) the azimuthal angle θ has more of a normal distribution within $[0, \pi]$ with a mean value at $\theta = \pi/2$ radians. For anisotropic scattering such as for Coulomb (c) or polar optical phonon scattering (d) there is a clear preference towards smaller scattering angles. For ionised scattering we have assumed here an impurity concentration of $n_{II} = 10^{16} cm^{-3}$ and the carrier having an energy ϵ=10meV. For polar optical phonon (d) ϵ=200meV and ϵ'=164.4meV after the emission of a phonon.

5.3.3.5 Calculation of Final Estimators

We have now reached the point that after consecutive accelerations and scattering events acting upon the carriers the criteria which have been set for steady state are satisfied. The final mean value of the estimators Q can be defined by summing the individual contributions at each subhistory [8]

$$\langle Q \rangle = \frac{1}{T} \sum_r \int_0^{t_r} Q(t')dt' \tag{5.29}$$

where t_r is the free flight duration and T is the total simulation time. The estimators that we are basically interested in is the average drift velocity v_d and the average energy ϵ of the carrier. The former within the free flight time t_i is given by

$$\langle v_d \rangle_{t_r} = \frac{1}{\hbar} \frac{\Delta\epsilon}{\Delta \mathbf{k}} \tag{5.30}$$

where $\Delta\epsilon$ and $\Delta\mathbf{k}$ are the infinitesimal energy and wavevector difference during the free flight time t_r. Equation 5.30 is valid because $v_d = \frac{1}{\hbar}\frac{\partial\epsilon}{\partial\mathbf{k}}$. Also, from 5.11 we have

$$\Delta \mathbf{k} = \frac{(-q)\mathbf{F}t_r}{\hbar} \tag{5.31}$$

Using 5.31 we can write 5.30 as

$$\langle v_d \rangle_{t_r} = \frac{\Delta\epsilon}{(-q)\mathbf{F}t_r} \tag{5.32}$$

and the average steady-state drift velocity is given by

$$\begin{aligned} \langle v_d \rangle &= \tfrac{1}{T} \textstyle\sum \langle v_d \rangle_{t_r} t_r \\ \langle v_d \rangle &= \tfrac{1}{(-q)\mathbf{F}T} \textstyle\sum \Delta\epsilon \\ \langle v_d \rangle &= \tfrac{1}{(-q)\mathbf{F}T} \textstyle\sum (\epsilon_f - \epsilon_i) \end{aligned} \tag{5.33}$$

where ϵ_i is the energy at the start and ϵ_f the energy at the end of the free flight. If $\epsilon_f > \epsilon_i$ then $\langle v_d \rangle$ is positive, since the negative sign in the previous equations stands for the electron charge. The average steady state energy is

$$\langle \epsilon \rangle = \frac{1}{T} \sum \langle \epsilon \rangle_{t_r} t_r \tag{5.34}$$

where $\langle \epsilon \rangle_{t_r} = \frac{\epsilon_i + \epsilon_f}{2}$

5.4 Ensemble Electron Monte Carlo

So far we have seen that single Monte Carlo method has been used to evaluate the steady state characteristics under a static and uniform electric field. However, it is often the case where carriers move balistically or semi-balistically in small devices. Ballistic transport means that carriers move from one physical end to the other end of a semiconductor within a time which is comparable to their relaxation time. In such a case carriers do not manage to obtain their steady-state and non-stationary carrier transport such as velocity overshoot takes place. Apart from the time-dependent phenomena, the study of the space-dependent phenomena requires an alternative approach to that of SMC as well. An example is the evaluation of the diffusion coefficient, where at high electric fields the Einstein relation does not describe sufficiently the diffusion of hot electrons [8].

In order to tackle these problems an EMC method is used [22–25]. Ensemble refers to the fact that a group of N particles is used for the simulation. For example, the dynamical response of electrons to an external voltage will typically require the simulation of a few thousand electrons for a very short space of time δT, typically three orders of magnitude smaller than the simulation time for steady state.

In the introduction of Sect. 5.3 we mentioned that the system we are looking at is ergodic. Indeed, by simulating an ensemble of carriers for a very large time we should be able to get the same results as obtained within the SMC.

5.4.1 Description of the Algorithm

Before the algorithm is given a description of the core part of the ensemble simulation is shown in Fig. 5.15. We assume n particles (n=1,2,3,..N) each one of which is being simulated for a total time T. The time increases to the right as shown by the arrow. The horizontal lines show each particle's trajectory versus time. Subsequently, we assume a fixed time step Δt. This is something we define as a in input to the code and is typically a few femtoseconds. The vertical lines interrupting the trajectories correspond to this time step. The solid circles represent random scattering events for each one of the particles which may or may not occur during one Δt. Therefore, the time between two successive solid circles is the free flight time t_r. Also, it is physical that not all scattering events will take place at the same time within one step, which is represented by the small offset in the circles within a Δt. It may also be possible that within the same step a carrier will be scattered more than once. The sequence is that every particle is simulated until the end of one time step and then the next one follows. Over each Δt the motion of a particle is assumed to be independent of the other carriers in the ensemble.

The average value of a quantity Q is defined as the ensemble average at an instantaneous time $t = t_i$ over the N particles of the ensemble

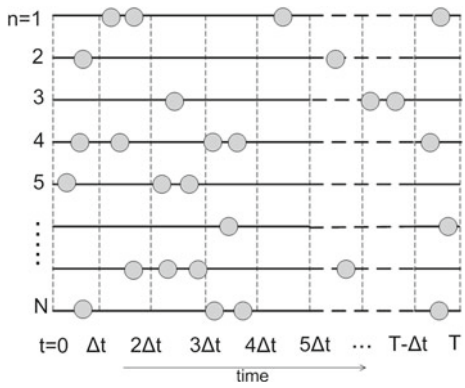


Fig. 5.15 Flowchart of the free flight time within an Ensemble Monte Carlo. The horizontal lines are the trajectories of the particles and the vertical lines are the times of observation. The circles are the scattering events taking place. Note that it is possible to have none, one or more than one scattering event within an observation period. The code is modified accordingly in each case

$$\langle Q \rangle_{t_i} = \frac{1}{N} \sum_{j=1}^{N} \langle Q \rangle_j (t = t_i) \tag{5.35}$$

We can choose $t_i = n\Delta t$ which corresponds to the fixed sampling time. Therefore, the average value of N particles will be

$$\langle Q \rangle_{n\Delta t} = \frac{1}{N} \sum_{j=1}^{N} \langle Q \rangle_j (n\Delta t) \tag{5.36}$$

The standard error s will be

$$s = \frac{\sigma}{\sqrt{N}} \tag{5.37}$$

where σ is the standard deviation with a variance σ^2 given by

$$\sigma^2 = \frac{N}{N-1} \{ \frac{1}{N} \sum_{j=1}^{N} (\langle Q \rangle_j)^2 - \langle Q \rangle^2 \} \tag{5.38}$$

This highlights the fact that when the ensemble size increases, there is always a trade-off between the "added" accuracy of the estimators and the "added" computational time.

In the description of the algorithm some of the steps or basic concepts such as the band structure of the material which acts as an input to the code, the scattering processes, or the selection method after scattering remain practically the same as in the SMC method.

1. *Input of "external" physical parameters and simulation characteristics*: Parameters such as electric field strength F, lattice temperature T, number of simulated particles N, time step of observation Δt, allowed maximum time of simulation t_{max} and other physical constants are defined. All values are deterministic.
2. *Definition of the physical system*: It is given deterministically and is similar to the corresponding part of SMC.
3. *Scattering parameters*: Also a deterministic input in the same way as described for SMC.
4. *Initial conditions of motion*: This step remains the same as well.
5. *Free flight time*: The free flight time is slightly differently from that defined in the SMC. At the beginning of the simulation the first free flight duration is derived from 5.21. This makes use of the maximum scattering rate $\Gamma = S_{j,\max}$ and of a uniformly distributed random number r between 0 and 1. This is repeated for all particles N. Also, the momentum vector $\mathbf{k}$ is known at this point.
6. *Time of scattering event-new free flight time*: Let us take a look at Fig. 5.15 and assume that a carrier is located at Δt. There are two cases depending on the duration of the free flight time t_r, i.e. whether it is bigger or smaller than the observation time interval Δt. Assuming that t_s is the scattering time (corresponding to the free flight time t_r), if $t_s > \Delta t$, then the particle drifts during Δt, whilst if $t_s < \Delta t$ the particle is scattered by some process (to be defined later) before Δt (in our example before $t = 2\Delta t$). In the second case the carrier's free flight time is $t_r = t_s - \Delta t$. In this case we must define a new scattering time t_s' and then we check again whether this is bigger or smaller than Δt.
7. *Drift process and collection of data for estimators*: These steps are the same as in SMC. The new wavevector is monitored.
8. *Selection of scattering process and determination of the post scattering state*: Also the same as in SMC.
9. *Logic check of the simulation time*: After scattering the time is increased by the observation time Δt. The code execution stops when the steady-state requirements we have set at the start of the program are met.

A flowchart with the basic steps of the EMC is given in Fig. 5.16. We note that the steps described above should be repeated for all particles N in the ensemble. Once this is done we can collect the average quantities in consideration of the whole ensemble according to 5.36.

5.5 An Example: Electron Motion in Bulk GaAs

In this section the SMC and EMC methods are applied for the case of GaAs. We shall assume the simplest case where two conduction band valleys are considered (Γ and L) (Fig. 5.17). Non-parabolicity in included in the standard form [9, 14–16]. A valley separation of $\Delta\epsilon_{\Gamma L}$=0.29eV is assumed. The scattering rates that have been considered are for the Γ valley: Polar optical phonon, acoustic deformation potential,

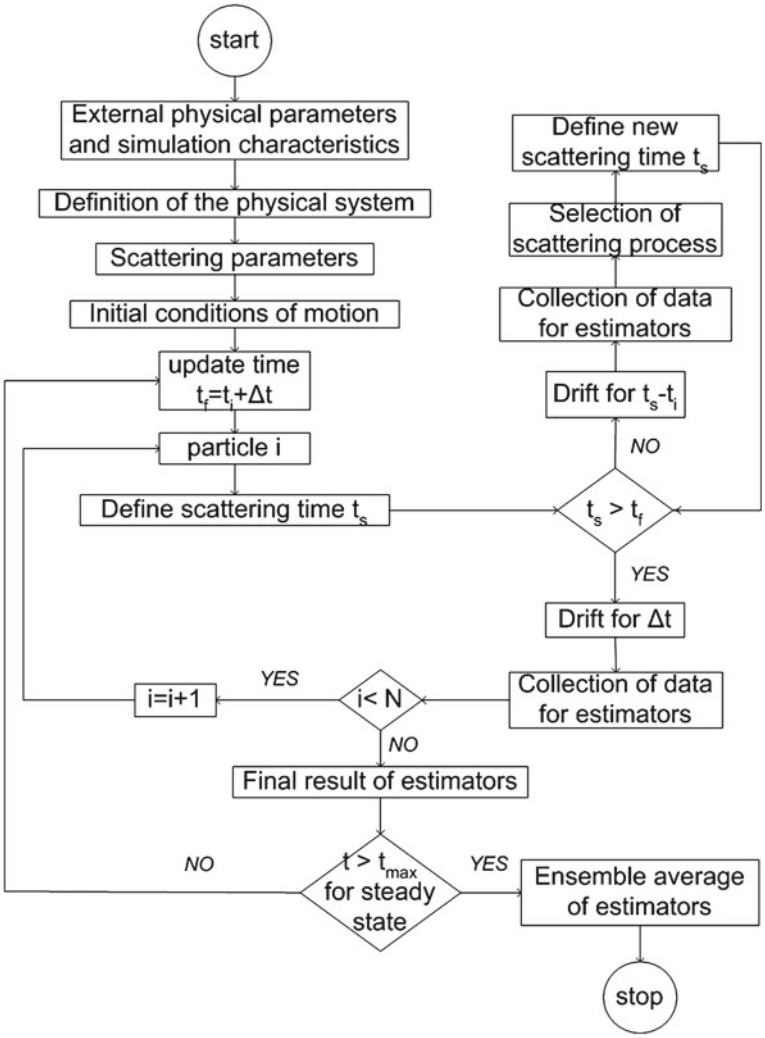


Fig. 5.16 Flowchart of the basic steps of the Ensemble Monte Carlo method

non-equivalent intervalley $\Gamma \rightarrow$L phonon scattering. For the L valley, all the previous ones plus the equivalent intervalley phonon scattering . Ionised impurity scattering is omitted. The formulation of the scattering rates is similar to that of Ref. [14].

We plot the steady-state characteristics which is the average drift velocity $\langle v_d \rangle$, the average energy $\langle \epsilon \rangle$ and the average occupation of the two valleys. To obtain these results we have used a simulation time of 1ns and 60 iterations. If we take a close look at the curves we can see that in some of them there is a "noise", which is associated with stability issues. In principle we can get smoother curves by increasing the simulation time (ex. to 2–3 ns) and/or the number of iterations (ex.100).

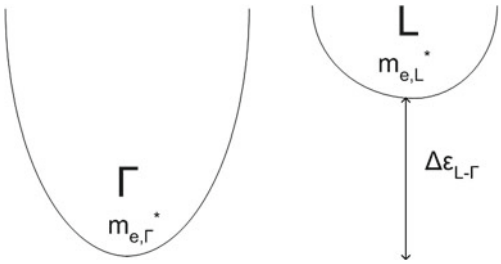


Fig. 5.17 Schematic conduction band structure of GaAs

The trade-off is the increased computational time which will be required. Therefore, this could be done only for regions where greater accuracy is necessary for example in getting the exact value of the overshoot electric field and the corresponding drift velocity.

For the Ensemble Monte Carlo the same assumptions as the band structure are made. Due to the small physical length of some heterojunction devices, non-thermally equilibrium carriers can be injected which do not have enough time to reach their steady state conditions. If the transit time is comparable to the relaxation time this can result in electrons travelling at higher velocities than at their steady-state. Physically, this is due to the nonequivalence of the energy relaxation time and the momentum relaxation time.

The increase of the transient velocity above the steady-state value is called overshoot and is shown in Fig. 5.18 [26]. We can see that the instantaneous velocity is bigger than the steady-state one as shown in Fig. 5.19(a). After a few picoseconds it tends to saturate towards the steady-state. The overshoot velocity depends on the applied voltage and the scattering processes. For example although the velocity $v_{d,max}$ for F=10kV/cm is bigger than the corresponding for 5kV/cm, the fact that more electrons in the first case reside in higher valleys where the energy relaxation time is smaller, results in a smaller steady state velocity. In (b) the average the distance x from the source is shown. The distance can also be extracted by monitoring the x, y and z components of the momentum $\mathbf{k}$.

Using bulk GaAs it has been shown how the solution for the electron distribution function can be derived without the need of having to tackle the complicated BTE. Both in the ensemble and in the single particle method, random numbers mapping on some deterministic process have used. We have seen how despite the randomness imposed, the physical system retains its signature characteristics.

From what has been described so far both SMC and EMC methods use a model as an input which in turn includes a description of the conduction band structure. But it has not been pointed out clearly yet what the limitations of this model are. Whilst for example it works well for simulating transport from low to high fields, it is expected to deviate more at extremely high fields. This problem is briefly discussed next.

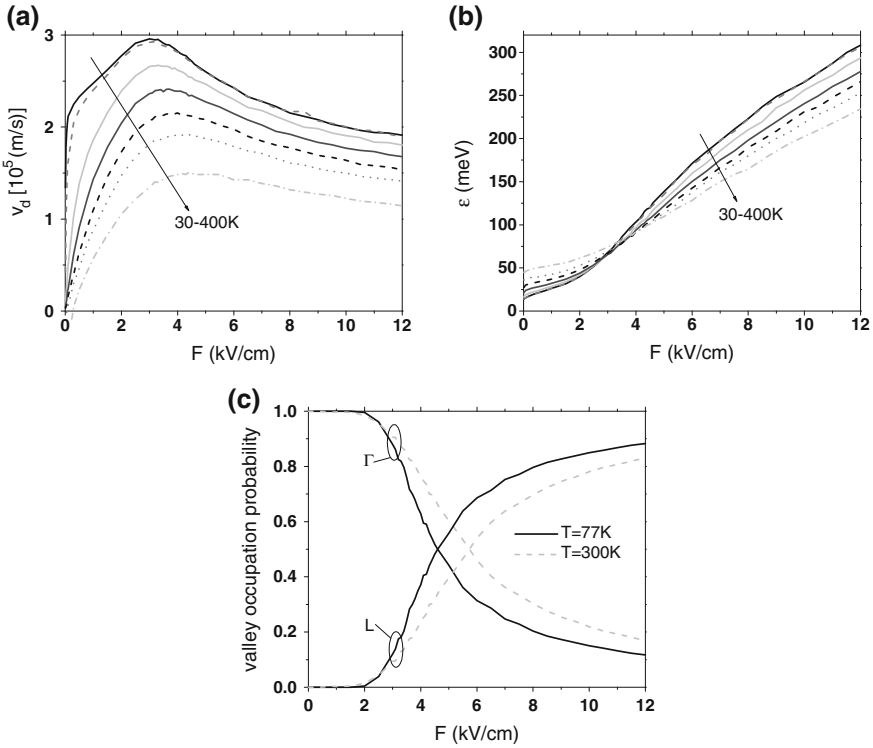


Fig. 5.18 Steady-state characteristics of the SMC in GaAs (**a**) average drift velocity v_d (**b**) average carrier energy ϵ and (**c**) average valley occupation versus applied electric field F for various temperatures

5.6 Monte Carlo Simulation at Very High Fields

It has been shown previously that the physical characteristics of the material under simulation are input in the beginning of the Monte Carlo code. Specifically, the conduction band structure is of importance because the scattering rates are based on it. For very low field transport, where carriers reside low in the conduction band we can assume that the latter has a parabolic shape. For moderate and high fields where electrons become hot, the concept of non-parabolicity must be included.

In the same way, for extremely high fields electrons acquire higher energies. Then the issue of whether the bands can be still approximated by a first order $\mathbf{k} \cdot \mathbf{p}$ non-parabolic type of approach becomes important. The short answer is that for very high fields a different treatment which will account for the full band is necessary.

This can be understood by looking at the scattering rates S. For example, phonon scattering is proportional to the density of states in the system in a linear fashion. Figure 5.20 shows the density of states $N(\epsilon)$ in Si calculated in two ways [27]. Similar trends apply for GaAs. The solid line corresponds to empirical-pseudopotential

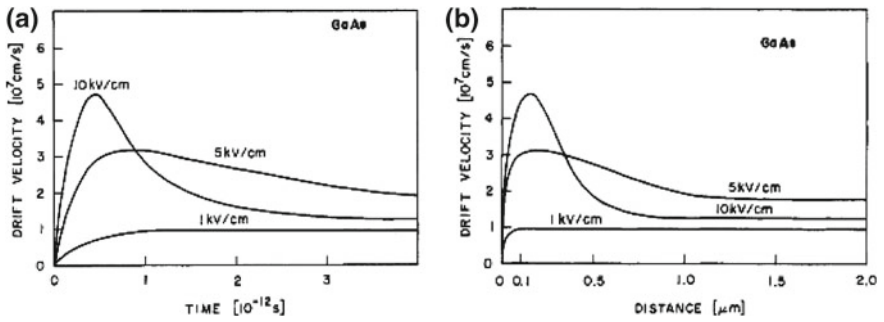


Fig. 5.19 Steady-state characteristics of the ensemble Monte Carlo in GaAs (**a**) average velocity v_d (**b**) average distance versus time for various applied fields (Reprinted figures with permission from Ref. [26], © 1972 IEEE)

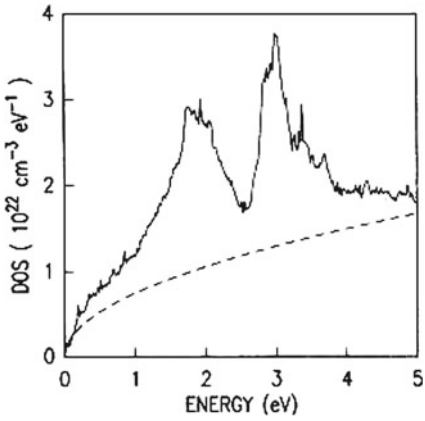


Fig. 5.20 Density of states of Si using empricial-pseudopotential calculations (*solid line*) and six ellipsoidal parabolic bands (*dashed lines*). The difference between the two approaches will have an impact for very high fields (Reprinted figure with permission from Ref. [27], © 1988 American Physical Society)

calculations accounting for the whole band structure, whilst the dashed line uses six ellipsoidal parabolic bands. A good agreement between the two exists only for very low carrier energies, which correspond to low-moderate electric fields. As energy increases there is an increasing difference between the two methods which is related to electrons moving under the influence of very high fields. This will have an impact on the quantities which are being calculated using the Monte Carlo technique. For example, Shichijo and Hess [28] state that for some valleys the method based on the effective mass m_e^* and the nonparabolicity starts to break down for energies above approximately 1eV.

The MC simulation method in the full band approach is different from the previous treatments; first, the band structure in the whole Brillouin zone is defined. For a given wavevector $\mathbf{k}$ the energy ϵ is derived by a quadratic interpolation utilising the energies and the gradients of the surrounding mesh points of the cubic element in the case of GaAs for example. Next, the scattering rates are evaluated using Fermi's golden rule, but the scattering extends over all bands ν and all wavevectors $\mathbf{q}$ in the first

Brillouin zone. We shall not give here the exact formulations of the phonon scattering rates, but the reader is advised to look for a detailed analysis at the work by Fischetti [27, 29], Hess et al. [28] and Bulutuay et al. [30, 31] in both zinc-blende and wurtzite semiconductors. For the selection of the final state an analytical way as was done earlier for simpler bands is not possible and instead a complicated process which involves a search in the whole Brillouin zone must be implemented. In order for the energy and the momentum to be conserved, some cubes are selected centred around $\mathbf{k}'$ that intersect the surfaces $\epsilon_\nu(\mathbf{k}') = \epsilon'$. Then the momentum difference between each possible final state and the initial momentum is computed. Subsequently, cubes that satisfy these conditions are chosen and their density of states is calculated. Then all densities are added up and a random number is used to choose one of them.

For a detailed description of how the full band structure can be incorporated in the MC code as well as how various other effects such as impact ionization rate can be included the IBM up-to-date simulator DAMOCLES is an excellent source [32].

5.7 Electron Transport in Dilute Nitrides

Dilute nitride semiconductors is a relatively new class of materials where a nitrogen atom has the tendency to substitute an isoelectronic atom of group V of the periodic table in a conventional III–V compound semiconductor. Even small amounts of nitrogen produce dramatic changes in the electronic properties. This is because nitrogen is a small atom with high electronegativity that perturbs strongly the host crystal structure. Results from nitrogen incorporation is the large reduction of the energy bandgap [33–35] which is much larger than the change observed when alloying conventional III–V semiconductors with other elements. In fact, N perturbs strongly only the conduction band leaving the valence band unaffected. Also, a significant increase in the electron effective mass [36–38] and a decrease in the mobility [39–44] is observed.

Figure 5.21(a) shows what happens when nitrogen is added in GaAs. The localised N state formed above the parabolic conduction band of GaAs interacts with it, resulting in its splitting and the formation of two mixed subbands ϵ_1 and ϵ_2 and the reduction of the energy bandgap. Here, we will focus on carrier transport of bulk GaN_xAs_{1-x}. Mobility in this material has been calculated by using the BTE [42, 43]. This can also be done by employing the MC method. However, GaAsN is a more challenging material system than the conventional GaAs with regard to the MC methodology. This is because of the unusual conduction band structure, namely a prominent non-parabolicity which is added to the standard non-parabolicity a_f described earlier in the GaAs section. This extreme non-parabolicity which is prominent at high $\mathbf{k}$ wavevectors (Fig. 5.21(a)) is associated with the strong mixing between the delocalised GaAs host states and the localised N-impurity. As long as the electrons are located away from this region we can still use the standard expression of non-parabolicity [9, 14–16]. For extremely high fields this assumptions should start breaking down.

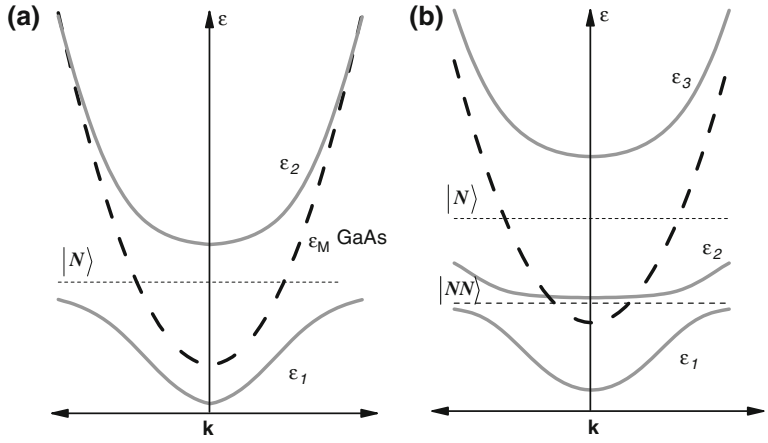


Fig. 5.21 **a** Single N state interacts with the GaAs band resulting in two mixed states ϵ_1 and ϵ_2. This is known as 2-band anticrossing (BAC) **b** Single and pair N state interact with GaAs yielding 3 mixed states, known as 3 BAC

The other unusual characteristic in dilute nitrides is that in the description of the electron motion we should include a mechanism that will account for the transfer of electrons from the lower conduction band ϵ_1 to the higher ϵ_2. Also, the carriers should be allowed to perform the inverse process, i.e. relax to a lower energy state. This could happen by assuming that electrons hop on the localised nitrogen state and then back off to the conduction band states. Because nitrogen is highly localised in real space it is delocalised in the **k**-space, meaning that scattering can take place in the momentum space within an allowed extent of **k** vectors and then by absorption or emission of a phonon transfer to a higher or lower subband [45, 46]. This could be used in junction with the energy broadening of the mixed subbands by employing a complex band structure scheme and the use of Green's functions [43, 47–49].

The situation becomes even more complicated for higher N concentration (typically N>0.4%) where apart from the single N state, formation of higher order nitrogen clusters (pairs, triples) takes place [50–52]. Figure 5.21(b) shows what happens when a nitrogen pair is present. These localised states mix with the GaAs band as well, producing a band structure that deviates even more from GaAs. The shape of this band structure should be appropriately accounted for within the MC algorithm.

5.7.1 Single Electron Monte Carlo in GaAsN

As opposed to the GaAs system the additional scattering process from the localised N state must be included in GaAsN. This can be incorporated in two ways which correspond to the two models shown in Fig. 5.22. In Model 1, nitrogen scattering is explicitly included as a separate scattering process and is given by [42, 43]

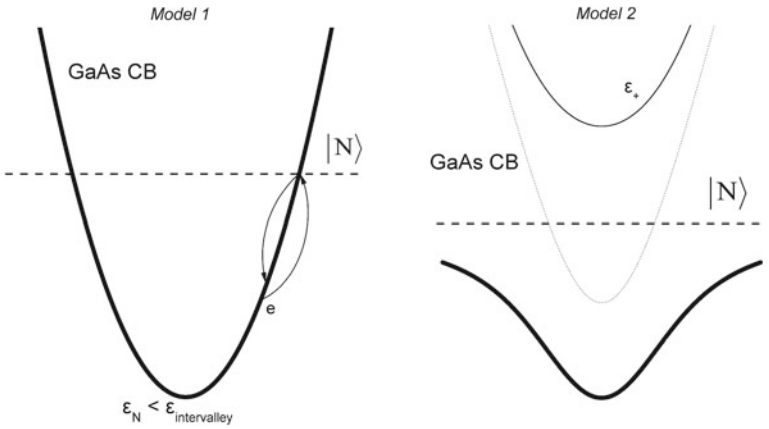


Fig. 5.22 Generic schematic of electron motion in GaAsN using Model 1, with a GaAs conduction band and a localised N state which scatters carriers and Model 2 where the role of nitrogen is manifested through the altered electron mass and the perturbed band structure

$$S_N(\epsilon) = \frac{\pi \alpha_0^3}{2\hbar} \frac{\beta_N^4 x}{(\epsilon - \epsilon_N)^2 + \Delta_N^2} \tag{5.39}$$

where β_N is the coupling constant between localised and delocalised states, x is the concentration of nitrogen, ϵ_N is the energy level of the nitrogen localised state, Δ_N the broadening of the state and α_0 the lattice constant. The electron mass m_e^* required in the MC algorithm is that of GaAs.

In Model 2, the effect of nitrogen is manifested by the altered m_e^* of GaAsN and the band dispersion of the mixed conduction band. The average drift velocity $\langle v_d \rangle$ exhibits Negative Differential Velocity (NDV) characteristic behaviour for $N = 0.1\%$ and saturation (Si-like behaviour) for higher concentrations [53]. Also, the average electron energy $\langle \epsilon \rangle$ at the overshoot point is found to be low and close to the conduction band minimum, something which allows us to use the standard form of non-parabolicity at least for these range of electric fields [45].

Model 1 is somehow easier to deal with due to the simpler description of the conduction band structure. Again, it gives characteristic NDV behavior as well as good agreement with experiment on low field mobility for up to $N = 0.4\%$ [54]. Here, nitrogen scattering is assumed elastic and isotropic. Therefore, the post scattering state will be given by 5.27.

5.7.1.1 Comparison with GaAs

Comparison of the steady-state characteristics of ultra-dilute GaAsN alloys with GaAs shows the detrimental effect that the addition of nitrogen has even at tiny fractions. Figure 5.23(a) shows the drift velocity versus the applied field for GaAs

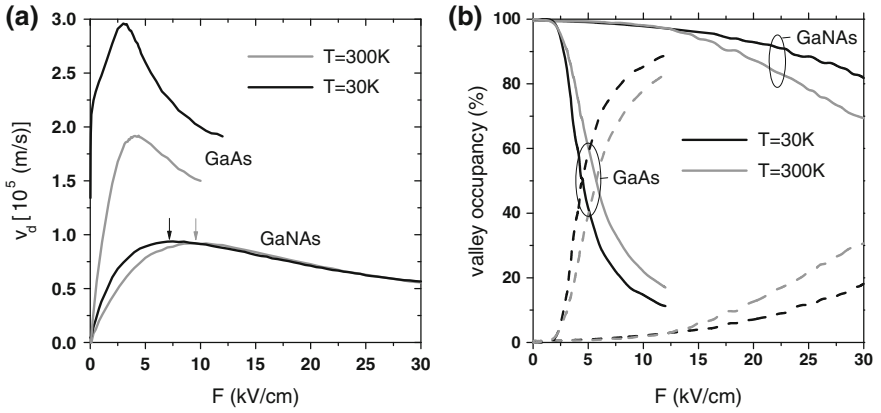


Fig. 5.23 **a** Average drift velocity versus applied electric field for GaAs and GaAsN for T=30K and 300K. The arrows indicate the field F required for overshoot **b** valley occupation for Γ (*solid*) and L valley (*dashed lines*) for GaAs and GaAsN for two temperatures. N impurity level acts as a barrier that scatters electrons strongly preventing them from moving higher in energy and transferring to the L valley

and $GaAs_{0.999}N_{0.001}$ for T=30K and 300K. In (b) the valley occupancy is shown (solid for Γ and dashed for L valley). The scattering processes considered in Γ valley are polar optical phonon, acoustic deformation potential, intervalley $\Gamma \rightarrow$L non-equivalent phonon and N scattering. For the L band similar to the previous excluding N scattering and adding intravalley scattering. Model 1 has been used for GaAsN.

From Fig. 5.23(a) we can see that in GaAsN the drift velocity decreases strongly compared to GaAs, whilst the electric field required to get the overshoot shifts towards higher values. Also, GaAsN shows a stronger temperature dependence something which has been observed experimentally as well [44, 55]. Details on the origin of the NDV for very small concentrations of nitrogen within GaAsN will not be given here, but we should say that it is a different effect from than the intervalley transfer observed in GaAs and is associated with the role of nitrogen in this material.

In conclusion, transport in dilute nitrides from the Monte Carlo point of view remains in a general context similar to the standard methodology employed earlier both for SMC and EMC. The difference is that the unusual conduction band structure needs to be accounted for in the code. Also, N scattering according to 5.39 should be added to the group of the scattering processes.

5.8 Quantum Monte Carlo

In Sects. 5.3 and 5.4.1, in the description of the SMC and EMC we said that although an electron may be in a bath of other electrons, it will drift under an external bias independently from the other electrons, which greatly simplifies calculations.

The same assumption is made within the simplest form of the Schrödinger wave equation which gives the energy eigenstates within a potential V.

In the most general case the Schrödinger equation accounts for a fully-interacting *many-electron* system given by [56]

$$H\Psi = -\sum_{i=1}^{N}\left(\frac{\hbar^2}{2m_e^*}\nabla_i^2 + \sum_a \frac{Z_a q^2}{4\pi\varepsilon_0|\mathbf{r}_i - \mathbf{d}_a|}\right)\Psi + \sum_{j>i}\frac{q^2}{4\pi\varepsilon_0|\mathbf{r}_i - \mathbf{r}_j|}\Psi = \epsilon\Psi \tag{5.40}$$

where Ψ is the N-electron wavefunction, $\mathbf{r}_i$ and $\mathbf{d}_i$ are the positions of the electrons and atoms respectively, and Z_a are the ionic charges, q the electron charge and ε_0 the vacuum permittivity. In the above equation the kinetic energy from the nuclei is neglected. Equation 5.40 is impossible to solve exactly.

This many-body problem can be simplified by assuming N electron equations, in which each electron moves within a *mean field potential* V which carries the signature of all other electrons which are present in the system. These single electron equations are

$$H\phi_i = -\frac{\hbar^2}{2m_e^*}\nabla_i^2\phi_i + V_i\phi_i = \epsilon_i\phi_i \tag{5.41}$$

where $i = 1, 2, \ldots, N$ and ϕ_i is the single electron wavefunction. To simplify things more, we can assume that the electrons are not interacting with each other, therefore the N-electron wavefunction can be written as

$$\Psi = \phi_1\phi_2 \ldots \phi_N \tag{5.42}$$

This is known as the *Hartree* approximation. The *Hartree-Fock* approximation extends the Hartree approximation to include the exchange interaction between electrons based on the Pauli's exclusion principle, which states that no two electrons with the same spin can occupy the same state simultaneously and that two electrons cannot have the same set of quantum numbers.

The target of all the previous equations is to find the ground state energy ϵ_0 of the system by a method based on the *variational principle*, which uses some trial wavefunctions Ψ which would ideally be equal to the ground state wavefunction Ψ_0.

Apart from the trial function method that was just described, there is another candidate theory which tries to tackle the many-body problem. This is the *density functional theory* (DFT). It is an exact theory and unlike the trial method which attempts to find the many-body wavefunction, it is the single electron charge density ρ which is the fundamental quantity. We shall not go in further detail in describing this theory and the improvements that have been made (Local Density Approximation, Generalized Gradient Approximation) to address the many-body problem. We will just point out that DFT gives exact solutions for many solids but fails to do so for some other materials.

This is the reason that another way of tackling the solution of 5.40 was suggested, which is based on a statistical approach, without having to reduce the

many-body problem to a set of single-particle simulations. This is the *Quantum Monte Carlo* (QMC) method. Essentially quantum Monte Carlo tries to solve the many particle equations without the approximations of a mean field V we described earlier. There are two basic versions: the *variational* and the *diffusion* QMC. The variational as indicated by the name, uses a set of trial wavefunctions and tries to solve some high-dimensional integrals. Choosing an optimized trial wavefunction is crucial for getting a wavefunction which is close to the ground state wavefunction Ψ_0. The diffusion QMC uses Green's functions to solve the many electron equations and in principle it is an exact method and in this case *importance sampling* is essential to make the simulation efficient. It is easier to apply this method to Bosonic systems than to Fermionic ones (and especially large Fermionic systems) and this constraint is associated with the ground state wavefunction of the fermonic system. A way to address this issue is by using a technique known as the *fixed-node approximation*.

Quantum Monte Carlo is by itself a huge area of research where various optimisation techniques and variations can be found. The aim of this section is to present briefly what it is, which problem it tries to solve and how it compares with the most important of the other existing methods of condensed matter physics. For more information above QMC, with references on the details of its development, on its statistical foundations, on its applications and on how it relates to the other trial methods the reader can look at Ref. [57–60]

5.9 Appendix: Random and Pseudorandom Numbers

In this chapter it has been pointed out that Monte Carlo is based on the generation of random numbers. We can now reveal that these numbers are not really as random as we may think. By random we typically think of an experiment equivalent to throwing a dice or collecting the numbers from a lottery. But is it possible for a computer to generate such random numbers for our simulations?

The answer is no, because computers typically use pseudorandom numbers. A pseudorandom number mimics the behavior of a random one and a pseudorandom number generator (PRNG) is an algorithm that uses a mathematical formula or precalculated tables to produce series of numbers which appear to be random. A really good PRNG can produce sequences of numbers with a long period which appear to be completely random.

What is then a truly random number and why do we not use one such for computer simulations? To do this we would need a truly random number generator (TRNG) connected to a computer. The randomness within a TRNG is based on physical phenomena that are completely unpredictable and aperiodic, such as the exact time of decay of a radioactive source, the atmospheric noise or the thermal noise. The problem with TRNGs is that they are very inefficient. One cannot produce a large quantity of random numbers which is necessary in many applications. Moreover, they are nondeterministic which makes them bad for simulations, because the same set of numbers cannot be reproduced (unless this happens by chance).

On the other hand, PRNGs are extremely efficient and deterministic. For an MC simulation where many random numbers need to be generated, efficiency is important, whilst being deterministic is good for testing or debugging because by having a fixed input we know the expected output. Also, PRNGs are periodic and at some point they will repeat, but the period is so long that practically PRNGs can be as good as TRNGs.

Therefore, it makes sense that the quality of a number generator will be important for any simulation. In the early years von Neumann used the *middle square method*. Later, PRNGs were based on the *linear congruential method* (LCM) due to its speed. These generators are still very popular. Other congruential generators are the *inverse* and *implicit* ones. An improvement of the LCG is the additive or multiplicative *lagged Fibonacci generator* (LFG), which is based on the generalisation of the Fibonacci sequence. However, for MC applications or any other study where a high quality of randomness is critical, especially the linear congruential and to a smaller extent the lagged Fibonacci generator, are not suitable. Instead, the *generalised feedback shift register* (GFSR) generator is used due to its long period and statistical randomness. Finally, so as to check the nonrandomness of any generator we can use some testing methods such as the *chi-square test* or the *Kolmogorov-Smirnov* test.

From this discussion it becomes clear that the selection of a PRNG is a non-trivial task and should not be treated as a black box especially when built in a programming language. The validity of the results of any simulation depends heavily on how good, reliable and suitable the generator is for the type of problem we are interested in. Modern softwares use their own PRNGs and it is worth spending some time to understand how they work and what their inherent limitations are. Sometimes it may be even useful to use a couple of different PRNGs to test the validity of the results. For a description of the TRNGs and PRNGs with applications and practical examples the reader is advised to check the online source of Ref. [61]. Also, for a comprehensive introduction to random number generation and their statistical testing as well as for a description of the uniform and non-uniform distributions used in MC Refs. [62, 63] are excellent sources.

References

1. Wolfram Mathworld: http://mathworld.wolfram.com/BuffonsNeedleProblem.html. Cited 17 June 2011
2. LANL-Histroy-People-Staff Biographies: http://www.lanl.gov/history/people/S_Ulam.shtml. Cited 2 April 2012
3. N. Metropolis, S. Ulam, J. Am. Stat. Assoc. **44**, 335 (1946)
4. R.W. Shonkwiler, F. Mendivil, in *Explorations in Monte Carlo Methods* (Springer Science+Business Media, Dordrecht, 2009)
5. P.N. Butcher, W. Fawcett, Phys. Lett. **21**, 489 (1966)
6. H. Budd, in *Proceedings of the International Conference on the Physics of Semiconductors*, 1967, Kyoto. J. Phys. Soc. Jpn Suppl. vol. 21, p. 420
7. T. Kurosawa, in *Proceedings of the International Conference on the Physics of Semiconductors*, 1967, Kyoto. J. Phys. Soc. Jpn Suppl. vol. 21, p. 464

8. C. Jacoboni, L. Reggiani, Rev. Mod. Phys. **55**, 645 (1983)
9. C. Jacoboni, P. Lugli, in *The Monte-Carlo Method for Semiconductor Device Simulation* (Springer-Verlag, Wien, 1989)
10. P.J. Price, in *Monte Carlo Calculation of Electron Transport in Solids*, ed. by R.K. Willardson and A.C. Beer, Vol. 14 of Semiconductors and Semimetals (Academic, New York, 1979), p. 249
11. A.D. Boardman, in *Computer Simulation of Hot Electron Behavior in Semiconductors Using Monte Carlo Methods*, Chap. 11 in Physics Programms (John Wiley, New York, 1980)
12. B.K. Ridley, *Quantum Processes in Semiconductors* (Oxford University Press, New York, 1988)
13. B.K. Ridley, *Electrons and Phonons in Semiconductor Multilayers* (Cambridge University Press, New York, 1997)
14. W. Fawcett, A.D. Boardman, S. Swain, J. Phys. Chem. Solids **70**, 1963 (1970)
15. E.M. Conwell, M.O. Vassell, Phys. Review **166**, 797 (1968)
16. H. Ehrenreich, Phys. Review **120**, 1951 (1960)
17. H.D. Rees, J. Phys. Chem. Solids **300**, 643 (1969)
18. E. Sangiorgi, B. Ricco, F. Venturi, IEEE Trans. Comput. Aided Des. **7**, 259 (1988)
19. R. Yorston, J. Comput. Phys. **64**, 177 (1986)
20. K. Tomizawa, *in Numerical Simulation of Submicron Semiconductor Devices* (Artech House Publishers, Cambridge, 2000)
21. M. Lundstrom, in *Fundamentals of Carrier Transport* (Cambridge University Press, Boston, 1993)
22. P. Price, IBM J. Res. Develop. **14**, 12 (1970)
23. K. Yokoyama, M. Tomizawa, A. Yoshii, IEEE Electron. Dev. Lett. **6**, 536 (1985)
24. D.K. Ferry, Phys. Lett. A **78**, 379 (1980)
25. W. Fawcett, in *Electrons in Crystalline Solids*, ed. by A. Salam (IAEA, Vienna, 1973), p. 531
26. J.G. Ruch, IEEE Trans. Electron. Dev. **19**, 652 (1972)
27. M.V. Fischetti, Phys. Rev. B **38**, 9721 (1988)
28. H. Shichijo, K. Hess, Phys. Rev. B **23**, 4197 (1981)
29. M.V. Fischetti, IEEE Trans. Electron. Dev. **38**, 634 (1991)
30. C. Bulutay, B.K. Ridley, N.A. Zahleniuk, Phys. Rev B **62**, 15754 (2000)
31. C. Bulutay, B.K. Ridley, N.A. Zahleniuk, Phys. Rev B **68**, 115205 (2003)
32. DAMOCLES: Monte Carlo simulation of semiconductor devices, http://www.research.ibm.com/DAMOCLES/. Cited 17 June 2011
33. M. Weyers, M. Sato, H. Ando, Jpn. J. Appl. Phys. **31**, L853 (1992)
34. M. Kondow, K. Uomi, K. Hosomi, T. Mozume, Jpn. J. Appl. Phys. **33**, L1056 (1994)
35. W. Shan, W. Walukiewicz, J.W. Ager III, E.E. Haller, J.F. Geisz, D.J. Friedman, J.M. Olson, S.R. Kurtz, Phys. Rev. Lett. **82**, 1221 (1999)
36. F. Masia, G. Pettinari, A. Polimeni, M. Felici, A. Miriametro, M. Capizzi, A. Lindsay, S.B. Healy, E.P. O'Reilly, A. Cristofoli, G. Bais, M. Piccin, S. Rubini, F. Martelli, A. Franciosi, P.J. Klar, K. Volz, W. Stolz, Phys. Rev. B. **73**, 073201 (2006)
37. E.P. O'Reilly, A. Lindsay, S. Fahy, J. Phys. Condens. Matter, 16, S3257, (2004)
38. A. Lindsay, E.P. O'Reilly, Phys. Rev. Letters **93**, 196402 (2004)
39. C. Skierbiszewski, Semicond. Sci. Technol. **17**, 803 (2002)
40. S.R. Kurtz, A.A. Allerman, C.H. Seager, R.M. Sieg, E.D. Jones, Appl. Phys. Lett. **77**, 400 (2000)
41. S. Fahy, O'Reilly. Appl. Phys. Lett. **83**, 3731 (2003)
42. S. Fahy, A. Lindsay, H. Ouerdane, E.P. O'Reilly, Phys. Rev. B **74**, 035203 (2006)
43. M.P. Vaughan, B.K. Ridley, Phys. Rev. B **75**, 195205 (2007)
44. A. Patanè, G. Allison, L. Eaves, M. Hopkinson, G. Hill, A. Ignatov, J. Phys. Condens. Matter, 21, 174209, (2009)
45. N. Vogiatzis, J.M. Rorison, J. Appl. Phys. **109**, 083720 (2011)
46. M. Seifikar, E.P. O'Reilly, S. Fahy, Phys. Status Solidu B **248**, 1176 (2011)
47. J. Wu, W. Walukiewicz, E.E. Haller, Phys. Rev. B **65**, 233210 (2011)

48. N. Vogiatzis, J.M. Rorison, J. Phys. Condens. Matter, **21**, 255801, (2009)
49. G.D. Mahan, in *Many-Particle Physics* (Plenum Press, New York, 1990)
50. P.R.C. Kent, A. Zunger, Appl. Phys. Lett. **79**, 2339 (2001)
51. P.R.C. Kent, L. Bellaiche, A. Zunger, Semicond. Sci. Technol. **17**, 851 (2002)
52. A. Patanè, J. Endicott, J. Ibez, P.N. Brunkov, L. Eaves, S.B. Healy, A. Lindsay, E.P. O'Reilly, M. Hopkinson, Phys. Rev. B **71**, 195307 (2005)
53. Y. Sun, M.P. Vaughan, A. Agarwal, M. Yilmaz, B. Ulug, A. Ulug, N. Balkan, M. Sopanen, O. Reentilä, M. Mattila, C. Fontaine, A. Arnoult, Phys. Rev. B **75**, 205316 (2007)
54. N. Vogiatzis, J.M. Rorison, Phys. Stat. Solidi B **248**, 1183 (2011)
55. S. Spasov, G. Allison, A. Patanè, L. Eaves, MYu. Tretyakov, A. Ignatov, M. Hopkinson, G. Hill, Appl. Phys. Lett. **93**, 022111 (2008)
56. N.W. Ashcroft, N.D. Mermin in textitSolid State Physics, vol. 2, Seminumerical Algorithms, (Harcourt College Publishers, Fort Worth, 1976)
57. A.J. James, Dissertation, Imperial College, (1995) http://www.imperial.ac.uk/research/cmth/research/theses/A.J.James.pdf. Cited 2 April 2012
58. M.L. Stedman, Dissertation, Imperial College, (1999) http://www.imperial.ac.uk/research/cmth/research/theses/M.L.Stedman.pdf. Cited 2 April 2012
59. R. Gaudoin, Dissertation, Imperial College, (1999) http://www.imperial.ac.uk/research/cmth/research/theses/R.Gaudoin.pdf. Cited 2 April 2012
60. (Quantum Monte Carlo and the CASINO Program) Available via University of Cambridge. http://www.tcm.phy.cam.ac.uk/~mdt26/casino2_introduction.html. Cited 2 April 2012
61. Random.org: True random number service, http://www.random.org. Cited 2 April 2012
62. D.E. Knuth, in *The Art of Computer Programming*, Seminumerical Algorithms, vol. 2 (Addison Wesley, Reading, 1998)
63. S. Tezuka, in *Uniform Random Numbers: Theory and Practice* (Kluwer Academic Publishers, Boston, 1995)

Chapter 6
Band Structure Engineering of Semiconductor Devices for Optical Telecommunications

Hélène Carrère and Xavier Marie

Abstract This chapter aims to provide an introduction to the main principles of band structure engineering of semiconductor devices. We show that it is possible to modify artificially the electronic structure of semiconductor materials. The combination of strain and quantum confinement can in particular lead to great improvements of the semiconductor laser characteristics. This explains that most of the commercial semiconductor lasers and semiconductor optical amplifiers for optical telecommunications (1.3 and 1.55 μm) are based on strained quantum wells.

6.1 Basics of Band Structure Engineering

We present in this section the main principles of band structure engineering which has been applied successfully to optimise the performances of semiconductor devices for optical telecommunications. In order to get more details the reader can refer to the excellent review papers written by Yablonovitch et al. [1], O'Reilly et al. [2] and Thijs et al. [3].

With the great progress in epitaxial growth it is possible to modify artificially the electronic structure of semiconductor materials. We must regard the natural electronic band structure of the semiconductor crystals (energy gap, carrier mass, etc.) as a starting point for application in devices. We will show that the combination of strain and quantum confinement can lead to great improvements of the semiconductor device characteristics (for a laser: lower threshold current, better quantum efficiency,

H. Carrère · X. Marie (✉)
Laboratoire de physique et Chimie des Nano-Objets,
INSA-CNRS-UPS Université de Toulouse, 135 avenue de Rangueil,
31077 Toulouse cedex, France
e-mail: marie@insa-toulouse.fr

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_6,

less temperature sensitivity). In other words, the semiconductor band structure can be changed in an artificial fashion to suit the device specifications [1].

The main ideas of band structure engineering were originally proposed by Adams and O'Reilly in United Kingdom and in parallel Kane and Yablonovitch in USA in 1986 [4, 5]. These authors predicted that the modified band structure of strained III–V quantum well structures should lead to significant benefits for diode laser performances, i.e. reduced threshold current, improved efficiency, improved temperature sensitivity and better high speed performance.

At the same time, the first strained epilayers were grown successfully. In 1982, Goldstein et al. managed to grow strained InGaAs/GaAs superlattices [6–8]. In 1984, Laidig et al. fabricated a strained quantum well semiconductor laser by adding indium into GaAs in order to reach the wavelength range 0.88–1.1 μm which was impossible to attain with classical GaAs/AlGaAs or InGaAs/InP standard lattice matched systems [9]. The first successful application of compressive strain for 1.5 μm InGaAs/InGaAsP multiple quantum well strained lasers was demonstrated in 1989 [10]. Nowadays the dramatic impact of the use of strained layers is well illustrated by the fact that they are used in almost all the optical devices based on III–V semiconductors for telecommunications (lasers and semiconductor optical amplifiers). Note that the use of strain in SiGe MOSFETs (which yields high hole mobility) will not be described in this chapter [11].

6.1.1 What is a Strained Semiconductor Layer?

To begin with, we will first explain what we mean by strained semiconductor structure and we will give the basic scheme for improving laser characteristics using this strain degree of freedom.

Let us consider the simple InGaAs/GaAs system. GaAs and InAs are two direct band gap semiconductor materials. The ternary bulk InGaAs has also a direct gap E_g. The lattice constant $a(x)$ of $In_xGa_{1-x}As$, which lies between $a_{GaAs} = 5.653$ Å and $a_{InAs} = 6.058$ Å is given by a Vegard type law: $a(x) = a_{GaAs} + 0.405.x$. Please refer to Fig. 6.23 of Sect. 6.5.1 for illustration.

When x is different from zero, GaAs and $In_xGa_{1-x}As$ have different lattice constants (note that it is not the case for GaAs and $Al_xGa_{1-x}As$ which have almost the same lattice constant). If a thin layer of InGaAs is grown on a thick layer of GaAs—by molecular beam epitaxy (MBE) for instance—the latter will impose its lattice constant in the layer plane. If we grow a GaAs layer above this structure, an elastically strained quantum well is thus obtained. The well is composed of a semiconductor which would normally have a larger lattice constant than the barrier material. In this example the InGaAs well is lattice-mismatched and the barrier is lattice matched (with respect to the GaAs substrate). We see in Fig. 6.1 that the lattice mismatch is accommodated by a tetragonal distortion of the layer: the lattice constant is different, parallel or perpendicular to the growth direction. A built-in axial strain is

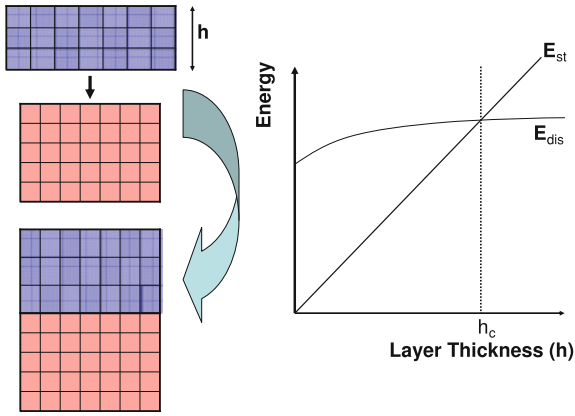


Fig. 6.1 Schematic representation of a strained semiconductor layer under biaxial compression. If the elastic energy stored in the strained layer (E_{st}) is smaller than the energy in a dislocation network relieving the strain (E_{dis}), the strained layer is thermodynamically stable ($h < h_c$) [2]

clearly present in the layer which is in biaxial compression: the layer is compressed in the x and y plane and relaxes by expanding along the z growth direction. In contrast, biaxial tension will occur when the mismatched layer has a smaller lattice constant than the one of the substrate. The order of magnitude of the lattice mismatch in actual devices is of the order of 1–2%.

One could wonder about the stability of such strained materials, we discuss below how these layers can be obtained while keeping crystal stability. From elasticity theory we know that the stored strain energy is linearly dependent on the layer thickness; we also know that there is a minimum energy associated with the formation of a dislocation and plastic relaxation [2]. Figure 6.1 displays the energy stored per unit area versus the layer thickness in a strained layer (E_{st}) and in a dislocation network relieving the strain (E_{dis}). It is clear that below a critical thickness (h_c), the elastically strained layer is thermodynamically stable and high quality growth can be achieved. Matthews et al. showed that the critical thickness h_c can be simply related to the strain ε and the Poisson ratio σ [12]:

$$\varepsilon = \frac{a\,(1-\sigma/4)\left[\ln\left(h_c\sqrt{2}/a\right)+1\right]}{2\sqrt{2}\pi h_c\,(1+\sigma)} \tag{6.1}$$

where a is the lattice parameter imposed by the substrate.

For the InGaAs/GaAs system, a good estimation of the critical thickness is given by: $\frac{\Delta a}{a} \cdot h \leq 20\,\text{nm}.\%$, where $\frac{\Delta a}{a}$ is the lattice mismatch [13].

It was initially feared that the excess energy (heat) which is dissipated in a laser structure would encourage dislocation formation. This would lead to rapid degradations of the material quality and laser characteristics. Life tests performed on various strained quantum well lasers (grown both on GaAs or InP substrates) demonstrated that it is not the case: very low degradation rates (even lower than comparable lattice-matched quantum wells) were measured [3].

The first advantage of strain is that it adds another degree of freedom to the combinations of materials that can be grown and a larger range of laser wavelengths can for instance be obtained with the same couple of materials. Without strain it would have been impossible to grow InGaAs/GaAs quantum well lasers on GaAs substrates which can reach wavelengths up to 1 μm.

Nevertheless, the main advantage of strain comes from the fact that the biaxial strain induces dramatic modifications of the electronic structure. In particular, many benefits of the strained lasers are due to the way in which the strain modifies the valence band structure. As we will show in the following a proper choice of the strain value can yield a reduction of the in-plane heavy-hole mass and hence a reduction of the corresponding density of states. The threshold current density will thus be lower than in conventional lattice-matched laser structures. The reshaping of the valence band induced by the strain will also lead to a reduction in the main loss mechanisms like Auger non-radiative recombination or intervalence band absorption (IVBA) [1].

6.1.2 Main Disadvantages of Lattice Matched III–V Semiconductor Lasers and Solutions Proposed by Band-Structure Engineering

In their pioneer work, Kane, Yablonovitch, O'Reilly and Adams pointed out clearly the non-optimised characteristics of the standard band structure of III–V semiconductors in order to fabricate very efficient diode lasers and they proposed the solutions that can be brought by the use of strain and quantum confinement [1–5]. We summarise here the main ideas.

6.1.2.1 Asymmetry of Conduction and Valence Band Masses

In the III–V semiconductors, there is a strong asymmetry between the very light conduction band and the heavy valence band masses. In the ideal situation, both masses should be as light as possible, the corresponding density of states (which is proportional to the mass in 2-D structures) would be small and the injected carrier density required to satisfy the Bernard–Duraffourg gain condition would be minimised [14]. As a consequence, a laser with a very low threshold current density could be obtained. The ideal situation of equal conduction (CB) and valence band (VB) mass is illustrated in Fig. 6.2a. One of the goals of band structure engineering for laser applications is to get close to this ideal band structure [1].

In a standard lattice-matched semiconductor laser (GaAs/AlGaAs Double Heterostructure for instance), the conduction band is filled with degenerate electrons but the holes in the valence band are non degenerate (i.e. the hole quasi-Fermi level is above the top of the valence band due to the heavy mass). The hole occupation probability at the top of the valence band is small (Fig. 6.2a).

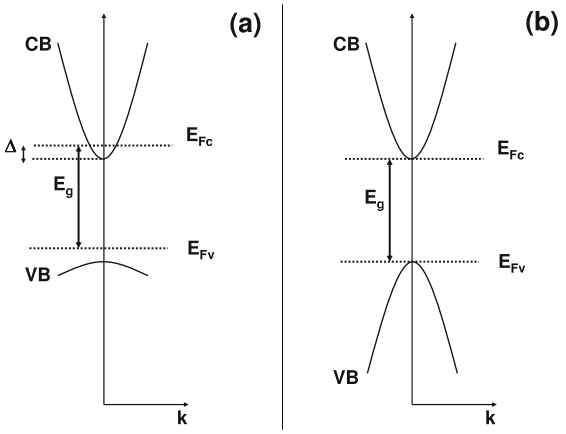


Fig. 6.2 Band structure of (a) a standard III–V semiconductor and (b) an "ideal" semiconductor with equal electron and hole masses. The Bernard–Duraffourg condition is minimally satisfied in both cases ($E_{\mathrm{Fc}} - E_{\mathrm{Fv}} = E_g$) [1]

In Fig. 6.2a, b, the Bernard–Duraffourg condition ($E_{\mathrm{Fc}} - E_{\mathrm{Fv}} \geq \hbar\omega \geq E_g$) is minimally satisfied; E_{Fc} and E_{Fv} are the conduction and valence band quasi Fermi levels.

In the ideal case (Fig. 6.2b), the carrier injection level n required to satisfy the Bernard–Duraffourg condition simply writes: $n = \int_0^\infty f(E) \cdot \rho(E)\mathrm{d}E$ with the density of states per unit area $\rho(E) = m_c/\pi\hbar^2$ (m_c is the carrier mass). One can check easily that:

$$n = \ln(2)\frac{k_B T m_c}{\pi\hbar^2} \tag{6.2}$$

In the standard situation (Fig. 6.2a), the conduction electrons are degenerate and one finds:

$$n = \frac{m_c \Delta}{\pi\hbar^2}, \quad \text{with} \quad \Delta = E_{\mathrm{Fc}}.$$

The holes are non degenerate and their density can be approximated by:

$$p = \int_0^\infty \mathrm{e}^{-(E-E_{\mathrm{Fv}})/K_B T} \cdot \frac{m_h}{\pi\hbar^2}\mathrm{d}E \quad \text{with} \quad E_{\mathrm{Fv}} = -\Delta$$

Equating n and p results in an equation for Δ which can be solved numerically. For a reasonable ratio $m_c/m_h \sim 1/6$, we get $\Delta \sim 1.43.K_B T$. Thus the carrier injection level required to reach the gain condition in the standard case writes as:

$$n = 1.43\frac{k_B T m_c}{\pi\hbar^2} \tag{6.3}$$

The ratio of carrier injection level between the two cases is $1.43/\ln(2) \sim 2$. The key result of this very simple model is that the carrier injection level for lasing is clearly

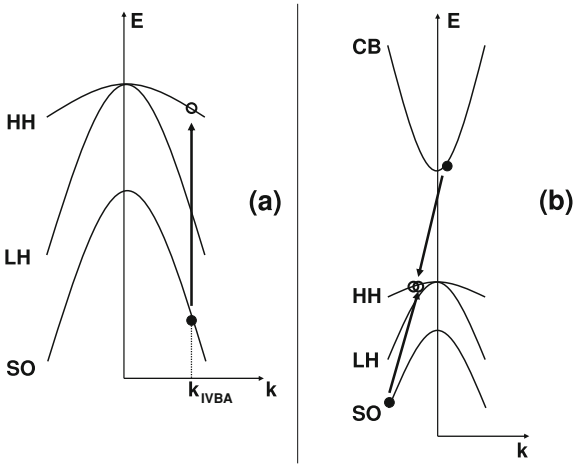


Fig. 6.3 Schematic representation of two laser loss mechanisms. (a) InterValence Band Absorption (IVBA) and (b) Auger recombination: Conduction-Heavy hole SO band-Heavy hole (CHSH) mechanism [1, 2]

reduced in the ideal band structure engineered material with equal hole and electron masses [1]. A lower threshold current density is thus predicted.

6.1.2.2 Losses Due to InterValence Band Absorption

Another problem in a III–V semiconductor laser operating in the near-infrared region is the free carrier absorption which is present above the lasing threshold. As shown in Fig. 6.3a for a bulk $In_{0.53}Ga_{0.43}As$ laser lattice matched to InP emitting at $\sim$1.5 µm ($E_g \sim 0.8$ eV), the emitted photon from the recombination of a conduction electron and a valence band hole can be absorbed in the valence band yielding a transition from the Spin-Orbit split off (SO) band to the heavy hole (HH) band [1]. This IVBA process can seriously compete with stimulated emission once lasing has started above the laser threshold. The absorption of the emitted photon energy—equal to the band gap energy (0.8 eV)—requires the transition to take place far from the Brillouin zone centre, where the energy difference between HH and SO bands is equal to the band gap energy (in the example of Fig. 6.3a, the energy difference between the top of the SO band and the top of the HH band is about 0.35 eV). The key point is that IVBA depends on the population of the off-zone centre heavy holes which can be influenced if one modifies the heavy hole mass by band structure engineering. A reduction of the heavy hole mass should yield a steeper variation of HH band, leading to a decrease of the IVBA efficiency and hence a strong reduction of this loss mechanism (see Sect. 6.2.3) [15].

6.1.2.3 Losses Due to Auger Recombination

We saw in Sect. 6.1.2.1 that a reduction of the heavy hole mass can yield a reduction of the threshold current density J_{th}. But J_{th} also strongly depends on non-radiative recombination processes, especially on Auger recombination. It is a three-body process which competes with radiative recombination; one of the most problematic mechanisms involves two HH recombining with one conduction electron leaving a hole behind in the SO band (called Conduction-Heavy hole-SO band-Heavy hole: CHSH), see Fig. 6.3b. In an oversimplified picture the Auger process is proportional to the cube of the carrier density ($\sim C \cdot n^3$). Obviously, even a factor two reduction in J_{th} (or n) can produce almost an order of magnitude reduction in Auger recombination. But the Auger coefficient itself (labelled C) can also be reduced by band structure engineering. As displayed in Fig. 6.3b, the Auger recombination is strongly dependent again on the population of the off-zone centre heavy holes since conservation of energy and momentum implies that the two arrows must be anti-parallel and of equal length. We can anticipate that a lowering of the heavy hole mass should also yield a reduction of the Auger losses due to the reduction of the population of the off zone centre heavy holes [1–3]. This will be detailed in Sect. 6.2.3.

6.2 Effects of Strain on the Band Structure

As recalled in Sect. 6.1.1, for sufficiently thin layer below the critical thickness h_c, the resulting biaxial in-plane strain causes a tetragonal deformation of the crystal lattice. This modifies drastically the electronic band structure. If we want to understand the benefits of strained layers, we have to describe in particular the way in which the strain modifies the valence band structure.

To illustrate these modifications we describe below the simple situation of strained InGaAs/GaAs layers grown on (001) substrates. We choose a coordinate axis in which the strain axis lies along the z-growth direction and the x- and y-axes in the strained layer plane.

Let us focus first on bulk strained layer (the effect of quantum confinement will be described in a second step).

6.2.1 Bulk InGaAs Under Biaxial Compression

The strain tensor has three non-vanishing components and simply writes as:

$$[\varepsilon] = \begin{pmatrix} \varepsilon_{xx} & & 0 \\ & \varepsilon_{yy} & \\ 0 & & \varepsilon_{zz} \end{pmatrix}, \quad \text{with} \quad \varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{//} = \frac{a_{GaAs} - a(x)}{a(x)} \tag{6.4}$$

What about the strain along z-direction?

The strain components are linked with the stresses by the usual Hooke law: $\sigma_i = C_{ij}\varepsilon_j$, where C_{ij} are the components of the elastic stiffness tensor. In T_d symmetry, the C_{ij} tensor writes as:

$$[C] = \begin{pmatrix} C_{11} & C_{12} & C_{12} & & & \\ C_{12} & C_{11} & C_{12} & & 0 & \\ C_{12} & C_{12} & C_{11} & & & \\ & & & 1/2(C_{11}-C_{12}) & & \\ 0 & & & & 1/2(C_{11}-C_{12}) & \\ & & & & & 1/2(C_{11}-C_{12}) \end{pmatrix} \tag{6.5}$$

Using the Hooke law, we get:

$$\sigma_z = 2C_{12}\varepsilon_{//} + C_{11}\varepsilon_z.$$

As there is no stress in the z-direction ($\sigma_z = 0$), the strain along z-axis is:

$$\varepsilon_z = -2\frac{C_{12}}{C_{11}}\varepsilon_{//} \tag{6.6}$$

It is useful for the calculations to consider that the biaxial stress is equivalent to the sum of a hydrostatic pressure and a biaxial tension along the z axis. From simple group theory considerations, the total strain can thus be resolved in a purely hydrostatic component ε_H:

$$\varepsilon_H = \varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz} = 2\left(1 - \frac{C_{12}}{C_{11}}\right)\varepsilon_{//} \tag{6.7}$$

and a tetragonal component ε_Θ:

$$\varepsilon_\Theta = 2\varepsilon_{zz} - \varepsilon_{xx} - \varepsilon_{yy} = -2\left(1 + \frac{2C_{12}}{C_{11}}\right)\varepsilon_{//} \tag{6.8}$$

Thus the Hamiltonian can be written as a linear combination of ε_H and ε_Θ:

$$H = H_H + H_\theta = U_H\varepsilon_H + U_\theta\varepsilon_\theta \tag{6.9}$$

where U_H and U_θ are electronic operators which act on the orbital part of the wavefunctions.

The hydrostatic component affects the band gap; the corresponding shift of the band edge writes as:

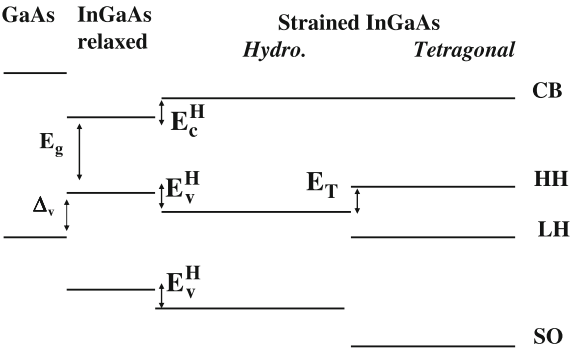


Fig. 6.4 Influence of strain (biaxial compression) on the band edge positions ($k = 0$) of bulk InGaAs

$$\begin{aligned} \text{CB: } E_c^H &= a_c \varepsilon_H \quad (a_c < 0) \\ \text{VB: } E_v^H &= a_v \varepsilon_H \quad (a_v > 0) \end{aligned} \tag{6.10}$$

a_c and a_v are the hydrostatic conduction and valence band deformation potential respectively.

For a biaxial compression ($\varepsilon_H < 0$), this means that the conduction band moves upward whereas the valence band moves downward, as sketched in Fig. 6.4. In contrast a biaxial tension will yield a reduction of the band gap.

The most important modification for band structure engineering comes from the axial component which splits the heavy and light hole states in $k = 0$ and has no effect on conduction band.

The corresponding shift of the HH band is:

$$E_T = -b_v \varepsilon_\theta \quad (b_v < 0), \tag{6.11}$$

where b_v is the tetragonal valence band deformation potential.

For a biaxial compression (tension), the HH band edge moves upward (downward). As a result the band edge positions ($k = 0$) for the conduction, heavy and light hole bands relative to the GaAs barrier valence band position write as:

$$\begin{aligned} \delta E_c &= \Delta_v + E_g + E_c^H \\ \delta E_{\text{HH}} &= \Delta_v + E_v^H + E_T \\ \delta E_{\text{LH}} &= \Delta_v + E_v^H - \frac{\Delta_{\text{SO}} + E_T - \sqrt{(\Delta_{\text{SO}} - E_T)^2 + 8E_T^2}}{2} \\ \delta E_{\text{SO}} &= \Delta_v + E_v^H - \frac{\Delta_{\text{SO}} + E_T + \sqrt{(\Delta_{\text{SO}} - E_T)^2 + 8E_T^2}}{2} \end{aligned}$$

where Δ_{SO} is the energy difference between HH and SO band edges. Considering $E_T \ll \Delta_{\text{SO}}$, one can write:

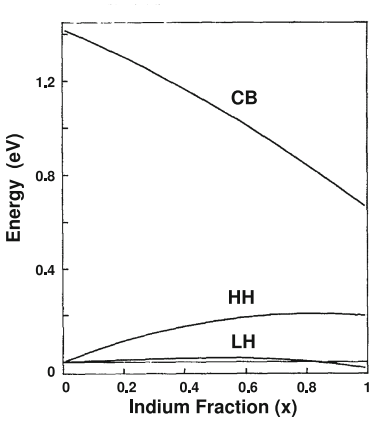


Fig. 6.5 Calculated position of the bands in strained $In_xGa_{1-x}As/GaAs$ versus the indium fraction; the reference energy is taken at the top of the binary GaAs valence band [16]

$$\begin{aligned} \delta E_{\mathrm{LH}} &\approx \Delta_v + E_v^H - E_T \\ \delta E_{\mathrm{SO}} &\approx \Delta_v + E_v^H - \Delta_{\mathrm{SO}} \end{aligned} \tag{6.12}$$

where Δ_v is the valence band offset between the two semiconductor materials.

Note that for compressive strain ($\varepsilon_\theta > 0$) the HH levels lie at the top of the valence band, as in the InGaAs/GaAs system sketched in Fig. 6.4. In contrast, for tensile strain ($\varepsilon_\theta < 0$), the LH levels are the top of the valence band.

Knowing the parameters $a_c, a_v, b_v, C_{ij}, E_g, \Delta_v$, the band edge positions in the $In_xGa_{1-x}As$ strained bulk material can be easily calculated as a function of the Indium content x, see Fig. 6.5. When x increases, the overall effect is a decrease in the band gap energy and an increase in the HH–LH splitting energy [16].

6.2.2 Electronic Band Structure in Strained Quantum Wells

Once the band edge positions in the strained bulk material have been obtained, one can simply use the envelop function approximation to describe the electronic structure in the quantum well. The first obvious effect of strain is that there is a different well depth for the heavy holes compared to the one for the light holes. Both quantum confinement and strain change the energy of the electron and hole levels. As the strain and thickness of the quantum well can be changed independently (provided that it is smaller than the critical thickness h_c), this will be very useful to optimise the characteristics of the optical devices.

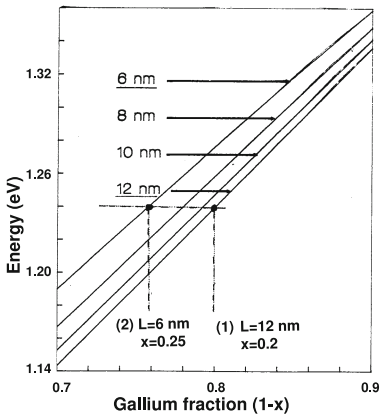


Fig. 6.6 Calculated energy transition E1–HH1 for different $In_xGa_{1-x}As/GaAs$ quantum well compositions and well widths L [17]

6.2.2.1 Optimisation of the Confined Level Positions

Band structure engineering means playing with the material parameters (composition, strain, quantum confinement) in order to optimise the device performances. Let us consider a first optimisation which simply relies on the position of the different carrier quantised levels in $k = 0$.

For a low laser threshold current, it is desirable to have a quantum well with only one populated level at the working temperature of the device. This requires a separation with the second quantised level to be greater than k_BT.

This condition is usually easily satisfied for conduction electron because of their low mass (in the infinite barrier height approximation the confinement energy writes as $E_n \approx \frac{\hbar^2}{2m}\left(\frac{\pi n}{L}\right)^2$, where m is the carrier mass and L the quantum well thickness). For holes, the condition is much more difficult to obtain.

Let us consider for illustration that one wants to fabricate an InGaAs/GaAs laser emitting at 1 μm (1.24 eV).

With the curve network displayed in Fig. 6.6, we see that our goal can be reached with an indium fraction of $x = 0.2$ and a well thickness $L = 12$ nm for example. But Fig. 6.7a shows that the second Heavy-hole level (HH2) lies only 15 meV above HH1 and will thus be populated at the device operating temperature. In Fig. 6.6 we note that the same emission wavelength can be obtained with another couple of parameters: $x = 0.25$ and $L = 6$ nm. This second choice is more favourable to get a low threshold current since the splitting in $k = 0$ between HH1 and HH2 is about 40 meV, i.e. larger than k_BT (see Fig. 6.7b).

The same kind of optimisation can be performed on strained $In_xGa_{1-x}As/InGaAsP$ quantum wells grown on InP which is the most common system used for 1.3 and 1.5 μm optical telecommunication devices (lasers and semiconductor optical amplifiers). As displayed in Fig. 6.8, 1.5 μm emission can be reached for only one couple of compositions (Indium fraction $x = 0.53$) and well width ($L = 6$ nm) if one

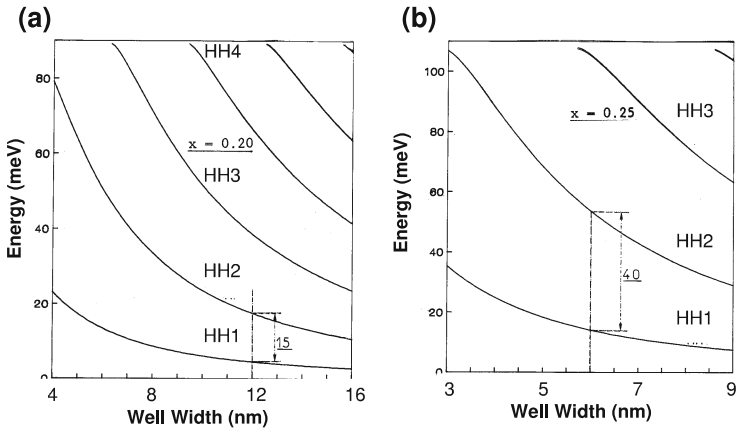


Fig. 6.7 Calculated confined heavy hole energies as a function of the well width for (a) $In_{0.2}Ga_{0.8}As$ and (b) $In_{0.25}Ga_{0.75}As$ quantum wells [17]

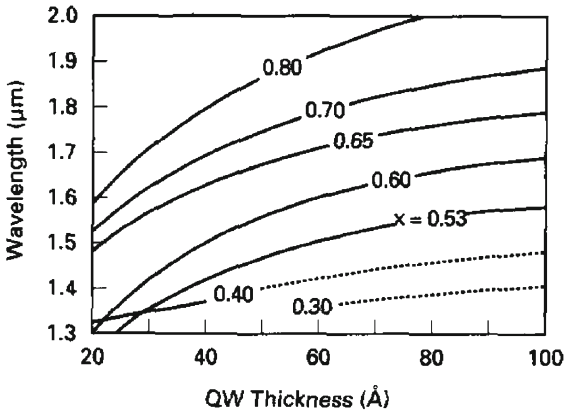


Fig. 6.8 Calculated emission wavelength of $In_xGa_{1-x}As/InGaAsP$ quantum wells as a function of the quantum well thickness [3]

considers only a lattice matched quantum well. The use of strain (both in compression $x > 0.53$ or tension $x < 0.53$) allows the optimisation of the structure thanks to the choice of many couples (x, L) which yield the same emission wavelength [3].

6.2.2.2 In-Plane Dispersion Curves and Density of States

The modification of the in-plane effective mass of the valence band is probably the most important advantage of electronic band structure engineering for the optimisation of device performances. We saw in the previous section that the axial strain breaks the cubic symmetry of the semiconductor which leads to a splitting in $k = 0$ between heavy and light hole bands (even in the absence of quantum confinement); the typical splitting is 60–80 meV for 1% lattice mismatch (for instance for $In_{0.15}Ga_{0.75}As$). The tetragonal distortion of the lattice leads to a highly anisotropic band structure

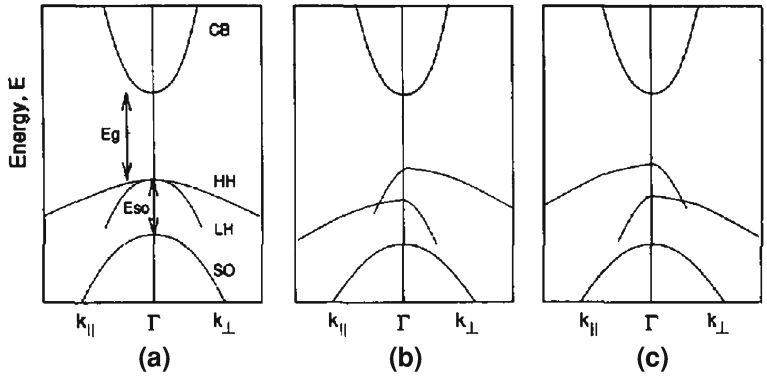


Fig. 6.9 Schematic representation of the band structure of (a) an unstrained bulk direct-gap cubic semiconductor; the same semiconductor under (b) biaxial compression or (c) tension [2]

for k different from zero. In other words, the mass along the growth direction will be different from the mass perpendicular to it. The consequences for the quantum well properties will be very important since the mass along the growth direction determines the quantum confinement energy while the density of states (DOS) is proportional to the in-plane quantum well mass.

The simplest method to calculate the carrier spectra near the conduction band minimum and valence band maxima is the $\mathbf{k} \cdot \mathbf{p}$ method which is widely used in band structure engineering applications [18–20] (see Chap. 1).

For a review of the very powerful techniques based on tight binding, pseudopotential or orthogonalised plane-wave methods, the reader can refer to Chap. 2.

With the $k \cdot p$ technique, the dispersion curves (carrier energy versus the in-plane wavevector) in strained layers can be calculated by solving the Bir and Pikus Hamiltonian [20]: $H^{\mathrm{BP}} = H^{\mathrm{KL}} + H^{\mathrm{strain}}$. The Kohn Luttinger Hamiltonian (H^{KL}) can be found in many good textbooks [18–21].

Considering the simple 8-band approach (CB, HH, LH and SO), the strain component simply writes as:

$$H^{\mathrm{Strain}} = \begin{pmatrix} E_c^H & 0 & 0 & 0 \\ 0 & E_v^H - E_T & 0 & \sqrt{2}E_T \\ 0 & 0 & E_v^H + E_T & 0 \\ 0 & \sqrt{2}E_T & 0 & E_v^H \end{pmatrix} \begin{matrix} u_1, u_2 \\ u_3, u_4 \\ u_5, u_6 \\ u_7, u_8 \end{matrix} \tag{6.13}$$

where u_i are the periodic parts of the Bloch functions.

A schematic representation of the calculated band structure of unstrained and strained (biaxial compression or tension) bulk semiconductor is presented in Fig. 6.9.

Under biaxial compression, in addition to the already described change in energy levels in $k = 0$, we note that the dispersion curves of the valence band are very anisotropic. At the top of the valence band lies the heavy holes which are characterised

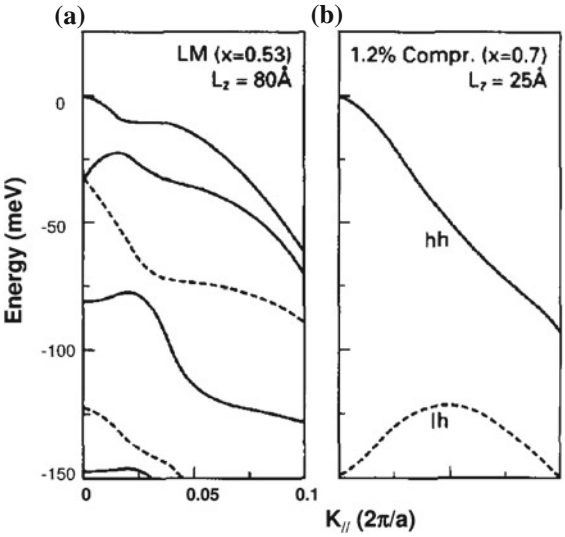


Fig. 6.10 Calculated band structure of $In_xGa_{1-x}As/InGaAsP/InP$ quantum well (a) lattice-matched $x = 0.53$ (b) 1.2% compressively strained $x = 0.7$. The well width L, is chosen for emission at 1.5 μm wavelength [3]

by a "heavy" mass along the growth axis $k_\perp$ and a "light" mass in the layer plane $k_{//}$. The opposite behaviour holds for the light holes. This raises a serious problem of terminology. The rule is the following: we label the bands by their mass along the growth direction; we could also label them by their angular momentum projection $J_z = \pm 3/2$ for HH and $J_z = \pm 1/2$ for LH.

As displayed in Fig. 6.9b, the main effect of biaxial compression is that the in-plane heavy hole band has a lighter mass compared to the lattice matched material [22, 23].

This is a key advantage to fabricate lasers with low thresholds since the DOS will be smaller and our goal of quasi equal electron and hole mass presented in Sect. 6.1.2.1 can be reached.

So far we commented on the dispersion curves in bulk strained material. The eigenvalue problem in the quantum well structure can then be solved by the transfer-matrix method, taking into account the interfacial discontinuity condition [24, 25].

Figure 6.10a presents the calculated in-plane dispersion curve of the valence band for two quantum well structures emitting at 1.5 μm [3]. The first one is a lattice-matched $In_{0.53}Ga_{0.47}As/InGaAsP/InP$ quantum well. The second is a 1.2% compressively strained $In_{0.7}Ga_{0.3}As/InGaAsP/InP$ with a different well width to get the same emission wavelength.

The first difference between the two structures is the dramatic increase of the energy separation between the HH1 and LH1 subbands. We also clearly observe the drastic modification of the in-plane heavy hole mass from ~0.7 m_0 for the unstrained quantum well to ~0.15 m_0 for the strained one.

Another example of valence band in-plane dispersion curve is displayed in Fig. 6.11 for a 5 nm strained $In_{0.77}Ga_{0.23}As/InP$ quantum well [1]. The dashed line

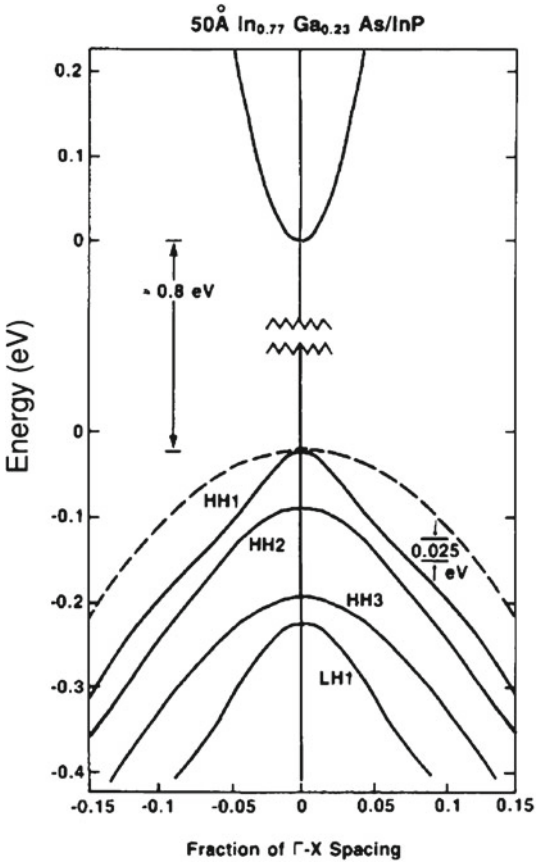


Fig. 6.11 Calculated band structure (energy as a function of the in-plane wavevector) for a 5 nm strained $In_{0.77}Ga_{0.23}As/InP$ quantum well [1]

is the heavy hole dispersion of lattice matched $In_{0.53}Ga_{0.47}As$ which has a similar band gap to the strained layer. Three advantages for laser operation can be identified:

- First advantage: light heavy hole mass, the mass of the HH1 subband is $\sim 0.09\, m_0$.
- Second advantage: the splitting between HH1 and HH2 is about 200 meV, this minimises the undesirable thermal occupation of HH2
- Third advantage: it deals with the loss mechanisms. In Fig. 6.11 we note a significant depression of the HH1 energy away from the zone centre. The energy difference between the HH1 subband and the dashed line away from the zone centre is about 75 meV. This leads to $\sim e^{-0.075/0.025} = e^{-3}$ reduction at room temperature of the off-zone centre heavy hole population due to band structure engineering. As explained in Sect. 6.1.2.2, InterValence Band Absorption depends strongly on this population and thus should be reduced by a similar factor. Similarly, the Auger effect (CHSH mechanism in particular) should be significantly reduced.

6.2.3 *Influence of Strain on the Loss Mechanisms*

6.2.3.1 InterValence Band Absorption

The magnitude of IVBA depends directly on the density of holes at k_{IVBA} (Fig. 6.3). If we assume for simplicity a parabolic band structure, the kinetic energy of the heavy holes (E_{HH}) and SO holes (E_S) taking part in IVBA write as:

$$E_{\mathrm{HH}} = \frac{\hbar^2 k_{\mathrm{IVBA}}^2}{2m_h} \quad \text{and} \quad E_S = \Delta_{\mathrm{SO}} + \frac{\hbar^2 k_{\mathrm{IVBA}}^2}{2m_{\mathrm{SO}}} \tag{6.14}$$

IVBA occurs if the emitted photon with energy E_g is absorbed between the SO and HH bands: $E_S - E_{\mathrm{HH}} = E_g$. Thus, the kinetic energy of the holes E_{HH} is:

$$E_{\mathrm{HH}} = \left(\frac{m_{\mathrm{SO}}}{m_{\mathrm{HH}} - m_{\mathrm{SO}}}\right).(E_g - \Delta_{\mathrm{SO}}) \tag{6.15}$$

When the HH mass is reduced towards that of the SO band, the corresponding hole energy for IVBA increases. Thus IVBA occurs at a much larger wavevector k_{IVBA} at which the hole occupation probability at room temperature is almost zero.

The reduction of IVBA efficiency has been checked experimentally through hydrostatic pressure measurements. This experimental technique is a very useful tool to investigate the loss mechanisms in semiconductor lasers. The effect of external hydrostatic pressure P_H is an increase in the bandgap energy (typically 10 meV/kbar) without affecting the subband dispersion. As a consequence the point in k space at which IVBA occurs moves to larger k values where the hole carrier density and hence the IVBA process is less probable. The measurement of the laser differential efficiency η changes with P_H provides a probe of the IVBA strength [2].

We see in Fig. 6.12 that the hydrostatic pressure causes a considerable increase in η (i.e. a reduction of IVBA) in bulk InGaAsP and in strained multiple quantum well devices [3]. This clearly proves the importance of IVBA for these 1.5 μm lasers. In contrast, when the same pressure is applied to compressively or tensile strained quantum well lasers, no such increase is observed. The interpretation is simply that IVBA was negligible in these quantum wells and so any further increase in bandgap due to external pressure had no additional effect.

6.2.3.2 Auger Recombination

Auger recombination, which is a very important loss mechanism for telecommunication lasers, involves three carriers; so the Auger current J_{Auger} varies approximately as: $J_{\mathrm{Auger}}(T) \cong C(T) n_{\mathrm{th}}^3$, where C(T) is the temperature-dependent Auger coefficient and n_{th} the threshold carrier density.

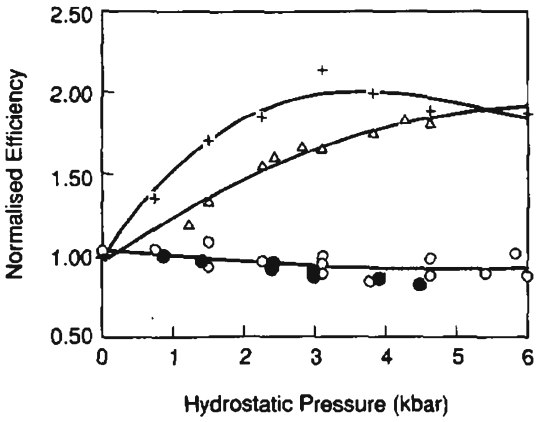


Fig. 6.12 Normalised efficiency as a function of an external hydrostatic pressure for (+) bulk InGaAsP, (Δ) unstrained $In_{0.53}Ga_{0.47}As$/InGaAsP, (○) 1.8% compressively strained $In_{0.8}Ga_{0.2}As$/InGaAsP, and (●)1.6% tensile-strained $In_{0.32}Ga_{0.68}As$/InGaAsP multiple quantum well lasers operating at 1.5 μm wavelength [3]

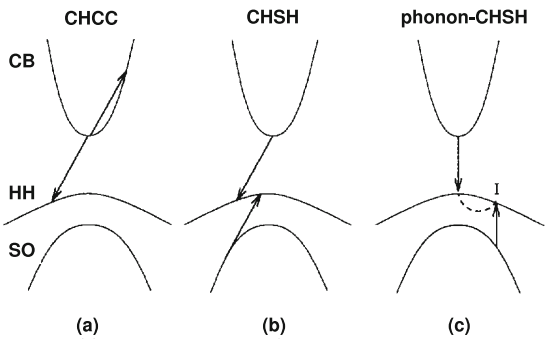


Fig. 6.13 Schematic representation of the three types of Auger recombination processes. (a) CHCC process (Conduction-Heavy hole Conduction–Conduction). (b) CHSH process (Conduction-Heavy hole Spin-orbit-Heavy hole); (c) Phonon-assisted CHSH [2]

The influence of strain and band structure engineering on Auger non-radiative recombination can be divided into two aspects: first, J_{Auger} is very sensitive to any reduction in n_{th} brought about by strain through the decrease of the in-plane valence band mass; second, strain may change the magnitude of the Auger coefficient C(T) itself.

There are three kinds of Auger recombinations as schematically sketched in Fig. 6.13 [2, 26, 27]:

- CHCC process (Conduction-Heavy hole Conduction–Conduction):
  The energy and wavevector released when an electron and a hole recombine is used to excite another conduction electron to higher conduction states.
- CHSH process (Conduction-Heavy hole Spin-orbit–Heavy hole):
  Two heavy holes recombine with one free electron leaving a hole behind in the spin-orbit split-off band. The efficiency of this process is proportional to the square of the off-zone centre HH population.
- Phonon assisted CHSH:
  It is a second order process which involves in addition a phonon: in this non-radiative recombination a conduction electron and a heavy hole recombine while

an electron passes from the SO band to the HH band with the simultaneous emission or absorption of a phonon for the wavevector conservation.

It is clear from Fig. 6.13 that the three processes depend strongly on the population of the off-zone centre HH population which can be modified by the reshaping of the valence band induced by strain and quantum confinement [1, 2]. Auger losses should be reduced in band structure engineered strained quantum well lasers. In a very simple picture, the Auger current corresponding to CHSH process can be written as:

$$J_{\text{Auger}} \approx \mathrm{e}^{-(E_a/K_BT)} \cdot n_{\text{th}}^3, \quad \text{with the activation energy}$$
$$E_a = \frac{m_{\text{SO}}}{2m_{\text{HH}} + m_e - m_{\text{SO}}}(E_G - \Delta_{\text{SO}}) \tag{6.16}$$

If the in-plane heavy hole mass is reduced (as in compressively strained QW for instance), the activation energy E_a increases and we expect to get a reduced Auger current J_{Auger}.

A more realistic calculation requires to take into account all the possible transitions. If one considers only the transitions which can occur between quantum well states (bound-bound transitions) [25], one has to calculate:

$$J_{\text{Auger}} \propto \sum_{\text{all states}} P(e, h_1, h_2, \text{so}).|M|^2.\delta(E), \quad \text{with}$$
$$P(e, h_1, h_2, \text{so}), \quad \approx f_c(k_e).f_v(k_{h1}).f_v(k_{h2}) \tag{6.17}$$

The matrix element for an unscreened Coulomb interaction writes as:

$$M \approx \iint \Psi^*_{k_{h1}}(r_1)\Psi^*_{k_{h2}}(r_2)\frac{\mathrm{e}^2}{4\pi\varepsilon_0\varepsilon_r\,|r_1 - r_2|}\Psi_{k\,\text{so}}(r_2)\Psi_{k\,e}(r_1)\mathrm{d}r_1\mathrm{d}r_2$$

Using this approach, the calculated Auger current can be compared in a lattice-matched ($\varepsilon = 0$) and a compressively strained Quantum Well ($\varepsilon = 1\%$) [26, 27]. Figure 6.14 shows that the Auger current is reduced as expected in the strained quantum well as a consequence of the reduced heavy-hole mass.

The experimental proof of lower Auger recombination in strained quantum well lasers can be obtained again by performing measurements under external hydrostatic pressures [2, 3]. A very large proportion of the current at threshold in classical 1.5 μm lasers is due to Auger non-radiative recombination. Thus the variation of J_{th} as a function of the hydrostatic pressure P_H should give information about the amplitude on the non-radiative Auger current.

Figure 6.15 presents the results of these measurements in bulk or 1.5 μm strained quantum well lasers. A decrease in the threshold current is observed in all lasers with increasing pressure. This is due to a decrease of the Auger process induced by a reduction of the off-zone centre HH population [3]. Nevertheless, the reduction is clearly greater in bulk compared to the one observed in strained quantum wells.

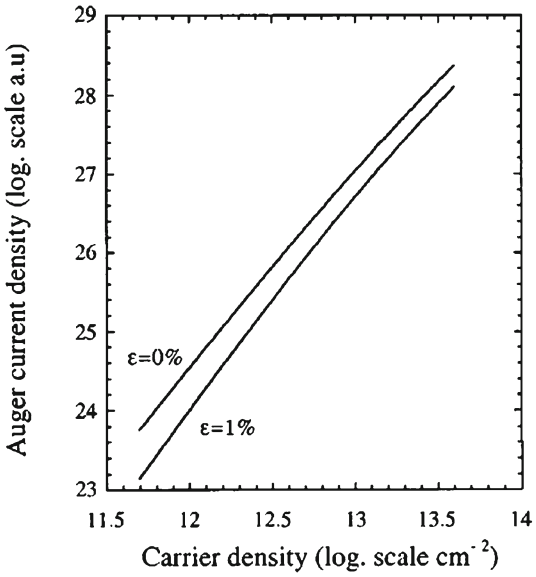


Fig. 6.14 Calculated Auger recombination (CHSH) current density versus carrier density for an unstrained 8 nm $In_{0.53}Ga_{0.47}As$/InGaAsP/InP quantum well and an 8 nm strained (ε = 1%) $In_{0.79}Ga_{0.21}As_{0.77}P_{0.23}$/InGaAsP/InP quantum well [26, 27]

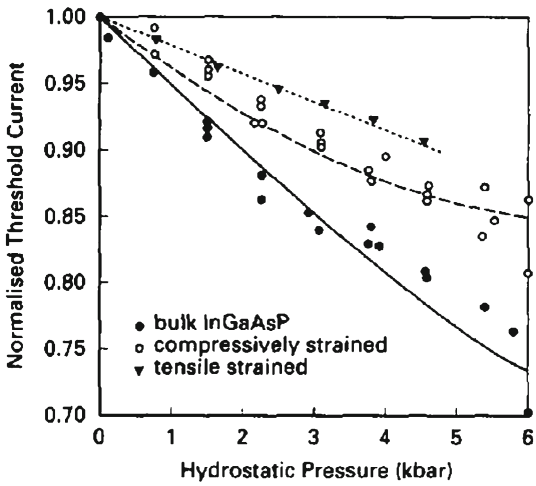


Fig. 6.15 Normalised threshold currents as a function of the hydrostatic pressure for (•)1.5 μm wavelength bulk InGaAsP, (○)1.8% compressively strained $In_{0.8}Ga_{0.2}As$ QW, and (▼)1.6% tensile-strained $In_{32}Ga_{68}As$ QW lasers [3]

This shows that the non-radiative current was already reduced in the strained quantum well laser and hence less reduction is observed by applying an external hydrostatic pressure.

In contrast to IVBA, the measured reduction of Auger process in strained quantum well laser is smaller than the one which could be predicted: in an oversimplified picture a reduction by a factor two of the threshold carrier density should yield almost an order of magnitude ($\sim 2^3$) decrease of Auger. In addition to the over-simplication of the model, the discrepancy can be attributed to the persistence of phonon-assisted Auger which is very difficult to suppress [28, 29].

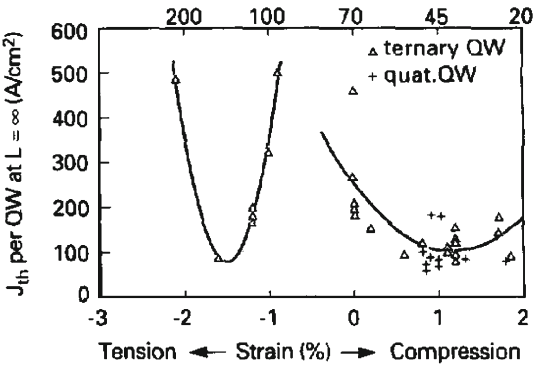


Fig. 6.16 Threshold current densities per quantum well deduced for infinite cavity length 1.5 μm lasers versus strain in InGaAs(P) [3]

6.2.3.3 Influence of Strain on the Temperature Sensitivity

The temperature sensitivity of a semiconductor laser is usually described by the T_0 parameter, where T_0, also called characteristic temperature, is related to the temperature dependence of the threshold current I_{th} by:

$$\frac{1}{T_0} = \frac{d}{dT}(\ln(I_{th})) \tag{6.18}$$

Auger recombination, which is temperature dependent, is usually the dominant cause of the poor temperature characteristics of long-wavelength diode lasers. As a consequence the temperature sensitivity of strained 1.5 μm lasers is improved compared to that of bulk or lattice-matched quantum wells [2, 3]. The temperature sensitivity of the laser will also depend on the possible escape of carriers out of the quantum well due to poor electron or hole confinement; this point will be discussed in Sect. 6.4.

6.2.4 Strain-Induced Changes of the Laser Threshold Current

Figure 6.16 summarises well the advantages of band structure engineering of semiconductor lasers for optical telecommunications. It displays the measured threshold current as a function of strain [3].

In the compressive strain branch, a clear reduction of the threshold current is observed with increasing compressive strain since the heavy hole mass monotonically decreases yielding a decrease of the DOS. The occupation of the off-zone centre HH population is also reduced leading to less loss mechanisms (IVBA and Auger). Both effects contribute to the measured reduction in threshold current [2, 3].

In the tensile strain branch, for moderate strain, the highest light hole band is affected by strong valence band mixing, yielding a rather "heavy" mass and large DOS; this explains the large measured threshold current. However, with increasing

tensile strain, the HH–LH valence subband separation increases, which reduces the band mixing effects. This results in a decrease of the light hole mass and again of the loss mechanism. A dramatic drop of the threshold current density, down to $\sim 90\,\mathrm{A/cm^2}$, is observed. For larger strain (up to $\sim 2\%$), an increase of J_{th} is again measured, which can be attributed both to the poor conduction band offset [3, 22] and to a quantum well width close to the critical thickness.

The reduction of the laser threshold current is a good demonstration of the way band structure engineering can improve the optoelectronic device performances. But many other device features can be optimized through a proper choice of strain and quantum confinement in the active layers: the reduction of the hole mass can also yield a useful increase in the differential gain and hence an increase in the relaxation oscillation frequency and a possible decrease of the linewidth enhancement factor (see Chap. 7).

To finish this section, let us mention that strain compensated quantum wells are often used in real devices. This strain compensation is another tool to improve the performances and reliability. In order to reduce or even eliminate the net strain in the structure, the idea is to grow opposite strains in the wells and in the barriers. For instance, the quantum well with a well width L_W can be grown under biaxial compression (ε_W) and the barrier with a width L_B under tension (ε_B) [30]. Thus the average strain writes as:

$$\varepsilon_{\mathrm{av}} = \frac{\varepsilon_W L_W + \varepsilon_B L_B}{L_W + L_B} \tag{6.19}$$

A proper choice of the parameter can yield $\varepsilon_{\mathrm{av}} \sim 0$. It has been shown that this can improve significantly the reliability of the laser when multiple quantum wells are used [14]. The possibility of growing strained barriers adds again an additional degree of freedom for the combinations of materials which can be used for the growth of the well and the barrier.

6.3 Gain Calculation in III–V Quantum Wells

6.3.1 Device Geometry

The following sections are dedicated to the modelling of laser and semiconductor optical amplifier active layers and to the optimisation of their performances using band structure engineering. We have chosen to investigate quantum well heterostructure devices in which light propagates in the layer plane, i.e. perpendicularly to the quantisation axis.

Assuming that light propagates along x-axis, electric wave can either oscillate along y-axis in transverse electric mode (TE) or along z-axis in transverse magnetic mode (TM) as shown in Fig. 6.17.

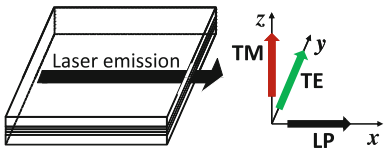


Fig. 6.17 Light propagation (LP along x-axis) and polarisation in edge emitting devices (TE polarization along y-axis and TM polarization along z-axis), after [2]

6.3.2 Carrier Wavefunctions in Quantum Wells

Using the envelope function model, the wavefunctions of carriers in the quantum well can be written as follows:

$$\Psi_m^C = F_m^C u_C \mathrm{e}^{ik_{\parallel}\cdot r_{\parallel}} \text{ for electrons}(C) \tag{6.20}$$

$$\Psi_n^{v,h} = g_n^v u_{v,h}\, \mathrm{e}^{ik_{\parallel}\cdot r_{\parallel}} \tag{6.21}$$

and

$$\Psi_n^{v,l} = f_n^v u_{v,l}\, \mathrm{e}^{ik_{\parallel}\cdot r_{\parallel}} \tag{6.22}$$

for heavy (V, h) and light (V, l) hole states, respectively.

F_m^C, g_n^v and f_n^v are the envelope functions of electron, heavy and light hole states in the quantum well, their variation and symmetry rely on the heterostructure geometry and on the subband indices (m, n);

u_C and u_v are the periodic Bloch functions of conduction and valence band; $k_{\parallel}$ and $r_{\parallel}$ are the wave vector and position vector in the layer plane.

6.3.3 Light–Matter Interaction and Optical Selection Rules

In the presence of an electromagnetic field, the one-electron Hamiltonian of a heterostructure can be written [18] as:

$$H = H_0 + \frac{e}{2m_0 c}\left[p \cdot A + A \cdot p\right] \tag{6.23}$$

where e and m_0 are the electron charge and mass, A is the vector potential and p is the electron momentum.

The dipole matrix element coupling electron and hole states writes as:

$$M_{\mathrm{nm}}\left(k_{\parallel}\right) = \langle \Psi_m^C | \varepsilon \cdot p | \Psi_n^V \rangle \tag{6.24}$$

Here ε is the light polarization.

For electron-heavy hole recombination, this latter expression can be rewritten as:

Table 6.1 Selection rules for interband transitions obtained from the absolute value of the matrix element $\langle u_C|\varepsilon \cdot p|u_V\rangle$ in $k_{\shortparallel} = 0$, after Bastard [18]

Polarization	ε_x	ε_y	ε_z	Transition type
Along x-axis	Impossible	$\frac{\Pi}{\sqrt{2}}$	Forbidden	$E \rightarrow HH$
Along y-axis	$\frac{\Pi}{\sqrt{2}}$	Impossible	Forbidden	
Along z-axis	$\frac{\Pi}{\sqrt{2}}$	$\frac{\Pi}{\sqrt{2}}$	Impossible	
Along x-axis	Impossible	$\frac{\Pi}{\sqrt{6}}$	$\frac{2\Pi}{\sqrt{6}}$	$E \rightarrow LH$
Along y-axis	$\frac{\Pi}{\sqrt{6}}$	Impossible	$\frac{2\Pi}{\sqrt{6}}$	
Along z-axis	$\frac{\Pi}{\sqrt{6}}$	$\frac{\Pi}{\sqrt{6}}$	Impossible	

$$M_{\mathrm{nm}}\left(k_{\shortparallel}\right) = \langle F_m^C|g_n^V\rangle\langle u_C|\varepsilon \cdot p|u_{V,h}\rangle \tag{6.25}$$

and for electron-light hole recombination:

$$M_{\mathrm{nm}}\left(k_{\shortparallel}\right) = \left\langle F_m^C|f_n^V\right\rangle\left\langle u_C|\varepsilon \cdot p|u_{V,l}\right\rangle \tag{6.26}$$

$\left|\left\langle F_m^C|g_n^V\right\rangle\right|$ and $\left|\left\langle F_m^C|f_n^V\right\rangle\right|$ are the envelope function overlaps of electronic $\left(F_m^C\right)$ and heavy $\left(g_n^V\right)$ and light $\left(f_n^V\right)$ hole states, respectively.

$\langle u_C|\varepsilon \cdot p|u_V\rangle$ depends on the wave polarisation and on the periodic Bloch functions of the conduction and valence band edges, its values are reported in Table 6.1 [18]:

Assuming propagation along x-axis (see Fig. 6.17), only TE mode is allowed (polarisation along y-axis) for electron-heavy hole recombinations ($E \rightarrow HH$). For electron-light hole recombination ($E \rightarrow LH$), both polarizations are possible (electric field along y-axis (TE) or along z-axis (TM)).

Hence, for TE mode, transition matrix element writes as:

$$\left|M_{\mathrm{nm}}^{\mathrm{TE}}(k)\right|^2 = \prod^2\left(\left|\left\langle F_m^C|g_n^V\right\rangle\right|^2 + \frac{1}{3}\left|\left\langle F_m^C|f_n^V\right\rangle\right|^2\right) \tag{6.27}$$

TM emission is only due to electron-light hole recombinations, but the corresponding TM transition is twice more intense than for TE polarization emission:

$$\left|M_{\mathrm{nm}}^{\mathrm{TM}}(k)\right|^2 = \frac{2\Pi^2}{3}\left|\left\langle F_m^C|f_n^V\right\rangle\right|^2 \tag{6.28}$$

$\Pi = \frac{-i}{m_0}\langle S|p_x|X\rangle = \frac{-i}{m_0}\left\langle S|p_y|Y\right\rangle = \frac{-i}{m_0}\langle S|p_z|Z\rangle$ is related to the Kane matrix element $E_p = 2m_0\Pi^2$ ($E_p \sim 20\,\mathrm{meV}$ whatever the III–V material is). $|S\rangle$, $|X\rangle$, $|Y\rangle$ and $|Z\rangle$ are the band edge Bloch functions of s-like conduction band and p-like valence band at Γ point.

6.3.4 Gain Calculation

Most characteristics of optoelectonic devices such as laser threshold current, laser linewidth or semiconductor optical amplifier optical bandwidth are related to the material gain of device active layers. One method to determine the material gain consists in calculating the first-order dielectric susceptibility $\tilde{\chi}(\omega)$ [31].

For a bulk material, $\tilde{\chi}(\omega)$ writes as:

$$\begin{aligned}\tilde{\chi}(\omega) &= \chi'(\omega) + i\chi''(\omega) \\ &= \frac{1}{V\varepsilon_0}\left(\frac{e}{m_0\omega}\right)^2 \sum_{C,V} |M_{CV}|^2 \frac{\left(f\left(E^C - E_{\mathrm{Fc}}\right) - f(E_{\mathrm{Fv}} - E^V)\right)}{\left(\hbar\omega - E^C - E^V - i\hbar\gamma_{\mathrm{int}}\right)}\end{aligned} \tag{6.29}$$

where $f(E^c - E_{\mathrm{fc}})$ and $f(E_{\mathrm{Fv}} - E^V)$ are the Fermi–Dirac conduction and valence band occupancy numbers. E^C and E^V are the carrier energies in conduction and valence band, respectively, and E_{Fc} and E_{Fv} are the quasi-Fermi levels of conduction and valence band, respectively. $\tilde{\chi}(\omega)$ is obtained by summing over the whole energy values in all subbands (C, V).

$\gamma_{\mathrm{int}} = \frac{1}{\tau_{\mathrm{int}}}$ is the reciprocal of intraband relaxation time.

The material gain is then calculated as follows:

$$G = \frac{\omega\mu_0 c\varepsilon_0}{n}\mathrm{Im}(\tilde{\chi}) \tag{6.30}$$

$$\begin{aligned}G = \frac{e^2\hbar}{m_0^2 nc\varepsilon_0}\frac{1}{\hbar\omega}\frac{1}{V}\sum_{C,V}|M_{CV}|^2\left(f\left(E^C - E_{\mathrm{Fc}}\right) - f(E_{\mathrm{Fv}} - E^V)\right) \\ \times\frac{\hbar\gamma_{\mathrm{int}}}{\left(\hbar\omega - E^C - E^V\right)^2 + (\hbar\gamma_{\mathrm{int}})^2}\end{aligned} \tag{6.31}$$

In the case of a quantum well with m electron and n hole subbands, the latter expression transforms [32]:

$$\begin{aligned}G = \frac{e^2\hbar}{m_0^2 nc\varepsilon_0}\frac{1}{\hbar\omega}\frac{1}{L}\sum_n\sum_m\int_0^{+\infty}\frac{k_{\mathrm{II}}}{\pi}\left|M_{\mathrm{nm}}(k_{\mathrm{II}})\right|^2\left(f\left(E_m^C(k_{\mathrm{II}}) - E_{\mathrm{Fc}}\right)\right. \\ \left. - f(E_{\mathrm{Fv}} - E_n^V(k_{\mathrm{II}}))\right)\frac{\hbar\gamma_{\mathrm{int}}}{\left(\hbar\omega - E_m^C(k_{\mathrm{II}}) - E_n^V(k_{\mathrm{II}})\right)^2 + (\hbar\gamma_{\mathrm{int}})^2}\mathrm{d}k_{\mathrm{II}}\end{aligned} \tag{6.32}$$

where L is the quantum well width.

However, due to the Lorentz distribution in the gain formula, this expression leads to two non physical effects: artificial absorption below band gap and underestimation

of Bernard–Duraffourg condition. It can be corrected as follows [26, 27, 33–35]:

$$G = \frac{e^2\hbar}{m_0^2 nc\varepsilon_0}\frac{1}{\hbar\omega}\frac{1}{L}\left(1 - \mathrm{e}^{\frac{1}{kT}(\hbar\omega - \Delta E_F)}\right) \times \sum_n \sum_m \int_0^{+\infty} \frac{k_{||}}{\pi} \left|M_{\mathrm{nm}}(k_{||})\right|^2 f\left(E_m^C(k_{||}) - E_{\mathrm{Fc}}\right)\left(1 - f(E_{\mathrm{Fv}} - E_n^V(k_{||})\right) \times \frac{\hbar\gamma_{\mathrm{int}}}{\left(\hbar\omega - E_m^C(k_{||}) - E_n^V(k_{||})\right)^2 + (\hbar\gamma_{\mathrm{int}})^2}\mathrm{d}k_{||} \tag{6.33}$$

$$\text{with } \Delta E_F = E_{\mathrm{Fc}} - E_{\mathrm{Fv}} \tag{6.34}$$

In the following sections, we describe the design and optimisation of 1.3 μm lasers (Sect. 6.4) and 1.55 μm semiconductor optical amplifiers (Sect. 6.5) using band structure engineering and gain calculation.

6.4 Uncooled Operation of 1.3 μm Lasers

Usual semiconductor lasers operating at 1.3 μm are fabricated using InP technology. The active layers of these devices are InGaAsP/InGaAsP [36] or InGaAlAs/InGaAlAs [37, 38] heterostructures grown on InP substrate with different quaternary compositions in well and barrier materials. Both technologies have their own advantages. One the one hand, InGaAlAs system offers large conduction and valence band offsets, resulting in an efficient carrier confinement in quantum wells and good performance of laser devices, even at high temperature. Its major drawback is the difficulty of integration technique with InGaAsP waveguides [39]. On the other hand, InGaAsP heterostructures are more easily fabricated, but the poor conduction band offset in this system leads to electron spillover out of the quantum wells, and hence to lower performances at operating temperatures. One last common disadvantage of these devices is the high cost of InP technology as compared to GaAs technology.

In 1992, Weyers et al. observed a strong reduction of band gap energy in dilute nitride GaAsN grown on GaAs substrate [40]. At that time, InP-based lasers emitting at 1.3 μm required thermoelectric coolers to overcome temperature losses due to poor electron confinement, and InGaAs/GaAs based lasers grown on [100]-oriented substrates were limited to 1 μm emission wavelength due to the large compressive strain induced by high indium contents necessary to reach these wavelengths. In order to overcome technological and economic drawbacks of InGaAsP system, Kondow et al. suggested to grow InGaAsN dilute nitrides on GaAs substrates [41]. Indeed, the band gap reduction induced by the introduction of nitrogen in GaAs or InGaAs matrix is related to a strong modification of the conduction band, whereas the valence band remains almost unchanged. This remarkable property could then allow extension of

wavelength emission far above 1 μm while ensuring a strong electron confinement in device active layers. Moreover, due to the much smaller size of nitrogen atom as compared to arsenic atom, the introduction of nitrogen is expected to partly compensate compressive strain in the quantum well.

Many dilute nitride devices have now been fabricated, but the poor material quality related to the introduction of nitrogen did not give rise to the expected commercial success. However, recent developments in dilute nitride laser growth have led to very encouraging results, showing performances comparable to the best lasers on InP substrates with better characteristic temperatures [42]. In the following, we describe one approach to predict and optimize laser properties of InGaAsN/GaAs devices.

6.4.1 Conduction Band of InGaAsN

The design of InGaAsN-based devices requires a deep knowledge of the alloy electronic properties and a development of accurate models. The dramatic band gap reduction induced by the incorporation of nitrogen in the host matrix has been intensively investigated since it was first published [40] and many experimental and theoretical studies have led to a good understanding of the material properties. High hydrostatic pressure experiments performed by Shan et al. have shown that incorporation of small amounts of nitrogen into conventional III–V compounds leads to a splitting of the conduction band into two subbands and an almost unchanged valence band structure [43, 44]. The observed effects were very nicely explained by a phenomenological model called Band AntiCrossing (BAC) considering a strong coupling between the extended conduction band states close to the zone centre and the localised nitrogen states. Many electronic properties of InGaAsN structures such as enlarged electron effective mass are well predicted using this simple two-levels BAC approach [45]. More sophisticated calculations based on the pseudopotential supercell technique have confirmed the localised–delocalised duality of the conduction band edge in III–V Nitride alloys [46–48] (see Chap. 2). However, if these calculations give an accurate description of conduction states, they are not easily expendable for device modelling. Hence, due to the good consistency between BAC model and experimental characterisation, we chose to use this latter. A simple description is presented below.

The localised nitrogen state E_N is resonant with the conduction band $E_M(k)$ and the interaction between these two states is represented by a coupling parameter V_{NM} which is composition dependent. The matrix writes as:

$$\begin{bmatrix} E_M(k) & V_{\mathrm{NM}} \\ V_{\mathrm{NM}} & E_N \end{bmatrix} \tag{6.35}$$

with:

$$V_{\mathrm{NM}} = C_{\mathrm{NM}}\sqrt{y}\ [44] \tag{6.36}$$

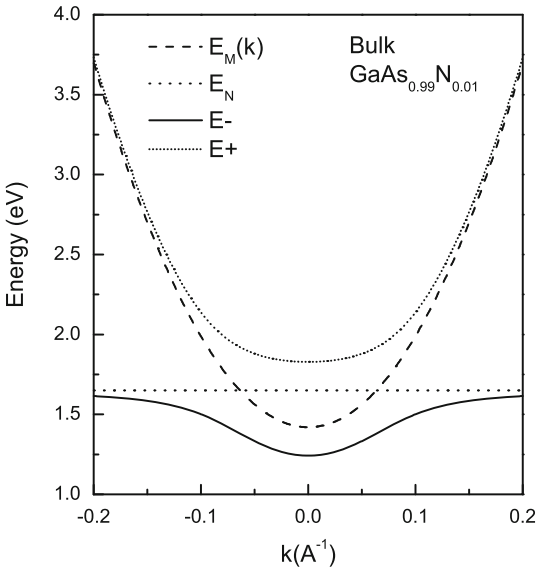


Fig. 6.18 Band anticrossing of GaAs matrix and localized nitrogen states in bulk GaAsN

$$C_{\mathrm{NM}} = 2.7 - 3.2, \text{depends on indium fraction } x \text{ [49]} \tag{6.37}$$

$$E_N = 1.65 - x(0.5 - 0.4x) \text{ [44]} \tag{6.38}$$

The energy of the nitrogen level E_N is assumed to be constant relative to the vacuum level whatever the InGaAsN composition is [44]. The In fraction (x) dependence of E_N in (6.38) reflects the variation of the valence band offset with respect to GaAs [16].

The strong interaction results in a splitting of the conduction band into two subbands E_+ and E_-:

$$E_{\pm}(k) = \frac{1}{2}\left[E_M(k) + E_N \pm \sqrt{(E_M(k) - E_N)^2 + 4C_{\mathrm{NM}}^2 y}\right] \tag{6.39}$$

The new band gap energy is the energy difference between the minimum of E_- conduction subband and the maximum of valence band, which is not affected by the introduction of nitrogen. The BAC-induced conduction band reduction is illustrated in Fig. 6.18 for bulk GaAsN with 0.01 nitrogen fraction using a parabolic dispersion law for GaAs matrix conduction band.

6.4.2 Gain Improvement of InGaAsN Structures is Obtained by . . .

6.4.2.1 . . . Designing Higher Barriers

InGaAsN/GaAs quantum well devices have been fabricated but their temperature performances are still limited. Tansu et al. proposed that apart from Auger recombinations, one of the factors contributing to the temperature sensitivity of InGaAsN/GaAs quantum well lasers could be the hole leakage into the barriers [50] due to a smaller valence band offset in InGaAsN/GaAs system than in InGaAsP/InGaAsP heterostructures. Large band gap GaAsP barriers may be then used in order to increase both the conduction and valence band offsets.

The presence of nitrogen in the quantum well is also responsible for non radiative recombination centres and an enlargement of the linewidth. Thus, the best quality quantum wells are usually obtained for high indium ($x > 0.3$) and low nitrogen contents ($y < 0.01$). These characteristics, combined with the growth of multiple quantum well structures, yield an emission at $1.3\,\mu m$ with a stronger intensity for room temperature operation. However, the number of quantum wells in such structures may be limited because of the high compressive strain (lattice mismatch: $\Delta a/a \approx 2\%$) in the quaternary layer. Hence, growth can be facilitated and the number of quantum wells increased using strain-compensated heterostructures, with barriers under tensile strain such as GaAsP ($\Delta a/a \approx -0.9\%$) for $(GaAs_{0.8}P_{0.2}/GaAs)$, as reported by Kawaguchi et al. [51] or Li et al. [52].

Tansu et al. have shown experimentally that adding phosphorus in the GaAs barrier induces a decrease in the threshold current density [50]. They have fabricated and compared $In_{0.4}Ga_{0.6}As_{0.995}N_{0.005}$/GaAs and $In_{0.4}Ga_{0.6}As_{0.995}N_{0.005}/GaAs_{0.85}P_{0.15}$ quantum well laser structures. The threshold current density increases with temperature for both structures, but the effect is reduced using GaAsP instead of GaAs barriers, and whatever the temperature, the threshold current density is reduced in GaAsP barrier devices (Fig. 6.19a).

Taking the strong coupling between the InGaAs conduction band and the localised nitrogen levels into account, the band structure of $In_{0.4}Ga_{0.6}As_{0.995}N_{0.005}$/GaAs and $In_{0.4}Ga_{0.6}As_{0.995}N_{0.005}/GaAs_{0.85}P_{0.15}$ quantum wells is calculated by solving the Lüttinger–Kohn Hamiltonian, including tetragonal strain and confinement (see Sect. 6.2.2.1). The eigenvalue problem is solved by the transfer-matrix method, taking into account the interfacial discontinuity condition [24, 25]. The valence band material parameters used for the calculations are the ones of InGaAs [16]. The material gain of both active layers is calculated using (6.33). We have reported in Fig. 6.19b the maximum material gain of these two structures as a function of temperature [53]. For both, the material gain decreases as temperature increases, but this trend is reduced for GaAsP barrier heterostructure. This gain improvement when hole confinement is enhanced may impact efficiently on threshold current density reduction.

GaAsP barriers seem very promising for laser applications. However, if the GaAsP barrier is grown as a bulk layer to form the optical confinement region of the laser, the phosphorus content should be optimised in order to both compensate the com-

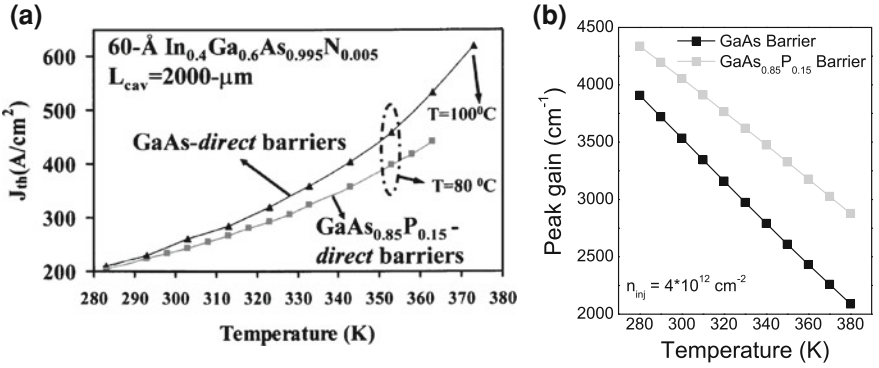


Fig. 6.19 (a) Temperature dependence of threshold current density for InGaAsN/GaAs (*black triangles*) and InGaAsN/GaAsP (*grey squares*) quantum-well lasers and [50] (b) calculated peak gain for InGaAsN/GaAs (*black triangles*) and InGaAsN/GaAsP (*grey squares*) for sheet carrier density $n_{\text{inj}} = 4 \times 10^{12}\,\text{cm}^{-2}$, after [53]

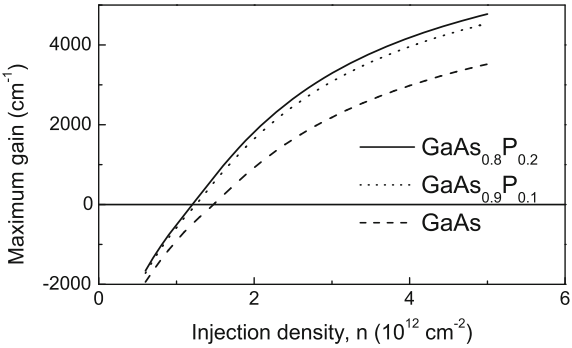


Fig. 6.20 Maximum gain as a function of sheet current density for GaAs (*dashed line*), $GaAs_{0.9}P_{0.1}$ (*dotted line*) and $GaAs_{0.8}P_{0.2}$ (*straight line*) barriers [53]

pressive strain in the QW and minimise the tensile strain in the barrier to prevent plastic relaxation of the material. Taking this into account, the maximum gain of $In_{0.3}Ga_{0.7}As_{0.99}N_{0.01}/GaAs_{1-z}P_z$ QW structures for $z = 0$, $z = 0.1$ and $z = 0.2$ can be compared. The results are shown in Fig. 6.20. It can be seen that the tensile strain in the barrier can be reduced without significantly affecting the material gain. Phosphorus concentrations in the barrier as small as 10% could be enough to fabricate efficient laser structures. The expected gain increase for a sheet current density of $n_{\text{inj}} = 4 \times 10^{12}\text{cm}^{-2}$ is of the order of 30% for phosphorus fractions between 0.1 and 0.2.

6.4.2.2 … and Self-Confinement of Carrier

In 1992, Barrau et al. [22, 54] have shown that coulombic attraction between electrons and holes plays a major role in carrier confinement. When carriers are injected, if one type of carriers is strongly localised in a deep well, as soon as their density increases, they attract more and more the other type of carriers, an effect which modifies the band

profile. The modelling consists in solving the coupled set of Schrödinger equations:

$$\left(-\frac{\hbar^2}{2m_e}\frac{d^2}{dz^2}+V_{\text{ini}}^{C}(z)+V(z)\right)F_m^C(z)=E_m^C F_m^C(z) \text{ for electrons} \tag{6.40}$$

$$\left(-\frac{\hbar^2}{2m_{\text{hh}}}\frac{d^2}{dz^2}+V_{\text{ini}}^{V,h}(z)+V(z)\right)g_n^V(z)=E_n^{V,h}g_n^V(z) \text{ for heavy holes} \tag{6.41}$$

$$\left(-\frac{\hbar^2}{2m_{\text{lh}}}\frac{d^2}{dz^2}+V_{\text{ini}}^{V,l}(z)+V(z)\right)f_n^V(z)=E_n^{V,l}f_n^V(z) \text{ for light holes} \tag{6.42}$$

and Poisson equation:

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2}V(z)=\frac{\mathrm{e}^2}{\varepsilon}\rho(z) \tag{6.43}$$

where m_e, m_{hh} and m_{lh} are the effective masses of electron, heavy and light holes; $V_{\text{ini}}^{C}(z)$, $V_{\text{ini}}^{V,h}(z)$ and $V_{\text{ini}}^{V,l}(z)$ are the initial square potential profiles for conduction band and heavy and light hole valence band; $F_m^C(z)$, $g_n^V(z)$ and $f_n^V(z)$ are the envelope functions of electron, heavy and light holes; and E_m^C, $E_n^{V,h}$ and $E_n^{V,l}$ are the energy eigenvalues for conduction and valence states. $V(z)$ is the induced potential profile.

For Poisson equation, e is the electronic charge, ε is the dielectric constant of the material and $\rho(z)$ is the total charge density in the heterostructure. $\rho(z)$ is calculated as follows:

$$\begin{aligned}\rho(z)=e\Bigg[&-\sum_m\left|F_m^C(z)\right|^2 f(E_m^C)+\sum_n\left|g_n^V(z)\right|^2 f(E_n^{V,h})\\&+\sum_n\left|f_n^V(z)\right|^2 f(E_n^{V,l})\Bigg]\end{aligned} \tag{6.44}$$

where $f(E_m^C)$, $f(E_n^{V,h})$ and $f(E_n^{V,l})$ are the occupancy numbers of conduction and heavy and light hole valence bands.

The algorithm consists in first solving Schrödinger equation in a square potential profile ($V_0(z)=0$), determining carrier wavefunctions and their occupancy numbers and determining Poisson coulombic potential $V_1(z)$ after the total charge density $\rho(z)$ in the square heterostructure. In a second step, the calculated $V_1(z)$ is added to the initial potential $V_{\text{ini}}^{C,V}(z)$. The described algorithm is repeated until $V_i(z)$ converges to an asymptotic value. Note that Poisson potential $V_i(z)$ is modified after each iteration as follows:

$$V_i(z)=V_i(z)+\alpha V_{i-1}(z) \tag{6.45}$$

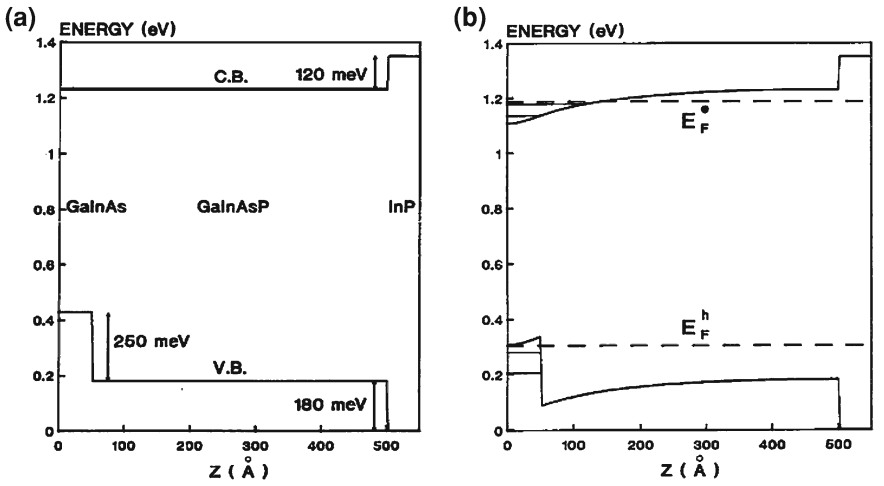


Fig. 6.21 (a) The band profile of a $Ga_{0.67}In_{0.33}As/Ga_{0.20}In_{0.80}As_{0.45}P_{0.55}$/InP undoped structure at thermal equilibrium. The structure is symmetric with respect to the origin. (b) The band profile at sheet carrier density $n_{inj} = 2 \times 10^{12}\,cm^{-2}$ (symmetric with respect to the origin). The first bound levels and the two pseudo Fermi levels are shown [54]

with $\alpha = [0 - 1]$ which is a parameter introduced in order to overcome convergence issues [55].

The effect of carrier self-confinement is highlighted in Fig. 6.21. Barrau et al. have studied a $Ga_{0.67}In_{0.33}As/Ga_{0.20}In_{0.80}As_{0.45}P_{0.55}$/InP heterostructure in which only holes are confined (Fig. 6.21a). For a sheet carrier density of $n_{inj} = 2 \times 10^{12}\,cm^{-2}$, as holes are confined in the valence band quantum well, they create an attractive coulombic potential for electrons, which superimposes to the initial flat band potential while a slightly repulsive potential for holes also appears in the valence band. A confinement effect in the conduction band then appears and electrons are trapped in the self-generated quantum well, giving rise to bound states (Fig. 6.21b). The higher the injected carrier density, the stronger the self-confinement effect of carriers Silver et al. have predicted that lasing could occur even in type II structures [56], this has been shown experimentally at low temperature in InAsSb/lnAs multiple quantum well laser structure emitting in the midwavelength infrared region [57].

In InGaAsN/GaAs(P) system, electrons are strongly confined, whereas holes may spillover towards the barriers due to a small valence band offset. Taking the strong confinement of electrons in the quantum well into account and solving both Poisson and Schrödinger equations, Healy et al. [58] have shown that the attractive potential created by electrons is almost sufficient to reach the highest expected gain. Their calculations are reported in Fig. 6.22. Neglecting self-confinement effects (non self-consistent calculation: NSC), they find an improvement of ~25% of the material gain when using GaAsP barriers instead of GaAs barriers. When solving self-consistently Schrödinger and Poisson equations (self-consistent calculation: SC), the obtained improvement is ~6% only.

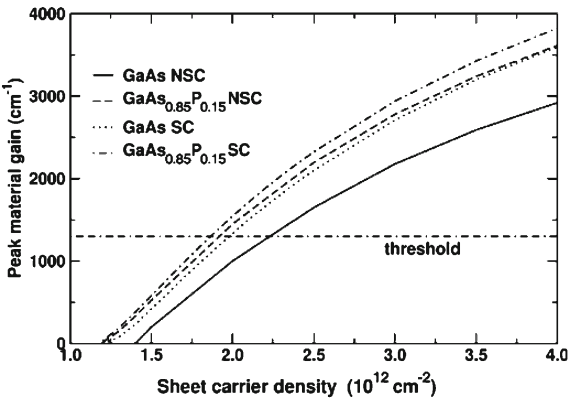


Fig. 6.22 Peak material gain calculated with (SC) and without (NSC) Poisson effects for a 6.4 nm GaInNAs quantum well with GaAs and GaAsP barriers as a function of carrier density at 300 K [58]

Hence, taking attractive coulombic effects into account is necessary to accurately design optoelectronic structures, whereas neglecting them will only give access to broad trends.

6.5 Large Bandwidth Semiconductor Optical Amplifiers

The needs in telecommunications have considerably increased in the past decade, leading to very high speed optical communications between metropolitan centres. By contrast, urban infrastructures have poorly evolved. In order to support the coming expansion of local exchanges such as video on demand, voiceover IP, interconnections with mobile phone applications or shared storage and calculation area networks, coarse-WDM (8-channel/$\Delta\lambda = 20$ nm) standard has been defined as an appropriate low cost solution. Up to now, no amplification is used in these networks, but the increasing communication rate will soon require the use of polarisation-independent broadband (150 nm width at -3 dB) amplifiers operating at 10 Gbits/s. We present in this section the design and optimisation of such semiconductor optical amplifier (SOA) using band structure engineering.

6.5.1 InGaAsP/InP Heterostructures

Due to their mature technology, InP-based SOAs are good candidates for small, cheap and integrated amplifiers. Indeed, quaternary InGaAsP can be grown lattice-matched, or under tensile or compressive strain on InP substrate, allowing to monitor light polarization (see Fig. 6.23). For instance, ternary alloy $In_{0.53}Ga_{0.47}As$ ($E_g = 0.75$ eV) is lattice-matched to InP. Tensile strain can be obtained either by adding phosphorus element, or decreasing indium content, and at the same time the band gap energy will be increased. However, if indium and arsenic have almost the same effect on band gap energy variation, strain is more impacted by indium than by

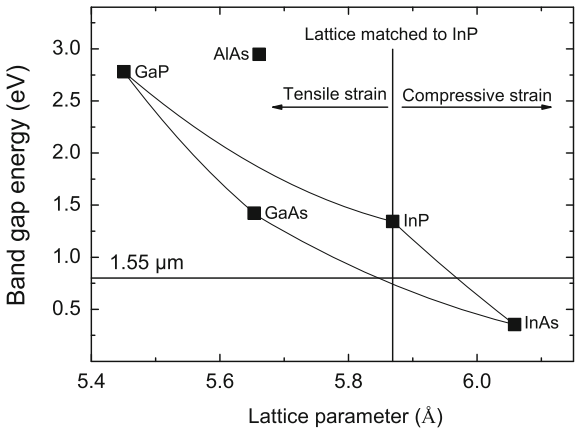


Fig. 6.23 Band gap energy of GaP, GaAs, InP and InAs binary and related ternary alloys as a function of lattice parameter. Note that for GaP and AlAs the energy differences between conduction and valence bands at Γ point are reported

arsenic variation. Then, aimed wavelength and strain values are reached by tuning the four element contents of InGaAsP quaternary.

Commercial SOAs using InGaAsP bulk or multiple quantum well active layers provide appropriate gain insensitivity to light polarisation, but their optical bandwidth is typically restricted to about 60–80 nm [59, 60]. However, due to the large number of parameters that can be tuned, multiple quantum well active layers are more suitable for wide bandwidth amplification applications. In the following, we describe the optimisation of multiple quantum well active layer in the view of reaching an optical bandwidth of 150 nm, with a polarisation-dependent gain (TE/TM ratio) lower than 2 dB.

6.5.2 How to Realize Polarisation-Independent Gain?

As previously described (see Sect. 6.3), E-HH transitions give rise to TE emission when TM polarisation originates from E-LH transitions. In order to have equal TE and TM emission contributions to the total material gain, the amplifier active layer must provide either both compressive and tensile strained quantum wells, or only one quantum well type in which first heavy and light hole states have almost the same energy level.

6.5.2.1 Typical InGaAsP Quantum Wells

We have represented the band alignment of InGaAsP heterostructures used in SOA active layers in Fig. 6.24. The barrier material can be either strained or lattice-matched to InP substrate. When strained multiple quantum wells are stacked, barriers under opposite strain can be used to counterbalance the total elastic energy and avoid

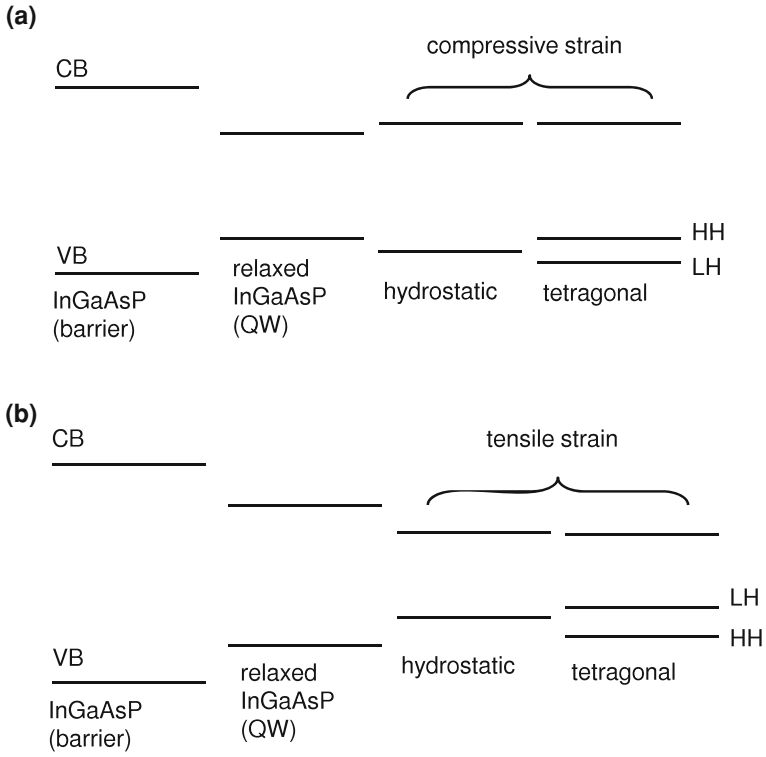


Fig. 6.24 Band alignment of InGaAsP/InGaAsP quantum well under (a) compressive and (b) tensile strain

relaxation. In Fig. 6.24a is reported the band alignment for compressively strained quantum well material, in this case, the heavy-hole band lies at the top of the valence band (see Sect. 6.2.1). On the contrary, in the case of tensile strain, the valence band maximum is a light-hole state (Fig. 6.24b). However, the first hole level in the quantum well is not only related to the valence band maximum; confinement, which is strongly mass-dependent, has to be taken into account in order to predict TE or TM polarisation of the electron-hole transitions.

6.5.2.2 Strain and Confinement Balance

In $k_{\parallel} = 0$, confinement energy E_p of level p, relative to the valence band maximum (VB_{HH} for heavy-hole states and VB_{LH} for light-hole states), writes in the simplest approach as:

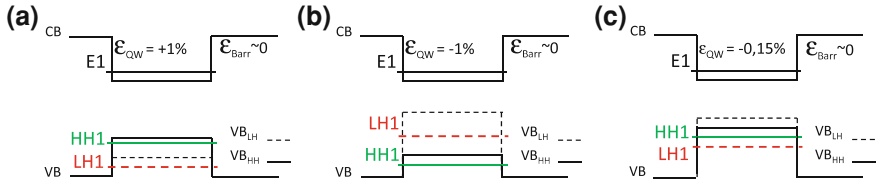


Fig. 6.25 Band alignment and hole confinement energies in InGaAsP/InGaAsP quantum wells under (a) compressive strain $\varepsilon_{QW} \sim +1\%$, (b) tensile strain $\varepsilon_{QW} \sim +1\%$ and (c) tensile strain $\varepsilon_{QW} \sim -0.15\%$. The barrier is lattice-matched to the InP substrate; its composition is $In_{0.8}Ga_{0.2}As_{0.45}P_{0.55}$

$$E_p = \frac{\hbar^2 \pi^2}{2m^* L^2} p^2, \quad p = 1, 2 \ldots \tag{6.46}$$

m^* is the effective mass of the holes, and L is the quantum well width. Given that the mass of heavy holes in the quantum well is larger that that light holes (along the growth axis), even in the case of tensile strain, the first quantised hole state in the quantum well might a heavy-hole state due to a more efficient confinement effect. We have reported in Fig. 6.25 the three different possibilities, for typical strain values: in the case of compressive strain $(\varepsilon_{QW} \sim +1\%)$, the first hole state in the quantum well is a heavy hole (a); in the case of sufficiently high tensile strain $(\varepsilon_{QW} \sim +1\%)$, the valence band maximum and the first hole state are light hole (b). On the contrary, in the case of slight tensile strain $(\varepsilon_{QW} \sim -0.15\%)$, in spite of a light-hole valence band maximum, the first hole state is a heavy hole (c).

6.5.2.3 Polarisation-Independent Operation

A very efficient solution to keep TE and TM mode equal over the −3 dB gain bandwidth consists in stacking both tensile and compressive quantum wells in the active layer. A simplified scheme of band line-up of such structures is reported in Fig. 6.26. These heterostructures with ∼1% tensile and compressive strain have shown very good results in terms of polarisation dependence both in the 1.55 μm range [61–63] and 1.3 μm range [64, 65]. However, the −3 dB gain bandwidth was limited to 70 nm. The amplified spontaneous emission (ASE) spectra of a 1.55 μm SOA containing five compressive quantum wells (5C, $L = 100$ Å) and four tensile wells (4T, $L = 140$ Å) are reported in Fig. 6.27 [62].

6.5.3 How to Increase Bandwidth?

In order to increase the optical bandwidth it has been suggested to insert multiple quantum wells with the same composition but with different well widths [66, 67].

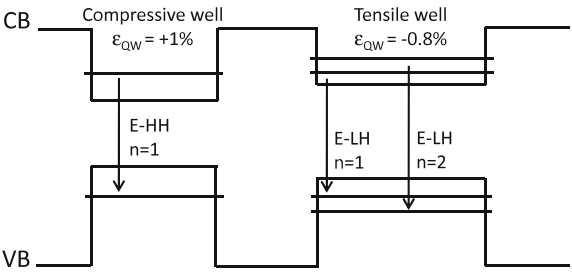


Fig. 6.26 Simplified band alignment of an active layer including both compressive and tensile strained quantum wells, after Newkirk et al. [61]

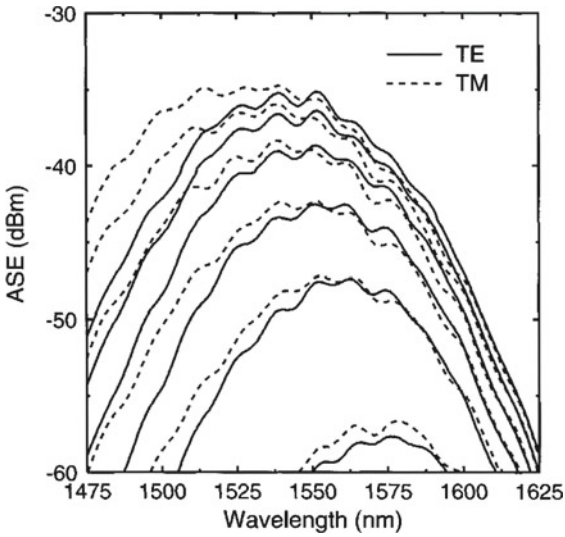


Fig. 6.27 Experimentally measured polarisation-independent amplified spontaneous emission (ASE) for a 5C+4T structure. The injected currents are from 50 to 300 mA at 50-mA intervals [62]

However, significant differences between well widths may induce appreciable difference in the density of states, and carrier redistribution within the active layer (that can be predicted by resolving self-consistently Schrödinger and Poisson equations [56, 62]) will occur, resulting in a narrower optical bandwidth than expected. Another approach for increasing bandwidth consists in using transitions between fundamental ($n = 1$) and excited states ($n \geq 2$) of both electrons and holes in the quantum well. Using this technique, Miller et al. have shown that amplification could be realised over about 100 nm bandwidth with 1% tensile-strain quantum wells [68]. In this study, due to the high value of tensile strain, light emission is due to the transition between electrons and light hole states, and is strongly TM polarised (see Fig. 6.28). However, this solution can be extended to slightly strained quantum wells, in which first heavy- and light-hole state levels have almost the same energy value [69]. Then, similar to bulk materials in which hole states are degenerated, TE = TM condition can be carried out and in the meantime, bandwidth enlargement becomes possible thanks to quantisation effects.

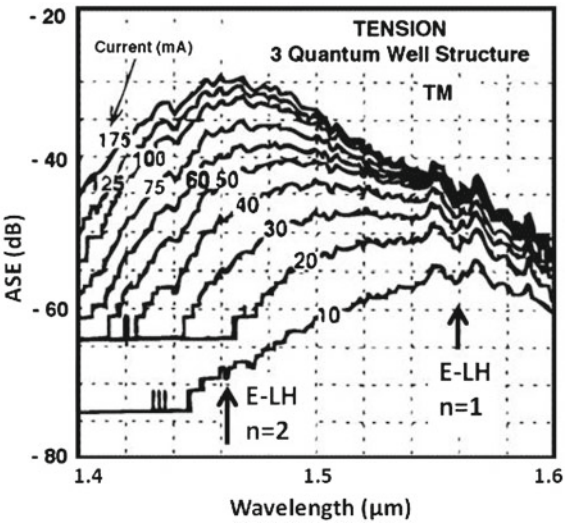


Fig. 6.28 Amplified spontaneous emission (ASE) spectra of tensile strained quantum well structures [68]

6.5.4 Band Structure and Gain

As discussed previously, suitable quantum well for TE = TM emission must be wide and slightly tensile strained in order to provide LH1∼HH1.

The chosen quantum well consists of a 14 nm $In_{0.53}Ga_{0.47}As_{0.96}P_{0.04}$ layer which is tensile strained on InP substrate, with a lattice-mismatch equal to −0.15%. The barrier material consists of $In_{0.8}Ga_{0.2}As_{0.45}P_{0.55}$ which is almost lattice-matched (0.05% compressive strain; quaternary $In_{0.8}Ga_{0.2}As_{0.45}P_{0.55}$ is named Q1.17 following its bulk emission wavelength at 1.17 μm) to InP substrate. These compositions have been chosen for several reasons: (i) the quantum well material band gap corresponds to an emission wavelength around 1.55 μm, (ii) the barrier material offers a good confinement for both electrons and holes, (iii) strains in the quantum well and barrier are opposite, which partly compensates the total strain in the device and (iv) a low strain value in the quantum well, which reduces the heavy/light hole energy splitting.

The band structure of InGaAsP/InGaAsP quantum well is calculated by solving the Lüttinger–Kohn Hamiltonian, including tetragonal strain and confinement effects. The eigenvalue problem is solved by the transfer-matrix method, taking into account the interfacial discontinuity condition [24, 25, 70, 71]. The band structure of this quantum well is reported in Fig. 6.29. Conduction states show almost parabolic dispersion laws, on the contrary, hole states are strongly non-parabolic, due to the strong heavy–light hole state mixing when $k_{//} > 0$.

The material gain has been calculated from the dispersion curves and the oscillator strengths of the different optical transitions [18]. We have plotted in Fig. 6.30 the calculated gain spectra for three carrier injection densities: $n_1 = 0.9 \times 10^{13}\,\mathrm{cm}^{-2}$, $n_2 = 1.2 \times 10^{13}\,\mathrm{cm}^{-2}$ and $n_3 = 2.3 \times 10^{13}\,\mathrm{cm}^{-2}$. Both TE (solid line) and TM (dashed

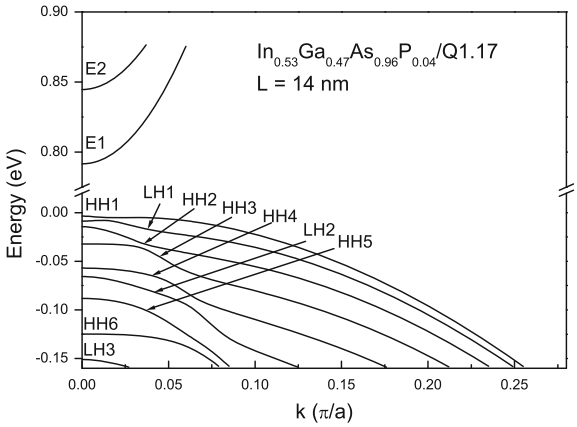


Fig. 6.29 Band structure of 14-nm $In_{0.53}Ga_{0.47}As_{0.96}P_{0.04}$/Q1.17 quantum well

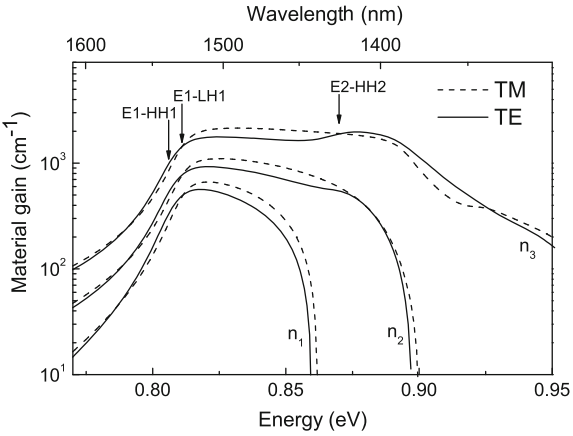


Fig. 6.30 Material gain of 14-nm $In_{0.53}Ga_{0.47}As_{0.96}P_{0.04}$/ $Q1.17$ quantum well for $n_1 = 0.9 \times 10^{13}cm^{-2}$, $n_2 = 1.2 \times 10^{13}cm^{-2}$ and $n_3 = 2.3 \times 10^{13}cm^{-2}$ carrier injection densities [72]

line) contributions are reported, the largest TM/TE ratio is about 1 dB on the −3 dB gain bandwidth. When carrier injection density is increased, population inversion on excited states of the quantum well is achieved, and transitions between $n = 1$ and $n \geq 2$ electron and hole levels are observed which leads to a significant increase of the optical bandwidth. All the involved transitions are labelled at their energy position on the calculated spectra. The lower energy transition occurs between first electron and first heavy hole levels (E1-HH1 at 0.806 eV; $\lambda = 1.538\,\mu m$), leading to a predominant TE gain, the second transition between first electron and first light hole levels (E1-LH1 at 0.811 eV−$\lambda = 1.528\,\mu m$) induces the rise of TM emission. Note that the gain polarisation insensitivity in the whole bandwidth can be achieved here thanks to the very large heavy–light hole mixing for all the valence subbands when $k > 0$.

Based on this band structure engineering, SOAs have been fabricated and characterised [72]; TE (solid line) and TM (dashed line) amplified spontaneous emis-

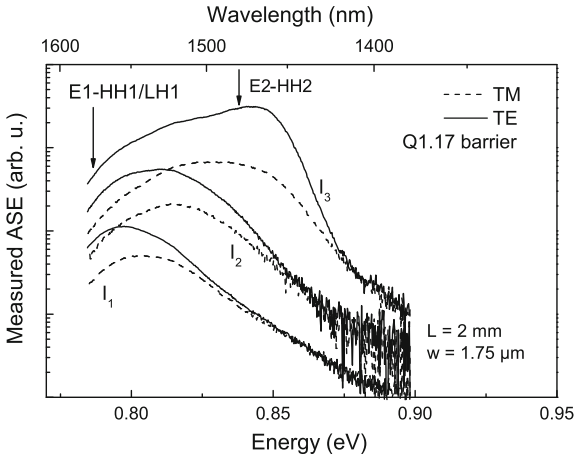


Fig. 6.31 Measured amplified spontaneous emission (ASE) power of 14-nm $In_{0.53}Ga_{0.47}As_{0.96}P_{0.04}$/Q1.17 quantum-well semiconductor optical amplifier for $I_1 = 50\,mA$, $I_2 = 100\,mA$ and $I_3 = 200\,mA$ [72]

sion are reported in Fig. 6.31. The same features as the ones given by the material gain calculation are observed experimentally. Due to their low energy splitting (E1 − LH1–E1-HH1 = 5 meV), E1-HH1 and E1-LH1 transitions cannot be distinguished experimentally at operating temperature. For the experimental spectra we have only reported the predominant transitions, i.e. E1-HH1/LH1 and E2-HH2. The calculated energy splitting between these transitions is 64 and 59 meV regarding E1-HH1 and E1-LH1, respectively. These values are in good agreement with the experimentally measured splitting, which is about 55 meV. The large difference between TE and TM measured amplified spontaneous emission is due to the device geometry which yields a larger amplification for TE optical mode and can be adjusted by improving the waveguide design [73]. It is important to note that TE/TM ratio is constant over a large optical bandwidth. An increase in the injection current (i.e. populating $n \geq 2$ conduction and valence states) results in a significant increase in amplified spontaneous emission bandwidth. The maximum measured amplified spontaneous emission bandwidth value of 98 nm is obtained for an injected current of 200 mA (Fig. 6.31). Another solution to further extend the bandwidth would be for instance to increase the splitting between the first and second electron levels, by reducing the quantum well width [72].

This chapter aimed to give an overview of the broad outlines of band structure engineering. We focused on optoelectronic applications and showed that device characteristics could be optimised by controlling microscopic parameters such as strain, composition and confinement. The exposed models and examples are of course non exhaustive and the literature dedicated to semiconductor physics and related applications is immensely rich for readers who wish to learn further.

References

1. E. Yablonovitch, E.O. Kane, J. Lightwave Technol. **6**, 1292 (1988)
2. E.P. O'Reilly, A.R. Adams, IEEE J. Quantum Electron. **30**, 366 (1994)
3. P.J.A. Thijs, L.F. Tiemeijer, J.J.M. Binsma, T.V. Dongen, IEEE J. Quantum Electron. **30**, 477 (1994)
4. A.R. Adams, Electron. Lett. **22**, 249 (1986)
5. E. Yablonovitch, E.O. Kane, J. Lightwave Technol. **LT-4**, 504 (1986)
6. L. Goldstein, M. Quillec, K. Rao, P. Henoc, J.M. Masson, J.Y. Marzin, J. de Physique **43**, 201 (1982)
7. L. Goldstein, F. Glas, J.Y. Marzin, M.N. Charasse, G. Leroux, Appl. Phys. Lett. **47**, 1099 (1985)
8. P. Voisin, M. Voos, J.Y. Marzin, M.C. Talargo, R.E. Nahory, A.Y. Cho, Appl. Phys. Lett. **48**, 1476 (1986)
9. W.D. Laidig, P.J. Caldwell, Y.F. Lin, C.K. Peng, Appl. Phys. Lett. **44**, 653 (1984)
10. P.J.A. Thijs, T.V. Dongen, Electron. Lett. **25**, 1735 (1989)
11. M.L. Lee, E.A. Fitzgerald, M.T. Bulsara, M.T. Currie, A. Lochtefeld, J. Appl. Phys. **97**, 011101 (2005)
12. J.W. Matthews, A.E. Blakeslee, J. Cryst. Growth **27**, 118 (1974)
13. T.G. Andersson, Z.G. Chen, V.D. Kulakovskii, A. Uddin, J.T. Vallin, Appl. Phys. Lett. **51**, 752 (1987)
14. M.G.A. Bemard, B. Duraffourg, Phys. Status. Solidi **I**, 699 (1961)
15. G. Fuchs, J. Horer, A. Hangleiter, V. Harle, F. Scholz, R.W. Glew, L. Goldstein, Appl. Phys. Lett. **60**, 231 (1992)
16. X. Marie, J. Barrau, B. Brousseau, Th. Amand, M. Brousseau, E.V.K. Rao, F. Alexandre, J. Appl. Phys. **69**, 812 (1991)
17. X. Marie, Ph.D. thesis, INSA, Toulouse, 1991
18. G. Bastard, Wave mechanics applied to semiconductor heterostructures. Les Editions de physique (1992)
19. G. Fishman, Semiconducteurs: les bases de la théorie k.p. Les Editions de l'Ecole Polytechnique (2010)
20. E.L. Ivchenko, G. Pikus, Superlattices and other heterostructures: symmetry and optical phenomena. Springer Ser. Solid-State Sci. **110**, 73 (1995)
21. C. Weisbuch, B. Vinter, Quantum Semiconductor Structures: Fundamentals and Applications (Academic Press, London, 1991)
22. J. Barrau, B. Brousseau, B. Calvas, J.Y. Emery, R.J. Simes, C. Starck, L. Goldstein, Electron. Lett. **28**, 551 (1992)
23. O. Issanchou, J. Barrau, X. Marie, J.Y. Emery, C. Fortin, L. Goldstein, IEEE J. Quantum Electron. **33**(12), 2277 (1997)
24. R. Epenga, M.F.H. Schuurmans, S. Colak, Phys. Rev. B **36**, 1554 (1987)
25. D. Ahn, S.L. Chuang, IEEE J. Quantum Electron. **26**, 13 (1990)
26. O. Gilard, F. Lozes-Dupuy, G. Vassilieff, J. Barrau, P. Le Jeune, J. Appl. Phys. **84**, 2705 (1998)
27. O. Gilard, Ph.D. thesis, Université Paul Sabatier, Toulouse, 1999
28. G. Fuchs, C. Schiedel, A. Hangleiter, V. Härle, F. Scholz, Appl. Phys. Lett. **62**, 396 (1993)
29. M. Silver, E.P. O'Reilley, A.R. Adams, IEEE J. Quantum Electron. **33**, 1557 (1997)
30. B.I. Miller, U. Koren, M.G. Young, D.M. Chien, Appl. Phys. Lett. **58**, 1952 (1991)
31. Y. Arakawa, A. Yariv, IEEE J. Quantum Electron. **21**, 1666 (1985)
32. W.L. Li, Y.K. Su, D.H. Jaw, IEEE J. Quantum Electron. **33**, 416 (1997)
33. P.M. Enders, IEEE J. Quantum Electron. **33**, 4 (1997)
34. S.L. Chuang, J. O'Gorman, A.F.J. Levi, IEEE J. Quantum Electron. **29**, 1631 (1993)
35. S.L. Chuang, IEEE J. Quantum Electron. **32**, 1791 (1996)
36. C.W. Hu, F.M. Li, K.F. Huang, M.C. Wu, C.L. Tsai, Y.H. Huang, C.C. Lin, IEEE Photonics Technol. Lett. **18**, 1551 (2006)

37. C.H. Zah, R. Bhat, B.N. Pathak, F. Favire, W. Lin, M.C. Wang, N.C. Andreadakis, D.M. Hwang, M.A. Koza, T.P. Lee, Z. Wang, D. Darby, D. Flanders, J.J. Hsieh, IEEE J. Quantum Electron. **30**, 511 (1994)
38. R. Paoletti, M. Agresti, D. Bertone, L. Bianco, C. Bruschi, A. Buccieri, R. Campi, C. Dorigoni, P. Gotta, M. Liotti, G. Magnetti, P. Montangero, G. Morello, C. Rigo, E. Riva, G. Rossi, D. Soderstrom, A. Stano, P. Valenti, M. Vallone, M. Meliga, J. Lightwave Technol. **24**, 142 (2006)
39. K. Shinoda, S. Makino, T. Kitatani, T. Shiota, T. Fukamachi, M. Aoki, IEEE J. Quantum Electron. **45**, 1201 (2009)
40. M. Weyers, M. Sato, H. Ando, Jpn. J. Appl. Phys. Part 2 **31**, L853 (1992)
41. M. Kondow, K. Uomi, A. Niwa, T. Kitatani, S. Wakahiki, Y. Yazawa, Jpn. J. Appl. Phys. **35**, 1273 (1996)
42. S.M. Wang, G. Adolfsson, H. Zhao, Y.X. Song, M. Sadeghi, J. Gustavsson, P. Modh, A. Haglund, P. Westbergh, A. Larsson, Phys. Status Solidi (b) **248**, 1207 (2011)
43. W. Shan, W. Walukiewicz, J.W. Ager III, E.E. Haller, J.F. Geisz, D.J. Friedman, J.M. Olson, S.R. Kurtz, Phys. Rev. Lett. **82**, 1221 (1999)
44. W. Shan, W. Walukiewicz, K.M. Yu, J.W. Ager III, E.E. Haller, J.F. Geisz, D.J. Friedman, J.M. Olson, S.R. Kurtz, H.P. Xin, C.W. Tu, Phys. Status Solidi (b) **223**, 75 (2001)
45. Z. Pan, L.H. Li, Y.W. Lin, Q. Dun, D.S. Jiang, W.K. Ge, Appl. Phys. Lett. **78**, 2217 (2001)
46. K. Kim, A. Zunger, Phys. Rev. Lett. **86**, 2609 (2001)
47. P.R.C. Kent, A. Zunger, Phys. Rev. Lett. **86**, 2613 (2001)
48. E. O'Reilly, A. Lindsay, S. Tomic, M. Kamal-Saadi, Semicond. Sci. Technol. **17**, 870 (2002)
49. R.J. Potter, N. Balkan, X. Marie, H. Carrère, E. Bedel, G. Lacoste, Phys. status solidi (a) **187**, 623 (2001)
50. N. Tansu, J.Y. Yeh, L.J. Mawst, Appl. Phys. Lett. **83**, 2112 (2003)
51. M. Kawaguchi, T. Miyamoto, A. Saitoh, F. Koyama, Jpn. J. Appl. Phys. **43**, L267 (2004)
52. W. Li, J. Turpeinen, P. Melanen, P. Savolainen, P. Uusimaa, M. Pessa, J. Cryst. Growth **230**, 533 (2001)
53. H. Carrère, X. Marie, J. Barrau, T. Amand, Appl. Phys. Lett. **86**, 071116 (2005)
54. J. Barrau, T. Amand, M. Brousseau, R.J. Simes, L. Goldstein, J. Appl. Phys. **71**, 5768 (1992)
55. F. Stern, J. Comput. Phys. **6**, 56 (1970)
56. M. Silver, E.P. O'Reilly, IEEE J. Quantum Electron. **30**, 547 (1994)
57. A. Wilk, M. El Gazouli, M. El Skouri, P. Christol, P. Grech, A.N. Baranov, A. Joullié, Appl. Phys. Lett. **77**, 2298 (2000)
58. S.B. Healy, E.P. O'Reilly, IEEE J. Quantum Electron. **42**, 608 (2006)
59. M.J. Connelly, Opt. Quantum Electron. **38**, 1061 (2006)
60. S. Tanaka, K. Morito, Appl. Phys. Lett. **97**, 261104 (2010)
61. M.A. Newkirk, B.I. Miller, U. Koren, M.G. Young, M. Chien, R.M. Jopson, C.A. Burrus, IEEE Photon. Tech. Lett. **4**, 406 (1993)
62. M. Silver, A.F. Phillips, A.R. Adams, P.D. Greene, A.J. Collas, IEEE J. Quantum Electron. **36**, 118 (2000)
63. R. Matei, R. Maciejko, Y. Lizé, Conference on Lasers and Electro-Optics proceedings 3, 1990 (2005)
64. L.F. Tiemeijer, P.J.A. Thijs, T. van Dongen, R.W.M. Slootweg, J.M.M. van der Heijden, J.J.M. Binsma, M.P.C.M. Krijn, Appl. Phys. Lett. **62**, 826 (1993)
65. J. Jin, D. Tian, J. Shi, T. Li, Semicond. Sci. Technol. **19**, 120 (2004)
66. X. Zhu, D.T. Cassidy, M.J. Hamp, D.A. Thompson, B.J. Robinson, Q.C. Zhao, M. Davies, IEEE Photonics Technol. Lett. **9**, 1202 (1997)
67. H. Carrère, V.G. Truong, X. Marie, T. Amand, B. Urbaszek, R. Brenot, F. Lelarge, B. Rousseau, Microelectron. J. **40**, 827 (2009)
68. B.I. Miller, U. Koren, M.A. Newkirk, M.G. Young, R.M. Jopson, R.M. Derosier, M.D. Chien, IEEE Photonics Technol. Lett. **5**, 520 (1993)
69. P. Koonath, K.S. Kim, W.-J. Cho, A. Gopinath, IEEE J. Quantum Electron. **38**, 1282 (2002)

70. X. Marie, J. Barrau, T. Amand, H. Carrère, C. Fontaine, E. Bedel-Pereira, IEE Proc. Optoelectron. **150**, 25 (2003)
71. H. Carrère, X. Marie, J. Barrau, T. Amand, S. Ben Bouzid, V. Sallet, J.-C. Harmand, J. Phys.-Condens. Matt. **16**, S3215 (2004)
72. H. Carrère, V.G. Truong, X. Marie, R. Brenot, G. De Valicourt, F. Lelarge, T. Amand, Appl. Phys. Lett. **97**, 121101 (2010)
73. M. Itoh, Y. Shibata, T. Kakitsuka, Y. Kadota, Y. Tohmori, IEEE Photonics Technol. Lett. **14**, 765 (2002)

Chapter 7
Fundamental Theory of Semiconductor Lasers and SOAs

Mike J. Adams

Abstract This chapter aims to give a basic understanding of semiconductor lasers and semiconductor optical amplifiers (SOAs). Starting from the underlying physics of radiative emission, together with the elements of optical waveguide theory, simple approximations are found for optical gain, lasing threshold and cavity resonances. Rate equations are used to elucidate time-dependent laser behaviour and, in combination with a travelling-wave equation for spatial photon distribution, to describe the effects of saturation and crosstalk in SOAs.

7.1 Review of Key Concepts

This section is intended to give a review of the most significant parts of semiconductor theory that are needed for subsequent use in the remainder of the chapter. More details on specific aspects of some of these topics will be found in other chapters of this book. From the viewpoint of understanding semiconductor laser and optical amplifier behaviour, a brief discussion of radiative transitions will be presented first. The transitions of main interest are those that involve recombination of electrons in the conduction band with holes in the valence band. The topic of (unipolar) quantum cascade lasers [1] where the transitions are between states in conduction subbands formed as a result of size quantisation will not be discussed here.

M. J. Adams (✉)
School of Computer Science and Electronic Engineering,
University of Essex, Wivenhoe Park,
Colchester CO4 3SQ, UK
e-mail: adamm@essex.ac.uk

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_7,

7.1.1 Radiative Transitions

Radiative transitions are subject to requirements of conservation of energy and electron wavevector. The first of these requirements means that the photon energy $h\nu$ is equal to the difference between the upper and lower electron energy states involved in the transition. The second requirement states that the wave vector $\boldsymbol{k}_p$ of the photon is equal to the difference between the wave vectors of the two electron states. In practice, for a visible or near-infrared photon, the magnitude of $\boldsymbol{k}_p$ is of the order of $10^7\ \mathrm{m}^{-1}$, whereas the magnitudes of the electron wave vectors are typically at least a hundred times this value. Hence, $\boldsymbol{k}_p$ is usually negligible by comparison with the electron wave vectors, and this implies vertical transitions on the energy wavenumber diagram. The consequence of this for radiative emission is that direct-gap semiconductors are inherently more suitable as candidates for emitters than indirect-gap materials. This follows since it is usually only possible to pump electrons into the *lowest* conduction band minimum and holes into the valence band maximum. Thus in indirect-gap materials any radiative transition must involve an extra particle (usually a phonon) to provide conservation of wave vector, and this makes transitions less probable than single-particle transitions in direct-gap materials. In this context it is worth noting that the recent announcement of a Ge-on-Si laser [2] was accomplished by the use of tensile strain and *n*-type doping in order to compensate the energy difference between the direct and indirect conduction band minima. This example of band engineering caused the Ge material to behave like a direct-gap material so that optically pumped lasing could be achieved.

7.1.2 Spontaneous and Stimulated Emission

It is important to distinguish between the processes of spontaneous and stimulated emission. Let f_c and f_v be the occupation probabilities of states in the conduction and valence bands, respectively. Then the rates of spontaneous and stimulated emission are proportional to $f_c(1 - f_v)$ and $(f_c - f_v)$, respectively. To calculate the emission rates it is necessary to sum over all states that can emit a photon of energy $h\nu$, subject to the conservation of energy and wave vector discussed above. In the case of parabolic bands in a bulk semiconductor this leads to particularly simple expressions for the rates $r_{\mathrm{sp}}(h\nu)$, $r_{\mathrm{st}}(h\nu)$ of spontaneous and stimulated emission per unit energy per unit volume

$$r_{\mathrm{sp}}(h\upsilon) = P(h\upsilon - E_G)^{1/2} f_c(1 - f_v) \tag{7.1}$$

$$r_{\mathrm{st}}(h\upsilon) = P(h\upsilon - E_G)^{1/2}(f_c - f_v) \tag{7.2}$$

where E_G is the energy gap and the coefficient P has a weak dependence on photon energy. The reason for writing these equations in this simplified form is to gain insight

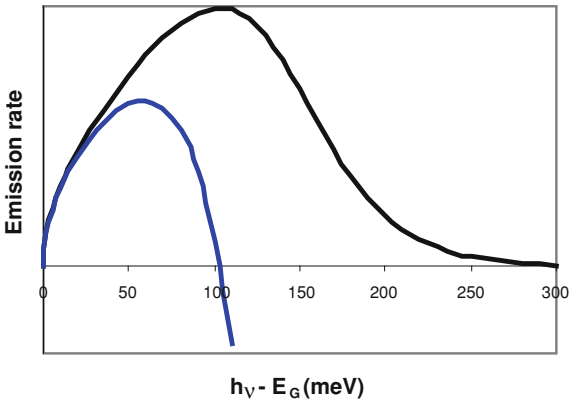


Fig. 7.1 Spontaneous (*black*) and stimulated (*blue*) emission spectra calculated for parabolic bands with k-conservation

into the spontaneous and stimulated spectra and their dependence on carrier density. The process is completed by noting that the occupation probabilities are given by quasi-Fermi distributions with quasi-Fermi levels F_c, F_v for states in the conduction and valence bands, respectively. The concentrations n, p of electrons and holes can be found from F_c and F_v in the usual way.

Figure 7.1 illustrates the difference between spontaneous and stimulated spectra calculated as described above. Spontaneous emission is a broadband process whose spectrum has a high-energy tail corresponding to that of the quasi-Fermi distribution functions in (7.1). However, the stimulated spectrum changes sign from positive to negative at a value of photon energy determined, from (7.2), by the condition $f_c = f_v$. When the quasi-Fermi distributions are substituted in this condition, the Bernard–Duraffourg condition [3] for population inversion (positive stimulated emission) is found.

$$F_c - F_v > h\upsilon \tag{7.3}$$

This is a necessary (but not sufficient) condition for lasing in semiconductors and holds independent of the model used to calculate the stimulated emission rate. On the low-energy side, both spontaneous and stimulated emission spectra are limited by the term $(h\nu - E_G)^{1/2}$ in (7.1) and (7.2); this term arises from the joint density-of-states for the transitions.

7.1.3 Optical Gain

In a semiconductor laser, the optical gain per unit length is a more useful parameter than the stimulated emission rate, but the two are linearly related and the spectra are very similar. Figure 7.2 illustrates the dependence of the gain spectrum on electron density, calculated by assuming charge neutrality ($n = p$). In this case, wavelength λ is used instead of photon energy; the relation between the two is

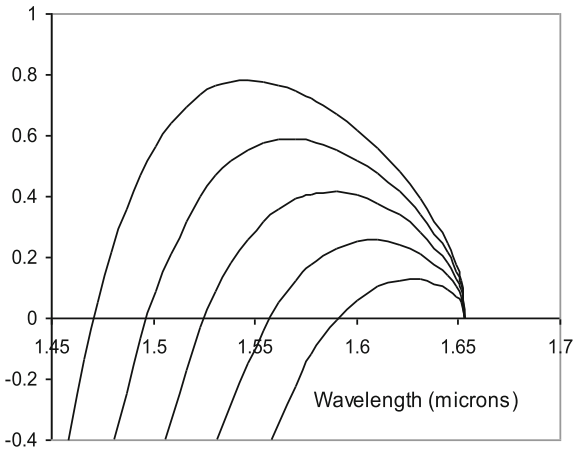


Fig. 7.2 Dependence of the gain spectrum on electron density calculated assuming charge neutrality ($n = p$). Values of n are 1.2×10^{18}, 1.4×10^{18}, 1.6×10^{18}, 1.8×10^{18} and $2.0 \times 10^{18}\,\mathrm{cm}^{-3}$ (lowest to highest *curve*)

$\lambda(\mu\mathrm{m}) = 1.24/hv(\mathrm{eV})$. In Fig. 7.2 the wavelengths of the gain maxima exhibit a linear dependence on n, and in addition it is found that a linear relation is also a good approximation for the dependence of peak gain on n. The latter observation leads to the widely used approximation for the peak gain g_m

$$g_m = a(n - n_o) \tag{7.4}$$

where a is the differential gain and n_o is the transparency concentration.

When quantum well (QW) material is used in semiconductor lasers, the well-known step-like density of states has a very beneficial effect in that more carriers are available at energies close to the (effective) band edge than in the case of bulk (3-D) semiconductors. This makes QWs more efficient at generating optical gain than bulk semiconductors (more gain per electron). Figure 7.3 shows QW gain spectra calculated using again the simplest single subband model with wave vector conservation and parabolic bands. Although this model is not rigorously accurate, from the results two general observations can be made which aid our understanding of QW lasers: (1) the wavelength shift with n is much less in QW than in bulk and (2) gain saturation is much stronger in QW than in bulk. As a result the peak gain variation with n is different and a better approximation is given by [4]

$$g_m = g_o \ell n \left(\frac{n}{n_o} \right) \tag{7.5}$$

where g_o is the gain at transparency.

It should be stressed that the models for gain and recombination discussed above are very simple and represent the minimum necessary to proceed to a description of lasers and SOAs. More sophisticated models for these processes, taking into account

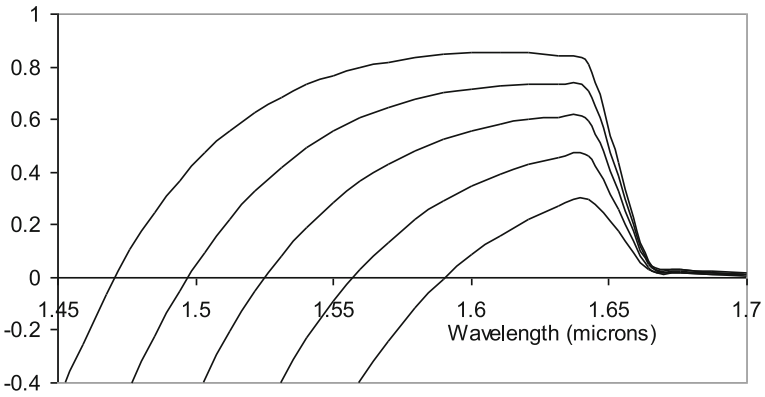


Fig. 7.3 QW gain spectra calculated for single subband model with k-conservation and parabolic bands for the same values of n as in Fig. 7.2

more accurate band structure information, higher subbands in QWs, many-body effects, etc. can be found in a number of textbooks [5–7].

7.2 Semiconductor Laser Structures

7.2.1 Heterostructures

The condition (7.3) for stimulated emission implies that a high concentration of electrons and holes must be present simultaneously as a prerequisite for lasing action. The first semiconductor lasers achieved this condition by a strongly forward-biased p–n junction; the quasi-Fermi level separation $(F_c - F_v)$ is then equal to the voltage dropped across the junction multiplied by the electron charge. The diode can thus produce population inversion in the vicinity of the junction, but it has a poorly defined active region and suffers from carrier wastage due to escape into the neighbouring n- and p-regions. As a consequence, the current densities required to produce lasing were extremely high (of the order of 10^6 A/cm^2) and continuous wave (cw) operation at room temperature was not possible. In order to achieve better confinement of electrons and holes to the active region, a double heterostructure is now used. In this structure a thin layer of lower band-gap material (e.g. GaAs) is placed between materials of higher band gap (e.g. AlGaAs); the heterojunctions form barriers to electrons and holes and thus confine the carriers to the central active layer. This structure, using GaAs–AlGaAs materials, was the first to produce cw lasing at room temperature in 1970 [8, 9]. The emission wavelength was determined by the energy gap of GaAs (1.43 eV) to lie around 0.85 μm.

When lasers are used as sources for optical fibre communication systems, the wavelengths of interest are dictated by the properties of the silica fibre. Optical attenuation in silica fibres is low in the region $1.0 - 1.6\,\mu\text{m}$ with a minimum at $1.55\,\mu\text{m}$. The wavelength of zero chromatic dispersion in such a fibre is normally $1.3\,\mu\text{m}$, but can be shifted to $1.55\,\mu\text{m}$ (or, alternatively, reduced to a low value over the range $1.3 - 1.55\,\mu\text{m}$) by careful design of the refractive index distribution to partially compensate material dispersion by waveguide dispersion. These constraints on wavelength resulted in the development of semiconductor lasers based on materials other than GaAs–AlGaAs, the most commonly used system being InGaAsP–InP. The choice of suitable materials is limited by the requirement of lattice-matching to achieve strain-free heterojunctions, as well as the energy gap of the active region corresponding to the emission wavelength required. The quaternary InGaAsP can be grown lattice-matched on InP substrates and thus lends itself well to the formation of double heterostructures. Another material system that offers potential for lasers is GaInNAs grown on GaAs substrates [10], but there has been as yet no commercial development of dilute nitride lasers. The use of limited amounts of strain to reduce the hole effective mass and hence allow reduced laser thresholds [11, 12] has become commonplace in today's QW lasers and SOAs (see Chap. 6).

7.2.2 Optical Waveguides

The use of the double heterostructure has a second benefit to laser operation in addition to the primary one of carrier confinement. The wider band gap confining layers have a lower refractive index than that of the active layer, so that the structure forms a planar dielectric waveguide which acts to confine the emitted radiation to the active layer. For example, for an emission wavelength of $1.55\,\mu\text{m}$, the refractive index of the InGaAsP active layer is about 3.57 and that of the InP commonly used as confining layers is about 3.17. The lowest order transverse mode of this waveguide has two possible polarisations, one with the electric field normal to the direction of propagation (transverse electric—TE) and the other with the magnetic field normal to the propagation direction (transverse magnetic—TM). Here, for simplicity, we consider only TE (which is normally the preferred polarisation in many lasers); more detailed descriptions of TE and TM modes can be found, for example, in [6]. For the purpose of analysis, it is convenient to group the active (confining) layer refractive index $N_1(N_2)$, the wavelength λ and the active layer thickness d into a single variable, the normalised frequency v, defined as

$$v = \frac{\pi d}{\lambda}\sqrt{N_1^2 - N_2^2} \tag{7.6}$$

As v decreases the number of modes that can propagate is reduced (the modes are "cut off"). For $v < \pi/2$, only the lowest order mode can propagate. This is the mode that is of interest for lasers. Using the numbers for the InGaAsP–InP laser

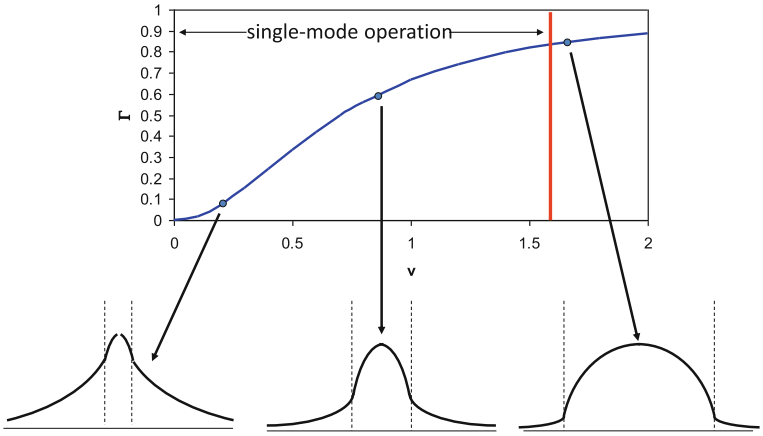


Fig. 7.4 Variation of optical confinement factor with normalised frequency v for a symmetric slab waveguide. Transverse intensity distributions are also shown for three values of v

given above results in single-mode operation for an active layer thickness d less than about $0.5\,\mu\text{m}$.

It is important to know the fraction of light intensity that is propagating in the core of the waveguide. This is the confinement factor, Γ, defined as the integral of the optical intensity over the active region divided by the corresponding integral over the total structure cross-section (essentially a dimension tending to infinity). A useful approximation that is accurate to about 1.5% is given by [13]

$$\Gamma = \frac{2v^2}{1 + 2v^2} \tag{7.7}$$

At the cut-off of the first higher order mode, this approximation yields $\Gamma = 0.83$. Figure 7.4 shows the variation of Γ with v; the transverse optical intensity distribution is also indicated schematically at three values of v. As v tends to zero the optical intensity spreads more and more into the cladding layers. With increasing v the intensity is more strongly confined to the waveguide core, which usually corresponds to the laser active layer. In the case of QW active layers, extra confining layers of refractive index intermediate between those of the QW and cladding indices are often included. Such a structure, often termed a 'separate confinement heterostructure' (SCH) gives better optical confinement than could be obtained with QWs alone, even in the case where multiple quantum wells (MQWs) separated by thin barrier layers, are used for the active material.

The heterostructure gives optical and electrical confinement in the *transverse* direction. In the *lateral* direction other structures are needed for confinement. Broad-area lasers suffer from high threshold current and filamentary behaviour (thin longitudinal regions of lasing determined by inhomogeneities). In the early years of laser development, lateral confinement was first achieved with the stripe-geometry

laser [14] which had a narrow ($5 - 20\,\mu$m) stripe contact defined by oxide insulation. Under the stripe there was a region of gain and elsewhere there was loss. This is "gain guiding", sometimes also accompanied by a weak index-guiding (or anti-guiding) effect. Such lasers suffered from unstable behaviour with increasing current, and commonly exhibited departures from linearity ("kinks") in the light-current (L-I) characteristic. More modern lasers include lateral optical and electrical confinement by the use of structures such as the buried-heterostructure (BH) [15] and ridge-waveguide laser [16]. The technological aspects of these structures are not the primary concern here, but it is relevant to mention the simplest analysis that can be used to aid understanding of how the optical confinement is achieved. Figure 7.5 illustrates this approach for the case of the ridge-waveguide laser. This 2-D waveguide problem is difficult to solve exactly, but a very useful simplification can be made by separating the structure into 3 separate 1-D problems (2 vertical and 1 horizontal, as illustrated in Fig. 7.5). Structure A (a vertical 4-layer asymmetric waveguide) can be solved to give an "effective index" (scaled propagation constant) which is lower than that of structure B (a 4-layer waveguide in which one layer is of a different thickness than the one in A). These two effective indices can be used to form a 3-layer symmetric waveguide in the horizontal direction, as indicated in Fig. 7.5. The solution of this waveguide yields the confinement factor for the lateral direction which can then be multiplied by the corresponding confinement factor for the vertical direction to give a result for the complete structure. It is important to note that the total optical confinement factor is the only waveguide characteristic that is of interest for laser design and modelling.

7.3 Semiconductor Laser Cavities

7.3.1 Fabry-Perot Cavity

The simplest form of semiconductor laser structure employs a Fabry-Perot (FP) cavity formed by cleaved facets at each end of the device. The relatively high refractive index (more accurately, the effective index from the solution of the lateral and transverse waveguide structure) of the semiconductor gives sufficient power reflectivity, R (typically 30%), at the facets to produce a resonant cavity. In order to model the cavity, define a modal gain per unit length, g,

$$g = \Gamma g_m - \alpha_\ell \tag{7.8}$$

where α_ℓ is the loss per unit length (which may be decomposed into contributions from the active layer and the passive confining layers, if desired) and the other symbols are as defined previously. The modal gain governs the exponential growth of the optical intensity in the longitudinal direction of the FP cavity. Using this simple model and allowing for reflections at the facets, the amplification G (defined as the

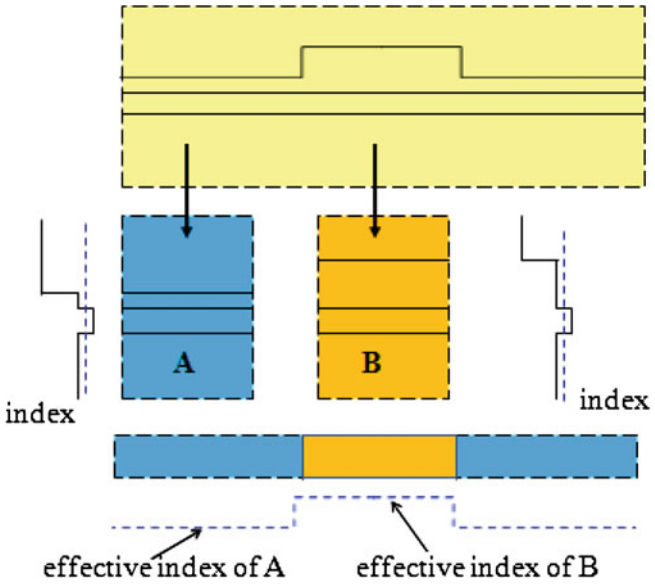


Fig. 7.5 The "effective index" method illustrated for a ridge-waveguide laser. The real structure is first separated into 3 separate 1-D structures (2 vertical and 1 horizontal). Structure A has an "effective index" (scaled propagation constant) lower than that of B. This results in a horizontal effective index waveguide

ratio of optical power output to input) can readily be derived in the form

$$G = \frac{(1-R)^2 e^{gL}}{(1-Re^{gL})^2 + 4Re^{gL}\sin^2\phi} \tag{7.9}$$

where L is the cavity length, R is the facet reflectivity (assumed the same at each end) and ϕ is the phase, given by $\phi = 2\pi NL/\lambda$ where N is the effective refractive index.

The FP resonances occur at values of the wavelength (λ_M) for which the argument of the sine function in (7.9) is a multiple of π, that is

$$\lambda_M = \frac{2NL}{M} \qquad (M = \text{an integer}) \tag{7.10}$$

Another way of stating the resonance condition (7.10) is that the effective optical cavity length (NL) must contain an integral number of half-wavelengths. Typical cavity lengths L are of the order of a few hundred microns and effective indices N are in the range 3.2–3.4, so that for a 1.55 μm laser it is clear that the integer M takes rather large values, typically of the order of 10^3. The spacing between the resonances, $\Delta\lambda$, can be calculated by differentiating (7.10). Noting that N is a function of wavelength, the result is

$$\Delta\lambda = \frac{\lambda^2}{2LN_g} \tag{7.11}$$

where N_g is the group refractive index (given by $N - \lambda dN/d\lambda$), which is typically in the range 3.7–4.0. For example, in a 300 μm-long laser, mode spacings of 0.3, 0.7 and 1.0 nm are found for the operating wavelengths of 0.85, 1.3 and 1.55 μm, respectively. Since the gain spectral bandwidth is usually many tens of nm (see Figs. 7.2 and 7.3), it is quite likely that more than one longitudinal mode will experience sufficient gain to satisfy the lasing threshold condition, and hence FP lasers often exhibit multimode output spectra.

7.3.2 Lasing Threshold and Power Output

The threshold for lasing is given from (7.9) by the condition that the denominator vanishes at the resonant wavelengths. Combining this result with (7.8) for g yields the threshold value for material gain, denoted g_{mth}, as

$$g_{\text{mth}} = \frac{1}{\Gamma}\left[\alpha_\ell + \frac{1}{L}\ell n\left(\frac{1}{R}\right)\right] \tag{7.12}$$

This equation states that the material gain at threshold is exactly balanced by the losses within the cavity as well as those through the end mirrors. Although this is a very good approximation, it neglects the small amount of spontaneous emission which couples into the lasing mode (this is the optical noise source which drives the oscillation). The threshold gain can be related to the carrier concentration at threshold, n_{th}, by (7.4) for bulk active media or (7.5) in the case of QW active layers. This carrier concentration, in turn, can be related to the threshold current density, j_{th}, by assuming that the carrier recombination (radiative and non-radiative) can be described by a lifetime, τ_e, so that the injected rate of electrons balances the recombination rate:

$$\frac{j_{\text{th}}}{ed} = \frac{n_{\text{th}}}{\tau_e} \tag{7.13}$$

where d is the thickness of the active material and e is the electron charge. Taking the case of a bulk active layer as an example and combining (7.4), (7.12) and (7.13) yields

$$j_{\text{th}} = \frac{ed}{\tau_e}\left\{n_o + \frac{1}{a\Gamma}\left[\alpha_\ell + \frac{1}{L}\ell n\left(\frac{1}{R}\right)\right]\right\} \tag{7.14}$$

Consider now a numerical example to see how this simple analysis can be used. For a laser emitting at 1.55 μm with an InGaAsP active layer ($N_1 = 3.57$) of thickness $d = 0.2\,\mu$m, and passive confining InP layers ($N_2 = 3.17$), the value of the normalised frequency is given from (7.6) as $v = 0.67$. The corresponding optical

confinement factor can be estimated from (7.7) as $\Gamma = 0.47$. For the lateral confinement, without going into the details of the effective index method as outlined above, let us assume a value of 0.85 for the confinement factor. Hence the total confinement factor is $0.85 \times 0.47 \approx 0.4$. Taking typical values of transparency concentration $n_o = 1 \times 10^{18}\,\mathrm{cm}^{-3}$, differential gain $a = 2.5 \times 10^{-16}\,\mathrm{cm}^2$, internal loss $\alpha_\ell = 30\,\mathrm{cm}^{-1}$, reflectivity $R = 0.3$, length $L = 300\,\mu\mathrm{m}$ and electron lifetime $\tau_e = 1\,\mathrm{ns}$, the equation gives the value of threshold current density as $j_{\mathrm{th}} \approx 5.4\,\mathrm{kA/cm}^2$. For an active width of $2\,\mu\mathrm{m}$, this would give a threshold current of about 33 mA.

Above the threshold, in the ideal case the lasing power output P is linearly related to the difference between the operating current, I, and the threshold current, I_{th},

$$P = \frac{I - I_{\mathrm{th}}}{e}\eta_D h\upsilon \tag{7.15}$$

where $h\nu$ is the photon energy and η_D is the differential quantum efficiency (or slope efficiency), defined as the fraction of photons escaping from the cavity. Hence η_D is given by the ratio of end-loss to total optical loss

$$\eta_D = \frac{\frac{1}{L}\ell n\left(\frac{1}{R}\right)}{\alpha_\ell + \frac{1}{L}\ell n\left(\frac{1}{R}\right)} \tag{7.16}$$

7.3.3 Distributed Bragg Reflectors

It has already been noted that FP lasers often exhibit multimode behaviour. For some applications, including optical communications, this is undesirable and a single-mode source is preferred. In order to select a single longitudinal mode (SLM), a grating, in the form of a distributed Bragg reflector (DBR), is often used in laser cavities. The simplest version is to replace one or both the laser facets by a DBR. The pitch of the grating, Ω, is related to the centre wavelength, λ_p, by

$$\Omega = p\frac{\lambda_p}{2N} \tag{7.17}$$

where p is an integer. For example, if $\lambda_p = 1.55\,\mu\mathrm{m}$ and $N = 3.3$, a first-order grating has a pitch of $0.23\,\mu\mathrm{m}$ and a second order grating has a pitch of $0.46\,\mu\mathrm{m}$. Higher orders than the second lead to strong coupling to radiation modes and are therefore rarely used, and then only for specific structures (grating-coupled lasers) to produce output normal to the grating for surface emission.

The properties of a DBR are determined by the coupling coefficient κ which measures the strength of coupling between forward and backward propagating waves in the grating. Figure 7.6 shows the grating reflection spectra for two values of the product κL_{DBR} where L_{DBR} is the length of the DBR. Larger values of this product

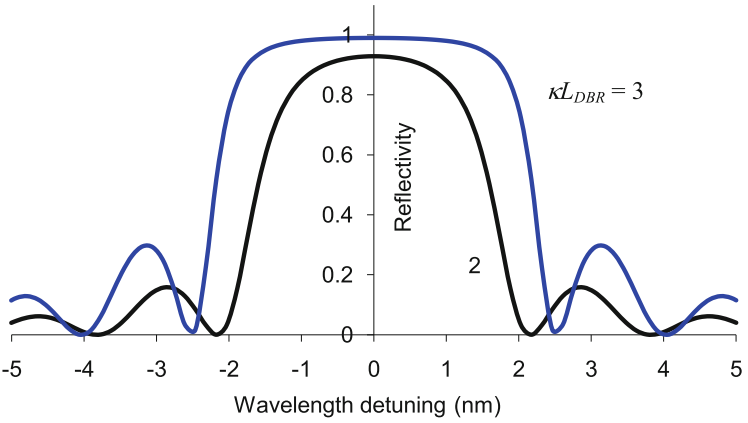


Fig. 7.6 Grating reflectivity spectra for two values of κL_{DBR}

give a peak reflectivity closer to unity and a more rectangular shape to the main lobe, together with higher side-lobes; smaller values of κL_{DBR} give lower reflectivities and a flatter response. In order to achieve SLM operation of a DBR laser, it is desirable to use a grating with narrow bandwidth and high reflectivity. The reflector at the other end of the cavity can be either a cleaved facet or a second DBR, this time with lower reflectivity for optimal output coupling.

DBR cavities are of particular interest for applications in tunable lasers. Injected current, temperature or an electric field produced by a reverse-biased junction can all be used to change the effective index N in the DBR, and hence, as is evident from (7.17), to change the centre wavelength λ_p. Injected current is the preferred mechanism, and the tuning range achievable in this way is usually limited to around 10 nm. In order to achieve a quasi-continuous tuning range (to overcome gaps at mode jumps), a phase section (with a current contact) is usually included to make a three-section tunable laser of the type shown schematically in Fig. 7.7. Wider tuning ranges can be obtained by using the four-section sampled-grating DBR (SG-DBR) laser which uses the Vernier effect for tuning. The SG-DBR reflectivity spectrum is a comb of wavelengths with the peaks spaced by an amount given by an equation similar to (7.11) where L is now interpreted as the grating sampling period. Using different sampling periods in the SG-DBRs at each end of a cavity results in two reflection combs with different wavelength spacings, as illustrated in Fig. 7.8. Lasing takes place at wavelengths where the reflection peaks align (the Vernier effect). Changing the current in either reflector causes the comb alignment to alter and hence tunes the wavelength. Tuning ranges of 50–70 nm can be achieved by this strategy. These structures also lend themselves well to monolithic integration with other optoelectronic components. More details of tunable lasers and associated devices can be found in [17].

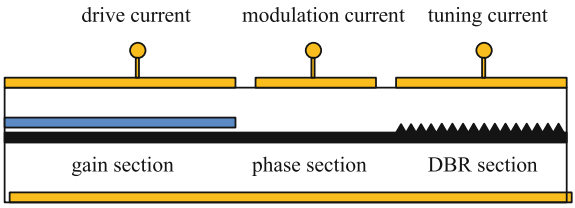
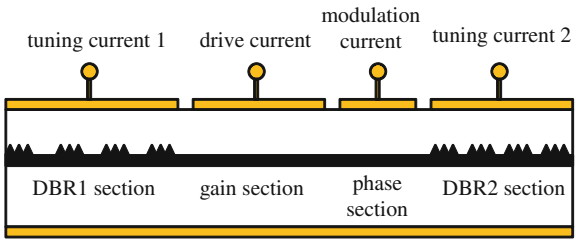


Fig. 7.7 Three-section tunable DBR laser (schematic)



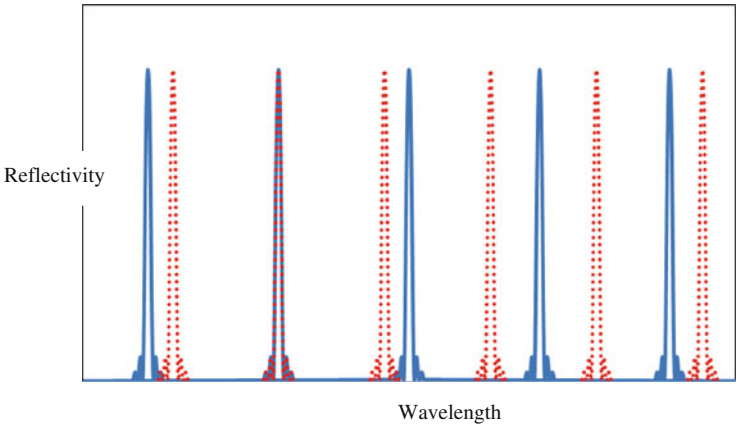


Fig. 7.8 Four-section widely-tunable laser: structure and reflectivity combs (schematic)

7.3.4 Distributed Feedback Lasers

The grating can also be incorporated along the length of the gain section, rather than as a separate reflector for the DBR laser. In this case the structure becomes a distributed feedback (DFB) laser where gain, g, and feedback (characterised by the coupling coefficient κ) occur in the same region of space, as illustrated in Fig. 7.9. In order to avoid non-radiative recombination due to defects introduced by the grating fabrication process, it is necessary to separate the corrugations from the active layer whilst still allowing them to interact strongly with the optical field. Usually this is achieved by the use of an SCH structure (see Sect. 7.2 above) where the grating

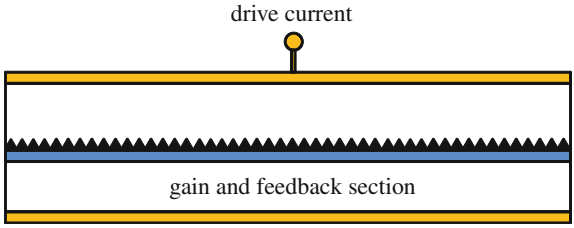


Fig. 7.9 Distributed feedback (DFB) laser (schematic)

is separated from the active layer by a layer of lower refractive index. Thus the corrugations interact with the evanescent tail of the optical distribution to provide feedback along the length of the cavity. Although the reflectivity for each corrugation is very small (typically about 0.03%), since there can be more than 500 of these along the length of the laser, the effective reflectivity resulting from the summation can be similar to that for an FP laser with cleaved mirrors. It is important in the DFB laser design to try to suppress the effects of reflection from cleaved facets, since these can adversely affect SLM performance. This is sometimes done by incorporating an absorbing region at one end of the device or, more commonly, by reducing the facet reflectivity through the use of anti-reflecting coatings or angled facets. More details of these approaches will be discussed below in relation to SOAs.

In general, the analysis of DFB lasers is somewhat complicated and there is no simple expression for lasing threshold. However, in the special case of strong coupling and low gain, $g \ll \kappa$, a useful simple approximate expression has been derived [18]

$$g_{\mathrm{mth}} = \frac{1}{\Gamma}\left[\alpha_\ell + \frac{1}{2L}\left(\frac{\pi}{\kappa L}\right)^2\right] \tag{7.18}$$

This equation for the DFB laser threshold has the same form as (7.12) for the FP laser threshold. In this case the interpretation is that the material gain at threshold is exactly balanced by the sum of the losses within the cavity as well as those due to coupling between the forward and backward propagating waves. The characteristic spectrum of an ideal DFB laser consists of two dominant longitudinal modes positioned symmetrically on either side of a stop band. The symmetry can be broken either by reflections from a facet, which is not easy to control since the phase of the reflection plays a strong role, or, more controllably, by the use of a quarter-wave shift in the grating. The latter is the preferred method to ensure SLM behaviour in DFB lasers. Within the constraints of the same approximation $g \ll \kappa$, the width of the stopband (in units of photon wavenumber) is approximately 2κ, and measurements of this width in DFB lasers below threshold are often used to obtain estimates of the grating coupling coefficient κ.

7.4 Transient Behaviour of Lasers

7.4.1 Static Properties

The main objective of this section is to analyse the temporal behaviour of semiconductor lasers using rate equations. However, as a first step towards this, let us first summarise the equations for cw laser operation. Below threshold, the stimulated photon density is zero and the electron concentration in the active region is linearly related to the current

$$n = \frac{I\tau_e}{eV} \tag{7.19}$$

where V is the volume of the active region. At threshold the photon density is still zero (neglecting the spontaneous emission) and the electron concentration reaches the value n_{th} which can be calculated from the threshold condition for material gain, g_{mth}, for example using (7.12) for an FP laser or (7.18) for a DFB laser. Above threshold the carrier concentration is clamped at n_{th} (if spatial non-uniformities are neglected), and the photon density S is given by

$$S = \frac{1}{g_{\text{mth}} v_g}\left(\frac{I}{eV} - \frac{n_{\text{th}}}{\tau_e}\right) \tag{7.20}$$

This equation is equivalent to (7.15), considering that here we are concerned with the photon density S inside the laser cavity, whereas (7.15) describes the power output P. It can easily be verified that the relation between these quantities for an FP laser is

$$P = SVh\upsilon \frac{v_g}{\Gamma L}\ell n\left(\frac{1}{R}\right) \tag{7.21}$$

where the final term on the RHS accounts for the loss rate of photons through both laser facets.

7.4.2 Rate Equations

The transient behaviour of semiconductor lasers can best be discussed in terms of rate equations describing the temporal evolution of electron concentration n and photon density S, each assumed uniform throughout the cavity. For a single-mode laser the simplest form of these equations is

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{j}{ed} - \frac{n}{\tau_e} - v_g a(n - n_o)S \tag{7.22}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = v_g \Gamma a(n - n_o)S - \frac{S}{\tau_p} \tag{7.23}$$

where j is the current density and τ_p is the photon lifetime in the cavity. For an FP laser the photon lifetime is related to the losses by

$$\frac{1}{\tau_p} = v_g \left[\alpha_\ell + \frac{1}{L} \ell n \left(\frac{1}{R} \right) \right] \tag{7.24}$$

The first term on the RHS of (7.22) represents the rate of pumping into the active region, the second term is the total radiative and non-radiative recombination rate and the third term is the stimulated emission rate. In (7.23) the two terms on the RHS represent, respectively, the stimulated emission rate and the photon loss rate. It is easily verified that in the steady state these equations can be solved to give (7.14) at threshold, and (7.19) and (7.20) below and above threshold, respectively, noting that current I is related to current density J by $Id = jV$.

7.4.3 Small-Signal Modulation

Some useful physical insights into the time-dependent behaviour of lasers can be found by considering the case of small-signal modulation of the current and using the rate equations to analyse the response. Using the subscript 's' to denote steady-state values of variables, the current density can be written as

$$j = j_\mathrm{s} + \Delta j e^{i\omega t} \tag{7.25}$$

where ω is the angular modulation frequency and the small-signal assumption means that $\Delta j \ll j_\mathrm{s}$. As already discussed the steady-state solutions of (7.22) and (7.23) are

$$n_\mathrm{s} = n_\mathrm{th} = n_o + \frac{1}{\Gamma a v_g \tau_p} \tag{7.26}$$

$$S_\mathrm{s} = \frac{\tau_p (j - j_\mathrm{th})}{ed} \tag{7.27}$$

The small-signal solutions for carrier and photon densities can be written as

$$n = n_\mathrm{s} + \Delta n e^{i\omega t} \tag{7.28}$$

$$S = S_\mathrm{s} + \Delta S e^{i\omega t} \tag{7.29}$$

where $\Delta n << n_s$, $\Delta S << S_s$. Substituting these solutions into the rate (7.22) and (7.23), and neglecting terms in $\Delta n \Delta S$ (in the small-signal approximation) yields

$$i\omega\Delta n = \frac{\Delta j}{ed} - \frac{\Delta n}{\tau_e} - v_g a\left[S_s\Delta n + (n_s - n_o)\Delta S\right] \tag{7.30}$$

$$i\omega\Delta S = v_g \Gamma a S_s \Delta n \tag{7.31}$$

These equations can be solved for the ratio $\Delta S/\Delta j$ in the form

$$\frac{\Delta S}{\Delta j} = \left(\frac{\Gamma\tau_p}{ed}\right)\frac{\omega_o^2}{\omega_o^2 + i\omega\omega_d - \omega^2} \tag{7.32}$$

where ω_o is the angular relaxation oscillation frequency (ROF)

$$\omega_o^2 = \frac{v_g a}{\tau_p} S_s \tag{7.33}$$

and ω_d is the damping frequency

$$\omega_d = \frac{1}{\tau_e} + \tau_p \omega_o^2 \tag{7.34}$$

The frequency dependence of the amplitude of the transfer function is given from (7.32) as

$$\left|\frac{\Delta S(\omega)}{\Delta S(0)}\right| = \frac{\omega_o^2}{\sqrt{(\omega_o^2 - \omega^2)^2 + \omega^2\omega_d^2}} \tag{7.35}$$

Figure 7.10 gives a plot of the variation of this function with modulation frequency $f = \omega/(2\pi)$. The maximum of the response occurs at the resonance frequency f_r, given by

$$f_r = \frac{1}{2\pi}\sqrt{\omega_o^2 - \frac{\omega_d^2}{2}} \tag{7.36}$$

The bandwidth or 3-dB frequency, $f_{3\text{dB}} = \omega_{3\text{dB}}/(2\pi)$, is given by the solution of

$$\left(\omega_o^2 - \omega_{3\text{dB}}^2\right)^2 + \omega_{3\text{dB}}^2\omega_d^2 = 4\omega_o^2 \tag{7.37}$$

Since the linear ROF, $f_o = \omega_o/(2\pi)$, is typically of the order of a few GHz, and τ_p, τ_e are of the order of 1 ps and 1 ns, respectively, it is often a good approximation to assume that ω_d from (7.34) is much less than ω_o. With this assumption an approximate solution of (7.37) is

$$f_{3\text{dB}} \cong \sqrt{3} f_o \cong \sqrt{3} f_r \tag{7.38}$$

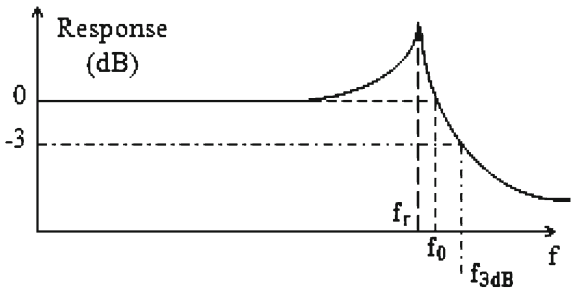


Fig. 7.10 Amplitude of the modulation transfer function for small-signal modulation

It is worth investigating the dependence of these frequencies on the current overdrive above threshold, $j - j_{\text{th}}$. Substituting (7.22) for S_s into equations (7.33) and (7.38), respectively, yields

$$f_o = \frac{1}{2\pi}\sqrt{\frac{v_g a(j - j_{\text{th}})}{ed}} \tag{7.39}$$

$$f_{3\text{dB}} \cong \frac{1}{2\pi}\sqrt{\frac{3v_g a(j - j_{\text{th}})}{ed}} \tag{7.40}$$

These results that both the ROF and the small-signal modulation bandwidth vary as the square root of the overdrive current have been tested experimentally many times in the literature and generally describe the measured behaviour very well.

7.4.4 Large-Signal Modulation

A limitation of the small-signal analysis is that it cannot deal with large-signal digital modulation (which is of fundamental importance when the laser is used as a source for optical communications). In this case, numerical solutions of the rate equations must be used to investigate the response. Some elementary insight can be obtained by considering the response of the laser to a step change in current. For example, if the laser is initially biased below threshold with current density j_b, and a current density j is applied at time $t = 0$, what is the optical response? In fact there will be a short but non-negligible delay between the leading edge of the current pulse and the resulting optical output. To see this it is only necessary to consider the carrier rate (7.22) below threshold

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{j}{ed} - \frac{n}{\tau_e} \tag{7.41}$$

The solution at time t is

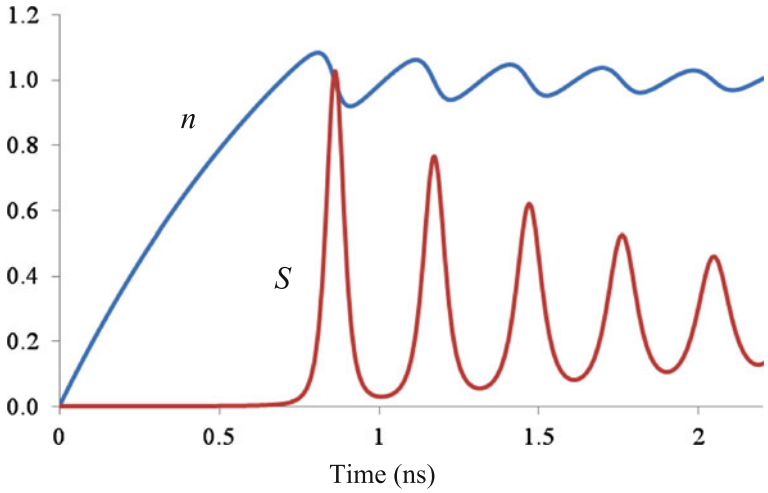


Fig. 7.11 Relaxation oscillations of electron and photon concentrations in response to a step current pulse

$$n = \frac{\tau_e}{ed}\left[j - (j - j_b)\exp\left(-\frac{t}{\tau_e}\right)\right] \tag{7.42}$$

The turn-on delay of the optical pulse, t_d, is defined as the time for the carrier density to reach its threshold value as given by (7.26). Solving (7.42) thus yields [19]

$$t_d = \tau_e \ell n\left(\frac{j - j_b}{j - j_{\text{th}}}\right) \tag{7.43}$$

It follows that the turn-on delay decreases as the bias current increases; when j_b approaches j_{th} the delay tends towards zero. Therefore, in order to modulate at high data rates, lasers are biased close to threshold to reduce turn-on delay. It is worth noting also that measured values of t_d as a function of current density can be used to gain an estimate of the recombination lifetime τ_e by using a logarithmic plot according to (7.43).

The laser output following the turn-on delay consists of a series of spikes whose amplitude decays with time; this is accompanied by the appearance of a damped saw-tooth oscillation in the electron concentration. This behaviour is illustrated in Fig. 7.11 where the temporal evolution of the photon and electron concentrations is simulated numerically. At first the carrier concentration increases rapidly until it exceeds the threshold value and a pulse of light is emitted. This, in turn, has the effect of reducing the carrier density so that the emission is extinguished. The process is repeated, with some damping, until the steady state is approached. The frequency and damping of these relaxation oscillations are given by (7.33) and (7.34) above.

7.4.5 Chirp

The effective refractive index in the active layer of the laser is a function of the carrier density. This leads to changes in the lasing wavelength ('chirp') during the relaxation oscillations or small-signal modulation. The effect is strongest during the initial turn-on period when the carrier density overshoots its steady-state value. To study laser chirp in more detail, consider a small change Δn in carrier density from the steady-state value n_s, and a corresponding small change $\Delta\lambda$ in wavelength from the steady-state value λ_s. There will be an associated change, ΔN, of the effective index N from its steady-state value N_s, given by

$$\Delta N = \frac{\partial N}{\partial \lambda}\Delta\lambda + \frac{\partial N}{\partial n}\Delta n \tag{7.44}$$

For a FP laser the resonant wavelength (λ_M) for mode M is given by (7.10). Using (7.10) and (7.44), it follows that

$$\frac{N}{\lambda_M}\Delta\lambda = \frac{\partial N}{\partial \lambda}\Delta\lambda + \frac{\partial N}{\partial n}\Delta n \tag{7.45}$$

which can be rewritten as

$$\Delta\lambda = \frac{\lambda_M}{N_g}\frac{\partial N}{\partial n}\Delta n \tag{7.46}$$

where N_g is the group index. This result is conveniently re-expressed in terms of another parameter, α, which relates the change in effective refractive index to that in gain, and is defined as

$$\alpha = -\frac{4\pi}{\lambda}\frac{\partial N/\partial n}{\partial g/\partial n} \tag{7.47}$$

This parameter is usually called the 'linewidth enhancement factor', in view of its importance in characterising single-mode laser linewidth. It was first introduced by C.H. Henry in a study of laser linewidth [20] and is therefore also known as 'Henry's α factor'. Using (7.47), the result for wavelength chirp from (7.46) becomes

$$\Delta\lambda = -\frac{\lambda^2}{4\pi N_g}\alpha\frac{\partial g}{\partial n}\Delta n \tag{7.48}$$

where the wavelength subscript 'M' has been omitted for simplicity. Note that for the linear gain model of (7.4), the differential gain $\partial g/\partial$n is given by Γa. With this model we can also use the small-signal modulation result (7.31) to re-express the chirp in terms of the change in photon density

$$\Delta\lambda = -\frac{\lambda^2}{4\pi N_g}\alpha\frac{i\omega}{v_g}\frac{\Delta S}{S_s} \tag{7.49}$$

The chirp can also be expressed as a change in frequency $\Delta\nu$, as

$$\Delta\nu = \alpha \frac{i\omega}{4\pi} \frac{\Delta S}{S_s} \tag{7.50}$$

It follows from this result that the chirp, or frequency modulation (FM) is simply proportional to the intensity modulation (IM), and this proportionality is governed by the modulation frequency ω and the linewidth enhancement factor. The FM index is defined as $|2\pi\Delta\nu/\omega|$, and the amplitude modulation (AM) index is defined as $|\Delta S/S_s|$. Thus the ratio F/A is given from (7.50) as

$$\frac{F}{A} = \frac{|\alpha|}{2} \tag{7.51}$$

Measurements of F/A as a function of modulation frequency can be used to find values of α [21]. Clearly, large values of α are desirable for FM applications, whereas low values are required to minimise chirp in IM transmission. Measured values of $|\alpha|$ in long-wavelength lasers are in the range 3–10, and exhibit a monotonic increase with wavelength towards the band-edge wavelength. Hence it is possible to design a DFB or DBR laser with a grating whose centre wavelength is used to select the desired value of $|\alpha|$.

7.5 Semiconductor Optical Amplifiers

The basic SOA structure consists of a semiconductor laser operated below threshold and arranged so that an input signal can be coupled into one end of the cavity and the transmitted signal coupled out of the other end; sometimes the device is operated in reflection and thus optical coupling is only required for the input end. The cavity structure can be of any type, e.g. FP, DBR or DFB, and the cavity will selectively amplify wavelengths corresponding to the resonant modes. A response that is more tolerant to input wavelength is usually desired, and this can be achieved by reducing the facet reflectivities of an FP cavity to sufficiently low values. In this case we speak of a near-travelling wave (NTW) amplifiers when the reflectivities are sufficiently low that the gain ripple due to residual FP modes is below a specified value, as compared with the ideal travelling-wave (TW) amplifier where the reflectivities would be zero.

7.5.1 Cavity Effects

In SOAs, as in lasers, the word "gain" can refer to more than one quantity. Hence it is important to distinguish between the material gain per unit length g_m (as given, for example, by (7.4) and (7.5) which are useful approximations), the modal gain per

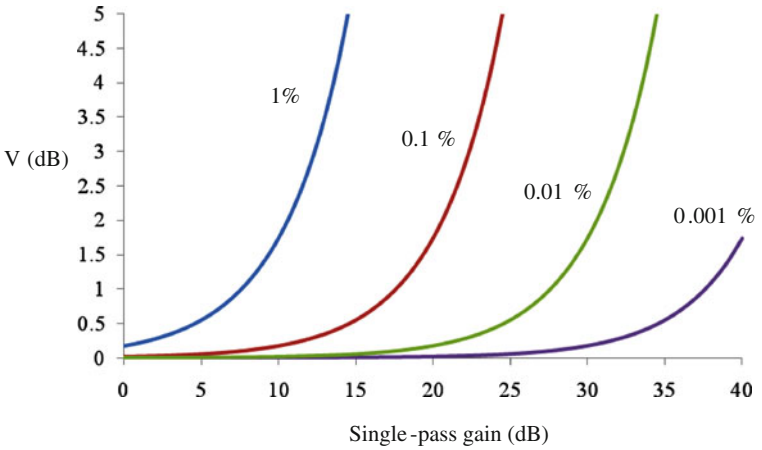


Fig. 7.12 Gain ripple for an FP cavity versus single-pass gain for reflectivities indicated

unit length g (as defined in (7.8) in terms of confinement factor and internal losses), and the transmission gain G as defined by (7.9) for FP and NTW amplifiers. There is also the gain in reflection, G_b, defined, again for the case of equal reflectivities at each end of the cavity, as

$$G_b = \frac{R\left(1 - e^{gL}\right)^2 + 4Re^{gL}\sin^2\phi}{\left(1 - Re^{gL}\right)^2 + 4Re^{gL}\sin^2\phi} \tag{7.52}$$

The peak-to-trough ratio V of the transmission gain is given from (7.9) as

$$V = \left(\frac{1 + Re^{gL}}{1 - Re^{gL}}\right)^2 \tag{7.53}$$

A plot of this quantity (in dB) versus single-pass gain (e^{gL}, also in dB) is given in Fig. 7.12, for a number of different values of reflectivity. It is clear that to obtain suitable low values of V with high gain, low values of reflectivity are required. For example, for 30 dB gain with less than 1 dB ripple, the reflectivity must be less than 0.01%.

Means of reducing the facet reflectivities include single-layer (quarter-wave) coatings using a material whose refractive index would ideally be equal to the square root of the effective index of the wave in the SOA (silicon monoxide is sometimes used as a coating). However, this gives a very narrowband response and the minimum occurs at different wavelengths for the two polarisations (TE and TM). Multilayer coatings can offer a broader wavelength range and give low gain ripple. Other methods of reducing reflectivities are the use of angled facets and 'window' or buried facets where the active region is terminated some distance away from the facet. For these

methods only a small fraction of the reflected light from the facet is captured into the active region of the SOA. Combinations of the latter techniques can give extremely low reflectivities and low gain ripple (e.g. less than 0.5 dB at a gain of 25 dB [22]) even without the use of anti-reflection (AR) coatings.

The resonances in an SOA are characterised by the full-width at half-maximum (FWHM) in frequency, denoted Δf. Writing the phase in terms of frequency f as $\phi = 2\pi NLf/c$ (where c is the speed of light), the FWHM for transmission is found from (7.9) to be

$$\Delta f = \frac{c}{N\pi L} \sin^{-1}\left[\frac{1 - Re^{gL}}{2(Re^{gL})^{1/2}}\right] \tag{7.54}$$

For example, for an uncoated SOA ($R = 0.3$) of length 200 μm with gain of 20 dB, the bandwidth is about 9 GHz. It is clear that the bandwidth decreases as the peak gain increases. A figure of merit for this is the gain-bandwidth product, defined as the product of the square root of the peak gain, $G_{\max}$, and the FWHM, Δf. Using the threshold condition and the approximation $\sin(x) = x$ (for small x) in (7.9) and (7.54) leads to the simple result [23]

$$\sqrt{G_{\max}}\Delta f = \frac{c}{2N\pi L}\left(\frac{1}{\sqrt{R}} - \sqrt{R}\right) \tag{7.55}$$

This is a particularly useful result in studying the design and analysis of SOAs, and especially for vertical cavity devices (VCSOAs)—see Chap. 8.

7.5.2 Saturation

SOAs offer high values of transmission gain, but at higher input powers the gain saturates as the available energy from the electrical current is used. To describe gain saturation in a simple manner, the rate equation (7.22) can be used in steady state, rewritten in the form

$$\frac{j}{ed} = \frac{n}{\tau_e} + \frac{(n - n_o)}{\tau_e}\frac{P_{\text{int}}}{P_{\text{sat}}} \tag{7.56}$$

where the photon density S has been expressed in terms of the internal power P_{int} in the SOA

$$P_{\text{int}} = SVh\upsilon\frac{v_g}{\Gamma L} \tag{7.57}$$

Note that (7.57) is similar to (7.21), except that here we deal with the internal photon density and hence the final term of (7.21) is omitted. The quantity P_{sat} is the saturation power defined as

$$P_{\text{sat}} = \frac{Vh\upsilon}{\tau_e a\Gamma L} \tag{7.58}$$

The solution of (7.56) gives a simple expression for the saturation of the material gain

$$g_m = \frac{g_{\mathrm{mo}}}{1 + \frac{P_{\mathrm{int}}}{P_{\mathrm{sat}}}} \tag{7.59}$$

where g_{mo} is the unsaturated material gain in the absence of an optical signal, given by

$$g_{\mathrm{mo}} = a\left(\frac{j\tau_e}{ed} - n_o\right) \tag{7.60}$$

For a TW amplifier we must allow for the spatial variation of power and saturation along the length of the device. A particularly simple case can be used for illustration by taking $\Gamma = 1$ and $\alpha_\ell = 0$, so that $g = g_m$. Then the equation for the growth of power P s a function of position z along the cavity becomes

$$\frac{\mathrm{d}P_{\mathrm{int}}}{\mathrm{d}z} = g\,P_{\mathrm{int}} = \frac{g_{\mathrm{mo}}\,P_{\mathrm{int}}}{1 + \frac{P_{\mathrm{int}}}{P_{\mathrm{sat}}}} \tag{7.61}$$

Defining the input power as P_{in} and and the single-pass unsaturated gain as $G_o = \exp(g_{\mathrm{mo}}L)$, (7.61) can be integrated from $z = 0$ (input) to $z = L$ (output), with the result

$$\frac{P_{\mathrm{in}}}{P_{\mathrm{sat}}} = \frac{\ell n\left(\frac{G_o}{G}\right)}{G - 1} \tag{7.62}$$

The solution of this equation is plotted in Fig. 7.13, assuming an unsaturated gainG_o of 20 dB. The plot shows how the saturated gain G varies with the ratio of input powerP_{in} to saturation power P_{sat}. From (7.58), taking typical values of $V/L = 0.2\,\mu\mathrm{m}^2$, $h\nu = 0.8\,\mathrm{eV}$, $a = 2.5 \times 10^{-16}\,\mathrm{cm}^2$, $\Gamma = 0.3$, and $\tau_e = 1\,\mathrm{ns}$, we find $P_{\mathrm{sat}} = 3\,\mathrm{mW}$. From Fig. 7.13, it is seen that, for this example, when P_{in} is 3 mW, the amplifier gain has reduced to about 4 dB.

7.5.3 Crosstalk

For some applications, as in the case of a wavelength division multiplexed (WDM) system, an SOA will be subject to a number of different input signals at different wavelengths. Crosstalk arises between these signals as a result of their interaction with the saturable gain as discussed above. To see the implications of this for the speed of response, consider the time-dependent rate equation

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{j}{ed} - \frac{n}{\tau_e} - \frac{(n - n_o)}{\tau_e}\frac{P_{\mathrm{int}}}{P_{\mathrm{sat}}} \tag{7.63}$$

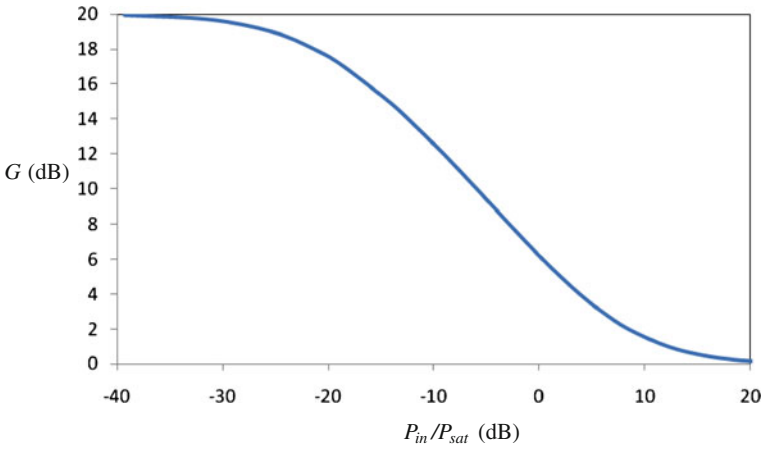


Fig. 7.13 Gain saturation with input power for an SOA, calculated from (7.62)

and assume an optical signal consisting of a constant term and a small-signal sinusoidal term

$$P_{\text{int}} = P_{\text{int s}} + \Delta P_{\text{int}} e^{i\omega t} \tag{7.64}$$

The corresponding variation of carrier concentration will follow a similar form

$$n = n_{\text{s}} + \Delta n e^{i\omega t} \tag{7.65}$$

and we are interested in finding the dependence of the small-signal amplitude Δn on the angular frequency ω. Performing a small-signal analysis of (7.63) by neglecting terms higher than first order in ΔP_{int} and Δn yields

$$\left| \frac{\Delta n P_{\text{sat}} \tau_e}{\Delta P_{\text{int}}} \right|^2 = \frac{(n_{\text{s}} - n_o)^2}{\omega^2 + T^{-2}} \tag{7.66}$$

where the time constant T is given by

$$\frac{1}{T} = \frac{1}{\tau_e} \left(1 + \frac{P_{\text{int s}}}{P_{\text{sat}}} \right) \tag{7.67}$$

It follows from (7.66) that the effects of crosstalk can be minimised by using amplifiers with large values of the saturation power; QW devices have shown the best performance in this respect. It is also clear from the form of (7.66) that the electron concentration, and hence the material gain, will faithfully follow the temporal variation of the optical power for angular frequencies below that corresponding to T^{-1}. However, for frequencies much above this value, the variation of the optical

signal is too fast for the gain to follow accurately, and the gain remains close to its steady-state mean value corresponding to an electron concentration n_s.

Whilst the effect of crosstalk between channels is very undesirable for WDM transmission systems, the underlying effect of gain saturation can also be used as the basis of wavelength conversion. Wavelength conversion is of interest for wavelength re-use and for simplifying switching nodes in WDM networks. Consider an amplifier with two input signals, one of which is at wavelength λ_1 and carries data, whilst the second is of constant power at wavelength λ_2. Now, provided the data rate is not too fast, as determined by the time constant T, then the modulation will be transferred to the signal at λ_2. Note also that, from equation (7.67), the time constant can be decreased, and hence higher data rates accommodated, by increasing the mean power $P_{\text{int s}}$ in the modulated signal. This is true for small-signal modulation assuming uniform optical power throughout the amplifier. If more realistic assumptions are made then the situation becomes more complicated, although the basic effect of speeding up the response by the use of a more intense signal still holds.

In order to model the transient response of wavelength conversion, (7.63) must be modified to allow for two signals (modulated and cw) and for the wavelength dependence of the differential gain and the transparency concentration. The resulting equation must be solved in parallel with a travelling wave equation similar to (7.61) in order to account for the spatial and temporal evolution of the optical signals. Solutions have been reported for models which split the amplifier cavity into longitudinal sections and by analytic small-signal solutions. These theoretical results show that TW effects are extremely important in determining the maximum bandwidth for wavelength conversion in SOAs. For the co-propagating case, the longitudinal dependence of gain can lead to considerable enhancement of the bandwidth. The low-frequency components are transferred from the modulated signal to the cw beam in the front portion of the amplifier (where the carrier lifetime is long due to the relatively low signal level), whilst the high-frequency components are transferred closer to the rear of the device (where the carrier lifetime is shorter due to the strong stimulated emission rate).

7.5.4 Polarisation

For applications in optical communications systems, SOAs are required to give the same gain independent of the polarisation of the input signals. However, this is difficult to achieve in practice since several physical effects in SOAs are polarisation dependent. In particular, the wave propagation constant and optical confinement factor are different for TE and TM polarisation. The cross-section of the active region is asymmetric (the width is typically an order of magnitude bigger than the thickness) and this leads to polarisation-dependent gain (PDG). To avoid this it would be necessary to aim for active regions of square cross-section, using etching and regrowth techniques. However, the fabrication for this is difficult and can lead to a reduction in saturation power, since a thick active layer gives a larger optical

confinement factor Γ, and this appears in the denominator of the expression for saturation power, (7.58). One SOA structure with near-square cross-section has been reported [24] with PDG less than 0.2 dB.

In QW active regions the selection rules for transitions between the conduction and valence bands add a further polarisation dependence to the optical gain. In this case strained layer epitaxy [11, 12] can be used to reduce or eliminate the polarisation sensitivity. If a material is grown that has a lattice constant less than that of the substrate, the result will be tensile strain. In the opposite case, compressive strain can be achieved. For example, the use of tensile strain in the barriers between QWs can be used to equalise TE and TM gain. More details can be found in Chap. 6 of this book. MQW SOAs have been reported with PDG less than 0.5 dB in the wavelength window around 1550 nm [25], and with PDG less than 0.6 dB over the entire 3-dB bandwidth of 56 nm around 1300 nm [26] (using tensile strain in the barriers combined with compressive strain in the QWs).

7.6 Conclusion

Whilst this chapter has surveyed elements of the basic theory of semiconductor lasers and SOAs, it should be emphasised that this is just the "tip of the iceberg" and that many of the topics covered here have been explored in much greater depth in other specialised publications. Nevertheless, it is hoped that the present minimalist treatment will suffice to give some insight into the design issues and measured behaviour of these devices. This concluding section is intended to highlight some of the more important theoretical expressions presented above.

The relations between material gain per unit length and electron concentration, for example (7.4) and (7.5), are fundamental to all aspects of lasers and SOAs. The corresponding definition of modal gain (7.8) in terms of optical confinement factor, as given by the approximation (7.7), is used to take account of waveguide effects and optical losses. The expression (7.9) for FP cavity gain is used to derive the resonant wavelengths (7.10) and lasing threshold condition (7.12), as well as SOA gain ripple (7.53) and optical bandwidth (7.54). The rate equations for electrons and photons (7.22) and (7.23) determine the transient behaviour of the laser, including the small-signal modulation response (7.35), the ROF (7.33), damping frequency (7.34) and turn-on delay (7.43). The linewidth enhancement factor (7.47) is used to describe the lasing wavelength chirp in terms of changes in carrier density (7.48) or photon density (7.49). The rate equation for carrier density is also used in the discussion of SOA saturation, leading to (7.59) for the saturation of material gain and (7.66) for the crosstalk between signals. However, more accurate treatments of SOAs must use travelling wave equations to describe the spatial dependence of photon distributions within the cavity, and (7.61) offers a simple example.

The relatively small number of key expressions noted here are all that is needed to furnish the simple theoretical description of device behaviour that has been adequate for the first half-century of semiconductor laser and SOA history. However,

as technology develops and more complex physics is utilised in device behaviour, more sophisticated theoretical tools become necessary. This, coupled with the widespread use of numerical simulations of lasers and SOAs, is leading to a much richer and more challenging theoretical landscape for future developments.

References

1. J. Faist, F. Capasso, D.L. Sivco, C. Sartori, A.L. Hutchinson, A.Y. Cho, Quantum cascade laser. Science **264**, 553–556 (1994)
2. J. Liu, X. Sun, R. Camacho-Aguilera, L.C. Kimerling, J. Michel, Ge-on-Si laser operating at room temperature. Opt. Lett. **35**, 679–681 (2010)
3. M.G.A. Bernard, G. Duraffourg, Laser conditions in semiconductors. Phys. Status Solidi **1**, 699–703 (1961)
4. P.W.A. McIlroy, A. Kurobe, Y. Uematsu, Analysis and application of theoretical gain curves to the design of multi-quantum-well lasers. IEEE J Quantum Electron. **21**, 1958–1963 (1985)
5. L.A. Coldren, S.W. Corzine, *Diode Lasers and Photonic Integrated Circuits* (Wiley, New York, 1995)
6. S.L. Chuang, *Physics of Optoelectronic Devices* (Wiley, New York, 1995)
7. W.W. Chow, S.W. Koch, *Semiconductor-Laser Fundamentals: Physics of the Gain Materials* (Springer, Berlin, 1999)
8. Z.I. Alferov, V.M. Andreev, D.Z. Garbuzov, Y.V. Zhilyaev, E.P. Morozov, E.L. Portnoi, V.G. Trofim, Investigation of the influence of the AlGaAs-GaAs heterostructure parameters on the laser threshold current and the realization of continuous emission at the room temperature. Fiz. Tekh. Poluprovodn. **4**, 1826–1829 (1970) (Sov. Phys. Semicond. **4**, 1573–1575 (1971))
9. I. Hayashi, M.B. Panish, P.W. Foy, S. Sumski, Junction lasers which operate continuously at room temperature. Appl. Phys. Lett. **17**, 109–111 (1970)
10. M. Kondow, K. Uomi, A. Niwa, T. Kitatani, S. Watahiki, Y. Yazawa, GaInNAs: a novel material for long-wavelength-range laser diodes with excellent high-temperature performance. Jpn. J. Appl. Phys. **35**, 1273–1275 (1996)
11. A.R. Adams, Band structure engineering for low-threshold high-efficiency semiconductor lasers. Electron. Lett. **22**, 249–250 (1986)
12. E. Yablonovitch, E.O. Kane, Reduction of lasing threshold current density by the lowering of valence band effective mass. J. Lightw. Technol. **LT-4**, 504–506 (1986)
13. D. Botez, Analytical approximation of radiation confinement factor for TE0 mode of a double heterojunction laser. IEEE J. Quantum Electron. **14**, 230–232 (1978)
14. J.E. Ripper, J.C. Dyment, L.A. D'Asaro, T.L. Paoli, Stripe-geometry double heterostructure junction lasers: mode structure and cw operation above room temperature. Appl. Phys. Lett. **18**, 155–157 (1971)
15. T. Tsukada, GaAs-Ga1-xAlxAs buried-heterostructure injection lasers. J. Appl. Phys. **45**, 4899–4906 (1974)
16. I.P. Kaminow, R.E. Nahory, M.A. Pollack, L.W. Stulz, J.C. DeWinter, Single-mode c.w. ridge-waveguide laser emitting at 1.55 μm. Electron. Lett. **15**, 763–765 (1979)
17. J. Buus, M.-C. Amann, D.J. Blumenthal, *Tunable Laser Diodes and Related Optical Sources*, 2nd edn. (Wiley, Hoboken, 2005)
18. H. Kogelnik, C.V. Shank, Coupled-wave theory of distributed feedback lasers. J. Appl. Phys. **43**, 2327–2335 (1972)
19. K. Konnerth, C. Lanza, Delay between current pulse and light emission of a gallium arsenide laser. Appl. Phys. Lett. **4**, 120–121 (1964)
20. C.H. Henry, Theory of the linewidth of semiconductor lasers. IEEE J. Quantum Electron. **18**, 259–264 (1978)

21. C. Harder, K. Vahala, A. Yariv, Measurement of the linewidth enhancement factor α of semiconductor lasers. Appl. Phys. Lett. **42**, 328–330 (1983)
22. A.E. Kelly, I.F. Lealman, L.J. Rivers, S.D. Perrin, M. Silver, Polarisation insensitive, 25 dB gain semiconductor laser amplifier without antireflection coatings. Electron. Lett. **32**, 1835–1836 (1996)
23. J. Piprek, S. Bjorlin, J.E. Bowers, Design and analysis of vertical-cavity semiconductor optical amplifiers. IEEE J. Quantum Electron. **37**, 127–134 (2001)
24. T. Ito, N. Yoshimoto, K. Magari, K. Kishi, Y. Kondo, Extremely low power consumption semiconductor optical amplifier gate for WDM applications. Electron. Lett. **33**, 1791–1792 (1997)
25. H. Ma, S.H. Chen, X.J. Yi, G.X. Gu, 1.55 μm spot-size converter integrated polarization-insensitive quantum-well semiconductor optical amplifier with tensile-strained barriers. Semicond. Sci. Technol. **19**, 846–850 (2004)
26. J.Y. Jin, D.C. Tian, J. Shi, T.N. Li, Fabrication and complete characterization of polarization insensitive 1310 nm InGaAsP-InP quantum-well semiconductor optical amplifiers. Semicond. Sci. Technol. **19**, 120–126 (2004)

Chapter 8
Vertical Cavities and Micro-Ring Resonators

Dimitris Alexandropoulos, Jacob Scheuer and Mike J. Adams

Abstract The scope of this chapter is to present the concepts of vertical cavities (VCs) and μ-ring resonators (MRs). The chapter commences with the motivation for progressing beyond conventional edge-emitting cavities emphasising on the potential of VC and MRs. The fundamental physics of VC and MRs is then analysed focusing on device design aspects. VCs are studied for optical amplifier applications. Lasing VCs are analysed in terms of polarisation dynamics. MRs in single and multi-ring configurations, like coupled resonator optical waveguides (CROWs) and side-coupled integrated spaced sequence of resonators, (SCISSORs) are discussed. Active MRs for lasers and amplifiers are investigated.

8.1 Introduction

A functional photonic device can be analysed in its three-structural elements: (1) the photonic material (2) the waveguide structure and (3) the resonator as shown in Fig. 8.1. Although a semiconductor photonic device is a complex physical system it is instructive to attempt a decoupling of the various effects so that we identify the main

D. Alexandropoulos (✉)
Department of Materials Science,
University of Patras, 26504 Patras, Greece
e-mail: dalexa@upatras.gr

J. Scheuer
Department of Physical Electronics,
School of Electrical Engineering,
Office: 232 Tel-Aviv University,
Ramat-Aviv, 69978 Tel-Aviv, Israel

M. J. Adams
School of Computer Science and Electronic Engineering,
University of Essex, Wivenhoe Park,
Colchester CO4 3SQ, UK

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8_8,

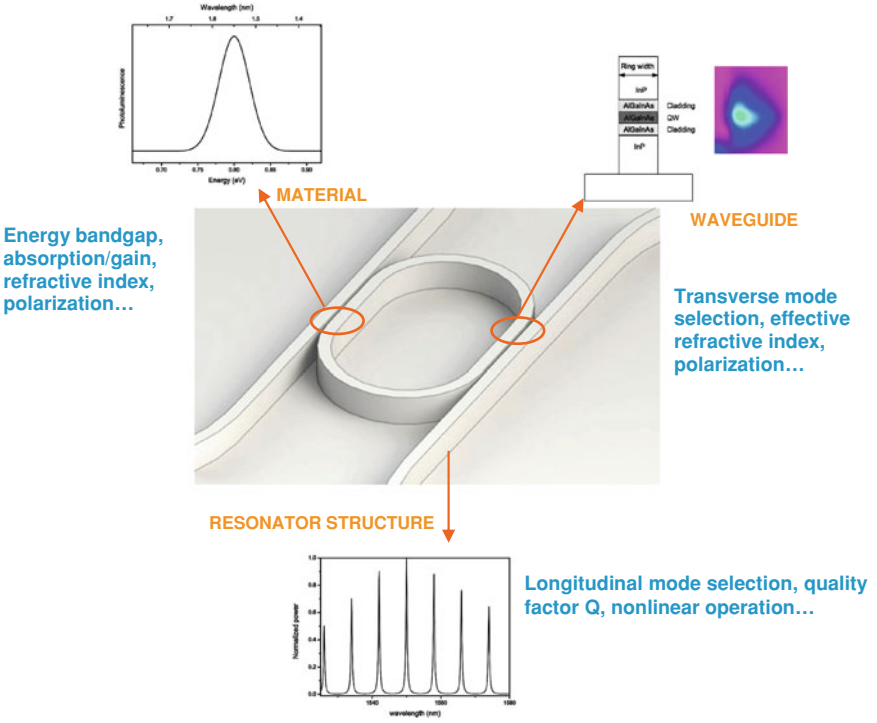


Fig. 8.1 A photonic device paradigm

consequences of the above elements to the device's performance. In broad terms, the photonic material contributes to the device material gain but also determines other optical properties like nonlinearities via e.g. carrier-dependent dispersion of the refractive index. The geometry of the waveguide structure, whose effects are usually lumped in the effective index, affects the transverse mode profile and the confinement factor, thus modifying the interaction of the oscillating field with the active material. Finally the resonator shapes the longitudinal mode structure and determines the free spectral range (FSR) and the quality factor, Q, of the cavity. In other words, the photonic material, the waveguide structure and the resonator are effectively three knobs that can be used to tune the photonic's device properties to the application's requirements.

The bulk of the chapters of this book has dealt with the physics of the gain material. In Chap. 7 the reader is introduced to photonic devices (semiconductor lasers and SOAs). In this chapter we develop the concepts presented there to apply to two novel resonator structures, those of vertical cavities (VC) and μ-ring resonators (MRs).

Schematics of VCs and MRs are shown in Fig. 8.2.What differentiates VCs from conventional edge emitters like Fabry–Perot (FP) cavities is that emission occurs along the direction of growth [1]. This feature as will be explained in more detail in Sect. 8.2 offers grounds not only for exciting physics but also for many applications. MRs consist of a circular waveguide and a bus waveguide [2]. The resonance is

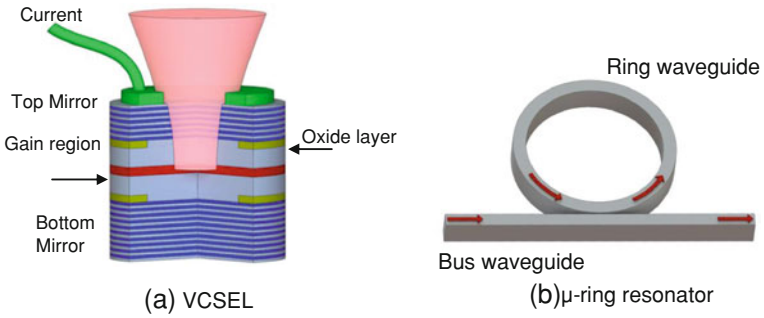


Fig. 8.2 Schematic of vertical cavities and μ-ring resonators. **a** VCSEL, **b** μ-ring resonators

created from the requirement that the field propagating in the bus waveguide and the ring waveguide have the same phase at the coupler. In this sense MRs are travelling wave devices as opposed to standing-wave devices (like FPs and VCs) where the resonance is created from the multiple internal reflections from the front and back mirrors. The fact that no facets or gratings are required for optical feedback makes MRs ideal for integration with other passive or active elements.

The rest of the chapter is divided into two parts, one dealing with VCs (Sect. 8.2) and the other with MRs (Sect. 8.3). For the VCs the fundamental concepts and design implications of these are analysed. The importance of high reflective distributed Bragg reflectors (DBRs) is outlined and the relevant theory is presented. Given that there are a few books that deal extensively with vertical-cavity surface-emitting laser (VCSEL) modelling the emphasis here is put on the design and applications of vertical cavity semiconductor optical amplifiers (VCSOAs). The first part closes with the advanced topic of the performance characteristics of spin-polarised VCSELs.

In the second part we deal with MR resonators. The analysis commences with the fundamental theory of MR by means of analytic relations. The waveguiding properties are crucial for the design of MRs and hence are treated in some detail. Single and multi-ring configurations are discussed and analysed. The modelling and design of active MR structures is then presented. Finally, the potential of MRs as processing elements operating below threshold is highlighted via an FP-like modelling approach.

8.2 Vertical Cavities

8.2.1 Basic Design Concepts

The distinctive difference of a Vertical Cavity from the edge emitting laser (EEL) counterparts is that the optical cavity is orthogonal to those of EEL as shown in Fig. 8.2a. The research and commercial interest in this peculiarity in the cavity orientation is fuelled by advantages inherent in the vertical geometry [3, 4]:

- VCs are particularly cost-effective devices that can be produced in high yields. Production cost is greatly reduced by the ability for on-wafer testing and monolithic fabrication of VCs. Also, it is possible to fabricate VCs in arrays and matrices, a feature that can be exploited, for e.g. parallel processing applications.
- The vertical geometry of the cavity allows fabrication flexibility absent from EEL. It is fairly straightforward to fabricate VC of cylindrical shape with wide emission surface. The optical beam emitted from this structure has a circular profile and low divergence, and therefore can be easily coupled to other optical components, e.g. standard telecom fibre.
- The small cavity volume and small cavity length affect the operational characteristics of VCs in several ways; VCSELs exhibit very low thresholds and VCs can be easily designed to support a single longitudinal mode. Both these features prove quite challenging for EEL.

Although initially VCs were intended for telecommunication applications, the aforementioned advantages have expanded the range of applications to biosensing, high density optical storage and imaging, to mention a few [5].

The VC literature is already very rich and includes a number of specialised books [3, 4]. The bulk of the published material concerns laser applications. In this chapter we put emphasis on VC applications as amplifiers and also highlight the emerging field associated with VCSEL polarisation properties. The interested reader is referred to, e.g. [3, 4] for a detailed account of VC for laser applications.

There are four different sets of main design parameters that need to be optimised for VC for operation either as lasers or amplifiers.

- Optical feedback. The mirrors must have sufficient reflectivity over the correct wavelength range.
- The resonance wavelength must be close to the Bragg wavelength of the mirrors.
- The active material's maximum gain must be aligned with the cavity resonance
- If multi-quantum wells (MQW) are used these must be positioned on the cavity antinodes.

However, apart from the above there are also some additional technological and operational challenges that need to be considered. VCSELs emit in only one longitudinal mode, as imposed by the small length of the cavity, but can emit in several transverse modes which can compromise performance. Also, confinement of photons and electrical current is very important for minimization of threshold current and efficiency maximisation. An additional issue is the heat generation due to the small active dimensions and current flow through the DBRs that induces variation of the refractive index. On top of these one should also add the requirement for sufficient output power. Producing a universal design that addresses all the above is indeed a difficult task, as it is often the case that a design scheme that remedies some parameters will deteriorate others. In this sense the appropriate VC structure is application-dependent.

The optical field can be confined using gain guiding, index guiding and antiguiding mechanisms [5]. Current leakage can be minimised using ion implantation to

define current paths through change of the electrical resistivity. An electrical path for the injection current can be provided by additional doped layers on either surface of the active layer (tunnel junction). The most popular VCSELs that are also massively produced are oxide aperture VCSELs. These are index guided structures that provide strong confinement of the optical field due to the refractive index difference between the oxide layer and that of the semiconductor layer. Also, oxide apertures, being insulating layers, confine injection current through the aperture. Overall oxide aperture VCSELs exhibit low threshold currents and enhanced efficiency but limit output power.

The purpose of the above short account of the various technological open issues in VC technology is to highlight the interrelation of the various parameters involved in VC device operation. A more detailed account of device details is certainly beyond the scope of the chapter at hand and can be found elsewhere [3, 4].

In the following sections we will develop a better understanding on how these four sets of design parameters affect performance.

8.2.2 Optical Feedback and DBRs

The small length of the cavity ($<1\,\mu$m) translates into small values of single pass gain. In order to reach lasing, the optical field must experience gain through many roundtrips. For this to happen, the mirrors must exhibit very high reflectivities, higher than 99%. This is by no means an easy task and requires mirror technology that goes well beyond the cleaved facets of simple EELs. High reflectivities can be obtained from (i) metal mirrors, (ii) stacks of dielectric layers and (iii) stacks of epitaxial layers [3, 4]. Of the three, the latter is the most attractive fabrication-wise as it permits monolithic growth of the whole structure providing that the Bragg stack material and the active material are compatible.

Usually the active material used for telecom wavelengths (1.3 and 1.55 μm) in EEL is InGaAsP [6]. The major drawback of this alloy is that it suffers from temperature-dependent losses and has motivated research in alternative material systems like AlGaInAs [7] and GaInNAs [8] as well as (Ga)InAs/GaAs Quantum Dots (QDs) [9] for EEL applications. Naturally, it would be expected that the active material technology of EEL, namely InGaAsP alloys, would be adopted for VCs as well. However, this is not favoured for a number of fabrication and performance issues. It is noted that the choice of material system for telecom VCSELs is a compromise between fabrication complexity and performance requirements.

It is preferable fabrication-wise that the VCSEL structure is monolithically fabricated. Using InP-based active material the requirement for monolithic fabrication obliges the use of compatible InP-based DBRs. In this case high reflectivities can only be achieved with InP-based DBRs with nearly 40 pairs due to the small refractive index difference of the DBR layers. Additionally InP-BDRs have poor thermal properties. GaInNAs poses an alternative solution as it is compatible with GaAs/$Al_{1-x}Ga_xAs$ DBRs with high refractive index contrast and thus few number of

layers. Despite the advances in GaInNAs material growth, it still lacks the quality compared with InGaAsP and AlGaInAs counterparts. The incompatibility of DBRs with appropriate refractive index difference with InGaAsP has motivated the development of wafer fusion technique for the fabrication of InGaAsP-based VCSELs with high contrast GaAs/$Al_{1-x}Ga_xAs$ DBRs and already there are commercially available VCSELs that employ this technique [10, 11]. Other commercial available VCSELs involve more sophisticated designs: Vertilas use one epitaxial AlInGaAs/AlInAs DBR and one dielectric DBR and incorporate a buried tunnel junction for optical and current confinement [12]. Successful monolithic fabrication of InP-based VCSELs with undoped-InAlGaAs–InAsAs DBRs (28 and 38 pairs for the top and bottom DBRs respectively) has been demonstrated and subsequently commercialised. In Ref [13] the limitation of large numbers of layers is relaxed with the use of InAlGaAs phase-matching layer and an Au metal layer in order to increase the reflectivity. Additionally, the use of undoped layers for the DBRs result in suppressed free carrier absorption loss.

8.2.2.1 Transfer Matrix Method for the Calculation of the DBR Reflectivity

DBRs are composed of alternating layers of high and low refractive index with layer thickness of λ/4, where λ is the wavelength in each layer. As mentioned before it is important to control the reflectivity of the DBRs and the wavelength of maximum reflectivity. Transfer matrix is a convenient method for the calculation of the DBR reflectivity. The optical fields propagating in (E^-) and out (E^+) of the ($i+1$)-layer are related to the fields propagating in and out of an adjacent i-layer by the matrix equation (see Fig. 8.3) [3], [14]:

$$\begin{bmatrix} E^+ \\ E^- \end{bmatrix}_{i+1} = \begin{bmatrix} \exp(-jk_ih_i) & 0 \\ 0 & \exp(jk_ih_i) \end{bmatrix} \begin{bmatrix} \frac{1}{t_i} & -\frac{r_i}{t_i} \\ -\frac{r_i}{t_i} & \frac{1}{t_i} \end{bmatrix} \begin{bmatrix} E^+ \\ E^- \end{bmatrix}_i \equiv M \begin{bmatrix} E^+ \\ E^- \end{bmatrix}_i \quad (8.1)$$

where $k_i = (2\pi N_i/\lambda) - j\alpha_{\text{abs},i}$, N_i the refractive index of the i layer and $\alpha_{\text{abs},i}$ the absorption of the i layer. The term:

$$\begin{bmatrix} \exp(-jk_ih_i) & 0 \\ 0 & \exp(jk_ih_i) \end{bmatrix} \quad (8.2)$$

accounts for the phase accumulation due to field propagation in the i layer of width h_i. The term:

$$\begin{bmatrix} \frac{1}{t_i} & -\frac{r_i}{t_i} \\ -\frac{r_i}{t_i} & \frac{1}{t_i} \end{bmatrix} \quad (8.3)$$

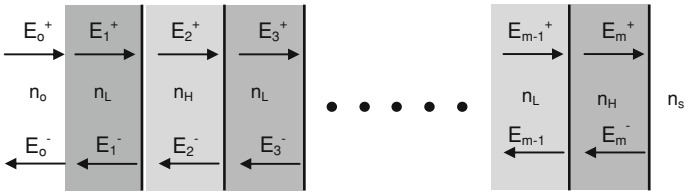


Fig. 8.3 Distributed Bragg reflector schematic

accounts for the effects of transmission (t_i) and reflection (r_i) that the field experiences at the interface of the two layers.

Using the transfer matrices the optical fields transmitted and reflected from the m layer are related to the optical fields transmitted and reflected from the 0 layer by:

$$\begin{bmatrix} E^+ \\ E^- \end{bmatrix}_m = \prod_{i=1}^{m-1} M_i \begin{bmatrix} E^+ \\ E^- \end{bmatrix}_i \equiv \begin{bmatrix} m_{11}\ m_{12} \\ m_{21}\ m_{22} \end{bmatrix} \begin{bmatrix} E^+ \\ E^- \end{bmatrix}_0 \tag{8.4}$$

The effective field reflectivity observed at the 0 layer is given by:

$$r_{\text{eff}} = \frac{E_0^-}{E_0^+} \equiv \frac{r_m m_{11} - m_{21}}{m_{22} - r_m m_{12}} \tag{8.5}$$

From (8.4) to (8.5) the effect number of layers m on the reflectivity characteristics can be calculated as shown in Fig. 8.4, for AlAs/GaAs DBRs (the refractive indices for AlAs and GaAs are 2.89 and 3.45 respectively).

8.2.2.2 The Assumption of Hard Mirrors

A convenient assumption that lends itself for integration with dynamic device models, is that of *hard mirrors*. According to this, the DBRs are approximated as single layer hard mirrors whose reflectivity is given by [15]:

$$R_{\text{DBR}} = \left(\frac{1 - qp^{2m-1}b}{1 + qp^{2m-1}b} \right)^2 \tag{8.6}$$

where: p, q and b are the low-to-high refractive index ratio of intermediate layer, first DBR interface and last DBR interface, respectively, given by $p = \frac{N_{\text{LOW}}}{N_{\text{HIGH}}}$, $q = \frac{N_{\text{LOW}}}{N_{\text{HIGH}}}\Big|_{1^{\text{st}}\text{DBR interface}}$ and $b = \frac{N_{\text{LOW}}}{N_{\text{HIGH}}}\Big|_{\text{last DBR interface}}$.

Application of the method requires the modification of the cavity length. The field penetrates and propagates through the DBR and thus experiences phase change. The effects of the propagation in the DBRs in the context of the assumption of hard mirrors are approximated by the length [15]:

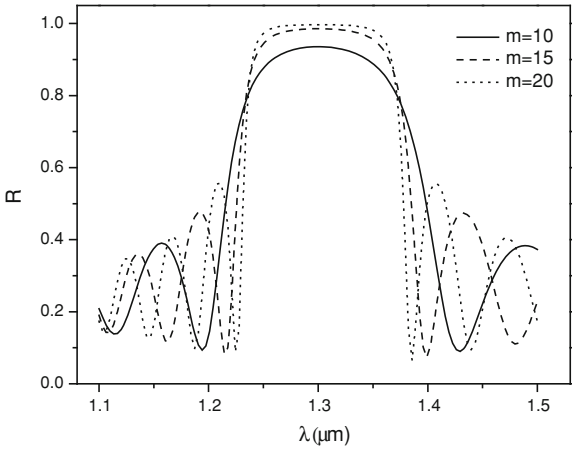


Fig. 8.4 Reflectivity of AlAs/GaAs DBR for various numbers of pairs as indicated in the figure

$$L_p = \frac{\lambda_c}{4N_c} \cdot \frac{q}{1-p} \cdot \frac{(1-b^2p^{2m-1})(1-p^{2m})}{1-q^2b^2p^{4m-2}} \tag{8.7}$$

In the above expression, λ_c is the cavity resonance in free space and N_c is the cavity refractive index. We will use this assumption in the analysis presented in Sects. 8.2.4 and 8.2.5.

8.2.3 Material Gain

The choice of material gain determines the wavelength of operation and it is dictated by the application. Whatever this is, the material gain peak must coincide with the cavity resonance. The fine-tuning can be achieved with bandstructure engineering presented in detail in Chap. 7 and involves the use of multi-quantum well structures (MQW). Given that in MQW-VCSELs the active material does not span the whole cavity, the maxima of the optical field (field antinodes) must be aligned with the MQWs. The gain enhancement factor ξ, discussed in the following section is a useful parameter that quantifies the interaction of the optical fields with the active material.

8.2.4 The Gain Enhancement Factor

The gain experienced by the optical mode oscillating in the cavity is given by (see 7.8, Chap. 7):

$$g = \Gamma g_{\mathrm{m}} - \alpha_{\mathrm{loss}} \tag{8.8}$$

where g_{m} is the material gain and α_{loss} the loss. Γ is the confinement factor that consists of three terms:

$$\Gamma = (\Gamma_x \Gamma_y) \cdot \Gamma_z \tag{8.9}$$

$\Gamma_{xy} = \Gamma_x \Gamma_y$ is the lateral confinement factor commonly used in EEL laser modelling (see 7.7 of Chap. 7) and Γ_z is the longitudinal factor [16]. For EEL Γ_z is usually not considered as it is unity for most cases, expect some (e.g. DBR lasers) where it is represented by the fill factor. In VCs on the other hand it is the *longitudinal factor* that is important while Γ_{xy} is unity due to large lateral dimensions (>5 μm). Γ_z is defined as the ratio of optical intensity confined in the active region along z, to the total intensity distributed in the cavity of length L:

$$\Gamma_{\mathrm{z}} = \frac{\int_{\mathrm{active}} E^2(z)\mathrm{dz}}{\int_{\mathrm{L}} E^2(z)\mathrm{dz}} \tag{8.10}$$

$E(\mathrm{z})$ is the electric field which is approximated by $E(z) = E_o \cos(\mathrm{kz})$. The longitudinal confinement can be calculated by substituting the expression for E(z) into (8.10). Integration can be simplified if it is performed over the length t, of one period and the result is multiplied by the ratio of the number of active segments (d/t) over the number of half wavelengths $L/(\lambda/2)$ (periods)

$$\Gamma_{\mathrm{z}} = \frac{\mathrm{d}\lambda}{2Lt} \frac{\int_t \cos^2(kz)\mathrm{dz}}{\int_{\lambda/2} \cos^2(kz)\mathrm{dz}} \tag{8.11}$$

Integration yields:

$$\Gamma_z = \frac{d}{L}\left\{1 + \frac{\sin\left(2\pi N_c \frac{t}{\lambda_c}\right)}{2\pi N_c \frac{t}{\lambda_c}}\right\} = \frac{d}{L} \cdot \xi \tag{8.12}$$

where

$$\xi = 1 + \frac{\sin(2\pi N_c t/\lambda_c)}{2\pi N_c t/\lambda_c} \tag{8.13}$$

ξ is termed as the *gain enhancement factor* [16] and attains values from 1 to 2. When the QWs are placed on the field maxima, interaction is maximised producing maximum gain. This is the case that corresponds to $\xi = 2$. When, on the other hand, the QWs are not aligned with the standing wave antinodes, then the part of the active material is rendered useless thus not contributing to the stimulated emission process. This case is described by $\xi = 1$. It is apparent that ξ is a crucial design parameter.

8.2.5 Vertical Cavity Semiconductor Optical Amplifiers (VCSOA)

The theoretical framework for the modelling of SOAs and EELs is presented in detail in Chap. 7. This is applicable to the cases of VCSOAs and VCSELs with the appropriate changes to account for the "peculiarities" of the vertical geometry i.e. the DBR reflectors through the assumption of hard mirrors and the modification of the confinement factor, and thus modal gain, to include the gain enhancement factor. As mentioned in the introductory comments of this chapter, VC for conventional lasing applications have been widely studied, hence we choose to focus on VC as amplifiers and also deal with manipulation of VCSEL polarisation states.

For optical amplifiers there are various degrees of model accuracy depending on the application and timescales of the phenomena exploited. The most popular methods are the rate equation (RE) method [17] and the Fabry Perot (FP) method [18]. The discrepancies between the two methods were resolved by Royo and co-workers [19].

The VCSOA gain in transmission and reflection are given by the known expressions for FP etalons (see 7.9 and 7.52 of Chap. 7)

$$G_T = \frac{(1-R_f)\cdot(1-R_b)G_s}{(1-\sqrt{R_bR_f}G_s)^2 + 4\sqrt{R_bR_f}G_s\sin^2\phi} \tag{8.14}$$

$$G_R = \frac{(\sqrt{R_f}-\sqrt{R_b}G_s)^2 + 4\sqrt{R_bR_f}G_s\sin^2\phi}{(1-\sqrt{R_bR_f}G_s)^2 + 4\sqrt{R_bR_f}G_s\sin^2\phi} \tag{8.15}$$

G_s is the single-pass gain given by $G_s = \exp[\xi g_{\mathrm{m}} t - a_{\mathrm{loss}} L]$. The phase shift ϕ, is described by:

$$\phi = \frac{2\pi N_c L}{hc}\left(E_{\mathrm{signal}} - E_r\right) + \frac{2\pi L}{hc}E_{\mathrm{signal}}\Delta N \tag{8.16}$$

E_r is the energy of the resonance. The first term in the r.h.s. of (8.16) is the linear phase shift due to the spectral difference (detuning) between the signal wavelength and the cavity resonance. The second term is the nonlinear phase change that gives rise to nonlinear effects. More specifically, the origin of the nonlinear effects can be traced to the dispersion of the refractive index, ΔN, with carrier concentration (n). There are many physical mechanisms that affect the refractive index [20]; the principal contributions to the carrier-dependent refractive index are (a) the bandgap shrinkage (b) the free carrier plasma effect, often termed as free carrier absorption and (c) the carrier induced shift of gain or absorption which alters refractive index through the Kramers–Kronig relation. For active media the dominant mechanism is the carrier-induced shift of gain and ΔN reads as:

$$\Delta N = (n - n_1)\frac{\mathrm{d}N}{\mathrm{d}n} \tag{8.17}$$

where n_1 is the carrier concentration when the input signal is absent and $\mathrm{d}N/\mathrm{d}n$ is the differential refractive index.

The RE approach involves the solution of the rate equations for photons S and carriers n (8.18, 8.19) [17]

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{j}{eL_{\mathrm{MQW}}} - \frac{N}{\tau_{\mathrm{e}}} - \frac{\xi\Gamma c}{N_{\mathrm{g}}} g_{\mathrm{m}} S \tag{8.18}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = R_{\mathrm{signal}} - \beta\Gamma R_{\mathrm{sp}} + \frac{\xi\Gamma c}{N_{\mathrm{g}}} g_{\mathrm{m}} S - \frac{S}{\tau_{\mathrm{p}}} \tag{8.19}$$

R_{signal} is the pumping rate related to input power $\mathrm{P_{in}}$, τ_{e} is the electron lifetime and τ_{p} is the photon lifetime. L_{MQW} is the total width of the Quantum Wells (QWs), j is the injected current density, e is the electron charge, β the spontaneous emission factor, Γ the longitudinal confinement factor (or fill factor) defined by the ratio of the active length over the cavity length (see 8.12), and R_{sp} is the spontaneous emission rate defined as $R_{\mathrm{sp}} = Bn^2$ where B is the bimolecular recombination constant. The details of the vertical cavity are expressed via ξ and τ_p. At steady state the photon density S (including amplified spontaneous emission photons) can be expressed in terms of material gain, $\mathrm{g_m}$. The resulting expression can be replaced in (8.18) to solve for carriers and hence g_{m} to calculate G_{s} and from (8.14) to (8.15) the amplifier characteristics.

In the context of the FP method, the rate that corresponds to signal photons and the term that corresponds to amplified spontaneous emission (ASE) photons, can be decoupled from emission rate expressed by the term $\frac{\xi\Gamma c}{N_{\mathrm{g}}} g_{\mathrm{m}} S$ of (8.18). The two different contributions can be found by applying the appropriate boundary conditions and solving the z-dependent field equation. FP method can be further simplified by applying the Longitudinally Average Travelling Wave Approach. (LTWA) by Adams [18]. The underlining assumption of LTWA is that the photon density (signal and amplified spontaneous emission) is uniform throughout the cavity and approximated by average quantities S_{sig} for signal photons and S_{spon} for amplified spontaneously emitted photons.

$$S_{\mathrm{spon}} = \left[\frac{(G_s - 1) \cdot \left[(1 - R_b)(1 + R_f G_s) + (1 - R_f)(1 + R_b G_s)\right]}{gL_c(1 - R_f R_b G_s^2)} - 2 \right] \times \frac{R_{sp} N_g}{gc} \tag{8.20}$$

$$S_{\mathrm{sig}} = \left[\frac{(G_s - 1) \cdot (1 - R_f)(1 + R_b G_s)}{\left(1 - \sqrt{R_f R_b} G_s\right)^2 + 4\sqrt{R_f R_b} G_s \sin^2 \phi} \right] \times \frac{P_{\mathrm{in}} N_g}{E_{signal} \Pi (W/2)^2 L_c g c} \tag{8.21}$$

In the above, R_f and R_b are the reflectivities of the front and back mirrors respectively, using the hard mirrors assumptions and L_c is the effective cavity length. Using (8.20) and (8.21) the rate equation for the carrier density can be modified as:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{j}{eL_{\mathrm{MQW}}} - \frac{n}{\tau_e} - \frac{\xi\Gamma c}{N_g} \left(g_{\mathrm{m}}(E_r) S_{\mathrm{spon}}(E_r) \beta(E_r) + g_{\mathrm{m}}(E) S_{\mathrm{sig}}(E) \right) \tag{8.22}$$

For given current density (8.22) yields the carrier and thus material gain in steady state. The single-pass gain, G_s, can then be calculated and substituted in the expressions for the VCSOA gain in transmission and reflection (8.14, 8.15). The amplifier gain of GaInNAs VCSOAs calculated following the methodology outlined above, is shown in Fig. 8.5.

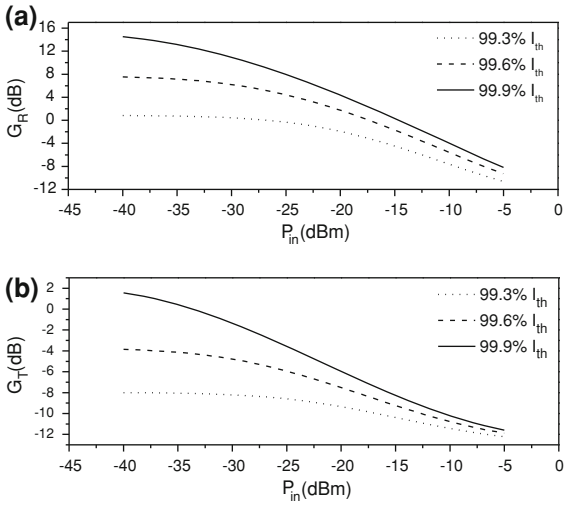


Fig. 8.5 VCSOA gain in **a** reflection and **b** transmission of an 8 QW $Ga_{0.65}In_{0.35}N_{0.025}As_{0.975}$/GaAs VCSOA structure for various currents. The maximum current is calculated to be $I_{th} = 12.9\,\text{mA}$. The back and front mirror reflectivities are $R_b = 0.985$ and $R_f = 0.999$ respectively. The effective cavity length is $4.22\,\mu\text{m}$

An alternative approach to the FP method bypasses the intermediate step of the RE for the solution of the carrier and photon rate equations to calculate the material gain for given current injection conditions, and thus G_R and G_T. Instead, it is approximated with as a function of the average optical intensity in the cavity I_{av}, the saturation intensity I_s and the saturation gain, g_o, via the expression (in the steady state) [21]:

$$gL = \frac{\Gamma g_o L}{1 + I_{\text{av}}/I_s} - \alpha_{\text{loss}} L \tag{8.23}$$

where g is the modal gain and I_s is given by $I_s = E/\Gamma\alpha_{\text{loss}}\tau_e$, see (7.58). The single-pass phase change is then approximated as:

$$\phi = \phi_o + \frac{g_o L\alpha}{2}\left(\frac{I_{\text{av}}/I_s}{1 + I_{\text{av}}/I_s}\right) \tag{8.24}$$

The ratio of the transmitted (I_{trans}) and reflected (I_{ref}) intensities with I_s are expressed in a straightforward manner from (8.14) and (8.15) as

$$\frac{I_{\text{trans}}}{I_s} = \frac{I_{\text{in}}}{I_s}\frac{(1 - R_f)\cdot(1 - R_b)G_s}{(1 - \sqrt{R_b R_f}G_s)^2 + 4\sqrt{R_b R_f}G_s \sin^2\phi} \tag{8.25}$$

$$\frac{I_{\text{ref}}}{I_s} = \frac{I_{\text{in}}}{I_s}\frac{(\sqrt{R_f} - \sqrt{R_b}G_s)^2 + 4\sqrt{R_b R_f}G_s \sin^2\phi}{(1 - \sqrt{R_b R_f}G_s)^2 + 4\sqrt{R_b R_f}G_s \sin^2\phi} \tag{8.26}$$

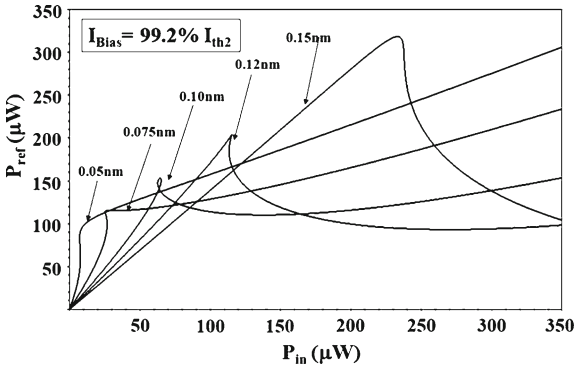


Fig. 8.6 VCSOA nonlinear characteristics: calculated reflected verses input power relationship for a 1550 nm VCSOA exhibiting different shapes of OB and nonlinear switching for applied bias current and initial wavelength detuning as indicated in the figure. (Reprinted with permission from [22] Copyright 2010, American Institute of Physics)

The ratio of average optical intensity to I_s is given by

$$\frac{I_{\text{av}}}{I_{\text{s}}} = \frac{I_{\text{trans}}}{I_{\text{s}}} \frac{(1 + R_{\text{b}} G_{\text{s}}) \cdot (1 - G_{\text{s}}^{-1})}{(1 - R_{\text{b}}) G_{\text{s}}} \tag{8.27}$$

Equations (8.25–8.27) can be solved numerically to give transmitted and reflected intensities. The powerful point of the method is that performance is analysed using device parameters that are readily available from experiment and therefore lends itself for fast estimations of device performance and comprehension of the effects of various magnitudes.

FP method is applicable for the cavities where the expressions of gain in transmission and reflection (8.14, 8.15) hold. For VC the FP method is applied along with the assumption of hard mirrors. The VC effects are then accounted for in the modified expressions for the reflectivities and effective cavity length. The FP method is applied in Fig. 8.6 to calculate the nonlinear reflected verses input power characteristics.

8.2.6 VCSEL Polarisation Properties and Spin-VCSELs

VCs unusual polarisation properties offer solid ground for interesting device physics and applications. The plurality in VC polarisation states arises from a combination of factors [23];

- VC emit in only one longitudinal mode (due to the small cavity length), but in several transverse modes.
- VCs most commonly have circular profiles hence based on the cavity waveguide profile, emit in both orthogonal modes that should be in theory degenerate.
- Quantum confinement is perpendicular to the emission axis which yields modified selection rules (see Fig. 8.7).

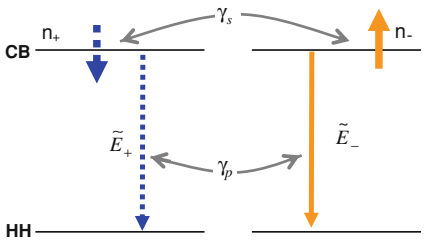


Fig. 8.7 Optical selection rules in the context of the SFM model. $n_+(n_-)$ carrier densities refer to spin-down (spin-up) populations which couple to the right- (left-) circularly polarised field component $E_+(E_-)$. $E_+(E_-)$ are coupled though the birefringence decay rate γ_p, whereas the $n_+(n_-)$ through the spin relaxation rate γ_s

- The relative orientation between optical emission and the crystal axes leads to birefringence thus relaxing degeneracy of the two orthogonal polarisations.
- The nonlinear susceptibilities induce self- and cross-saturation thus modifying the optical gain of the two polarisations contributing to the lift of degeneracy.

From a device viewpoint, the two orthogonal polarisations states may exhibit a plethora of nonlinear phenomena like switching and bistability that form the basis for various device functionalities. These can be triggered by applied bias current, optical injection and variation of device temperature.

The interdependence of the various factors listed above that affect VCSEL properties are best described with the Spin-Flip Model developed in [24]. The polarisation state of the emitted light depends on the angular momentum of the quantum states involved in the emission /absorption process and the details of the laser cavity. Electron–hole recombination occurs through two distinct carrier densities that differ in spin orientation. The two carrier densities (spin up and spin down) are coupled through spin-flip processes characterised with spin relaxation rate γ_s that tend to equalise the two. Recombination of electrons and holes with spin up yields right-circularly polarised light, whereas left-circularly polarised is emitted from recombination of electrons and holes with spin down. Left and right circularly polarised emissions are coupled through birefringence quantified with birefringence rate γ_p. The modes associated with the two circularly polarised emissions may experience different gain-to-loss ratio that leads to an amplitude anisotropy modelled with the gain anisotropy rate γ_α. All these effects are accounted for in the SFM model phenomenologically by means of the rates γ_s, γ_p, and γ_α.

The SFM rate equations are expressed more conventionally in terms of normalised carrier variables $N = (n_+ + n_-)/2$ and $m = (n_+ - n_-)/2$, where $n_+(n_-)$ is the spin down (up) carrier density. Optical pumping, is included through the circularly polarised pump components (η_+, η_-). Expressing the complex fields in terms of real and imaginary parts as $\bar{E}_\pm = E_{\pm,R} + i E_{\pm,I}$, the rate equations become:

$$\frac{\mathrm{d}E_{\pm,R}}{\mathrm{d}t} = \kappa\,(N \pm m - 1)\left(E_{\pm,R} - \alpha E_{\pm,I}\right) - \gamma_a E_{\pm,R} + \gamma_p E_{\mp,I} \tag{8.28}$$

$$\frac{\mathrm{d}E_{\pm,I}}{\mathrm{d}t} = \kappa\,(N \pm m - 1)\left(E_{\pm,I} + \alpha E_{\pm,R}\right) - \gamma_a E_{\pm,I} - \gamma_p E_{\mp,R} \tag{8.29}$$

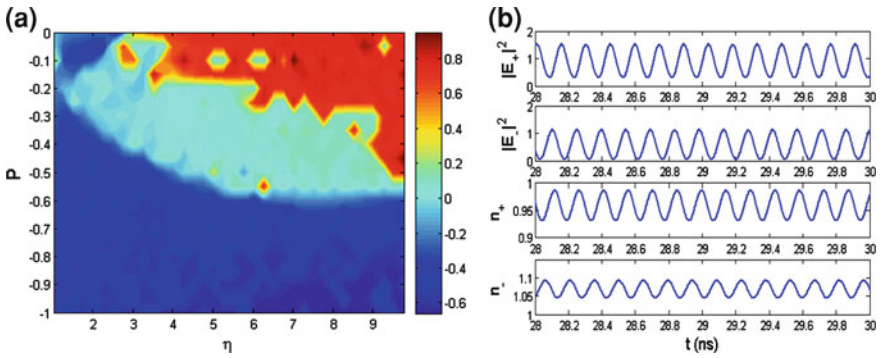


Fig. 8.8 **a** Stability maps and **b** time traces for spin-injected VCSELs. **a** ($\gamma_p = 20\,\mathrm{ns}^{-1}, \kappa = 125\,\mathrm{ns}^{-1}, \gamma_s = 10\,\mathrm{ns}^{-1}, \gamma = 1\,\mathrm{ns}^{-1}, \gamma_a = 0$ and $\alpha = 1$). Time series are calculated at the operating point $\eta = 2.4692$ and $P = -0.1231$. Reprinted with permission from [26] (© 2011 IEEE)

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \gamma \left[\eta_+ + \eta_- - \left(1 + \left|\bar{E}_+\right|^2 + \left|\bar{E}_-\right|^2\right) N - \left(\left|\bar{E}_+\right|^2 - \left|\bar{E}_-\right|^2\right) m\right] \tag{8.30}$$

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \gamma\left(\eta_+ - \eta_-\right) - \left[\gamma_s + \gamma\left(\left|\bar{E}_+\right|^2 + \left|\bar{E}_-\right|^2\right)\right] m - \gamma\left(\left|\bar{E}_+\right|^2 - \left|\bar{E}_-\right|^2\right) N \tag{8.31}$$

where $2\kappa = \tau_p^{-1}$ and $\gamma = \tau_e^{-1}$, with τ_p and τ_e as the photon and electron lifetimes, respectively, α is the linewidth enhancement factor, γ_p is the birefringence rate, γ_α is gain anisotropy rate and γ_s is the spin relaxation rate. The total normalised pump is $\eta = \eta_+ + \eta_-$, whereas the pump ellipticity, P, and the ellipticity of the output ε are defined as:

$$P = \frac{\eta_+ - \eta_-}{\eta_+ + \eta_-} \tag{8.32}$$

$$\varepsilon = \frac{\left|\bar{E}_+\right|^2 - \left|\bar{E}_-\right|^2}{\left|\bar{E}_+\right|^2 + \left|\bar{E}_-\right|^2} \tag{8.33}$$

The above equations can be easily modified to account for optical injection [25]. Equations (8.28–8.33) can be used to calculate the stability maps in the P, η plane for solitary optically pumped VCSELs (Fig. 8.8a) and corresponding time traces (Fig. 8.8b) for a specific operating point of Fig. 8.8(a) [26]. The dynamics are resolved using the Largest Lyapunov Exponent (LLE) method [27] whereby negative values of LLE correspond to stability, zero to oscillatory behaviour and positive values to regions of more complex dynamics tending towards chaos as the LLE increases.

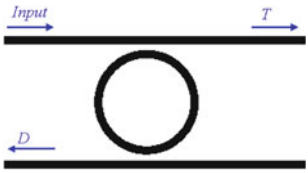


Fig. 8.9 Micro ring resonator in an Add/Drop multiplexer configuration

8.3 Microring Resonators

8.3.1 Fundamental Concepts

Micro ring resonators are formed by closing an optical path (e.g. a waveguide) upon itself to form a loop (see Fig. 8.9). Light can propagate either clockwise or counterclockwise along the closed loop while accumulating phase according to the wavelength and the effective index of the waveguide mode:

$$E(s) = E_0 \exp(ik_0 n_{\text{eff}} s) \tag{8.34}$$

where k_0 is the wavenumber in vacuum, n_{eff} is the modal index and s is a coordinate along the loop given by $R\Delta\phi$ where R is the radius of the micro ring and $\Delta\phi$ is the angle.

Because of the cyclic boundary conditions, the phase accumulation in a roundtrip along the resonator must be an integer multiple of 2π. Therefore, only a discrete set of frequencies, known as the resonance frequencies, can satisfy this condition and resonate in the micro ring:

$$\omega_m = mc/Rn_{\text{eff}} \tag{8.35}$$

where ω_m is the mth resonance frequency of the micro ring resonator and c is the velocity of light in vacuum. It should be noted that (8.36) is in fact a transcendental equation because n_{eff} is also frequency dependent. This point is highly important because it affects one of the fundamental properties of the micro ring resonator—the free spectral range (FSR, see below).

Practical waveguides also incorporate various propagation loss mechanisms such as absorption and Rayleigh scattering caused primarily by the roughness of the waveguide side walls [28]. Therefore, (8.34) describing the field at point s must be amended to include an exponential attenuation of the electric field:

$$E(s) = E_0 \exp(ik_0 n_{\text{eff}} s - \alpha s) \tag{8.36}$$

where α is the attenuation coefficient. The existence of propagation loss modifies the resonance condition by introducing an imaginary part to the resonance frequency or, equivalently, a cavity lifetime τ indicating the time a photon can resonate in the cavity before it is absorbed or scattered.

Probably the most important properties of a micro ring resonator are its FSR and quality factor (or equivalently, its Finesse). The FSR is the spectral separation between adjacent resonance frequencies and the quality factor (Q) is proportional to the cavity lifetime –$Q = \omega\tau$. The Q measures the cavity lifetime in units of the optical frequency periods [29]. Equivalently, the Finesse measures this lifetime in units of the roundtrip time, i.e. home many "roundtrips" the photon survives in the cavity before it is scattered or absorbed: $F = \tau/\tau_{rt}$ where τ_{rt} is the roundtrip time.

As mentioned above, the evaluation of the FSR is a somewhat subtle task because of the dependence of the effective index on the frequency. Approximating the dependence of the effective index on the frequency to the first order Taylor expansion, the FSR of a micro ring resonator is given by:

$$\Delta\omega_{\mathrm{FSR}} = c/Rn_g, \tag{8.37}$$

where n_g is the group index given by $n_g = n_{\mathrm{eff}}(\omega_0) + \omega_0 \cdot \mathrm{d}n/\mathrm{d}\omega$ and ω_0 is the angular frequency around which the effective index is expanded.

It is practically difficult to get access to an isolated micro ring resonator, i.e. injecting and extracting light into and out of it. Thus, in practical micro ring devices, I/O waveguide/s are coupled to the micro ring. Referring to Fig. 8.9, two I/O waveguides are coupled to the micro ring in order to inject and extract light into and out of the device. Note, that this is a specific configuration (often referred to as the add/drop multiplexer configuration) and configurations employing single or multiple I/O waveguides are equally possible.

Assuming ideal, loss-less, directional couplers with power coupling coefficients of κ_1 and κ_2 (see Fig. 8.9) and overall roundtrip loss of $L = 1 - \exp(-2\pi R \cdot \alpha)$, the field transmission function at the Through and Drop ports (see Fig. 8.9) are given by [30]:

$$D(\omega) = \frac{-\sqrt{\kappa_1\kappa_2}L^{1/4}\exp(i\varphi/2)}{1 - \sqrt{(1-\kappa_1)(1-\kappa_2)L}\exp(i\varphi)}, \tag{8.38a}$$

$$T(\omega) = \frac{\sqrt{1-\kappa_2} - \sqrt{(1-\kappa_1)L}\exp(i\varphi)}{1 - \sqrt{(1-\kappa_1)(1-\kappa_2)L}\exp(i\varphi)} \tag{8.38b}$$

where $\varphi = 2\pi \cdot \Delta\omega/\Delta\omega_{\mathrm{FSR}}$ and $\Delta\omega = \omega - \omega_0$ is the frequency detuning from the nearest resonance frequency (ω_0).

Figure 8.10 shows a theoretically calculated D and T spectra for $k_1 = 0.2$, $k_2 = 0.4$, $\alpha = 1\,\mathrm{dB/cm}$, $n_g = 1.57$ and ring radius $R = 20\,\mu\mathrm{m}$. As can be expected, the peaks (notches) at the Drop (Through) ports correspond to the resonance frequencies of the micro ring and the separation between successive resonance frequencies is the FSR. As can be expected, the FSR is determined by the circumference of the micro ring and the group index of the waveguide, where the longer the circumference and the larger the group index, the smaller the FSR. The spectral width of the resonance peaks are determined by *loaded* quality factor which is determined by the coupling coefficients of the propagation losses in the micro ring. This can be clearly seen from the denominator in (8.38)—note that the role of the signal attenuation per roundtrip, L, is identical to that of the coupling losses—$(1-\kappa)^{1/2}$.

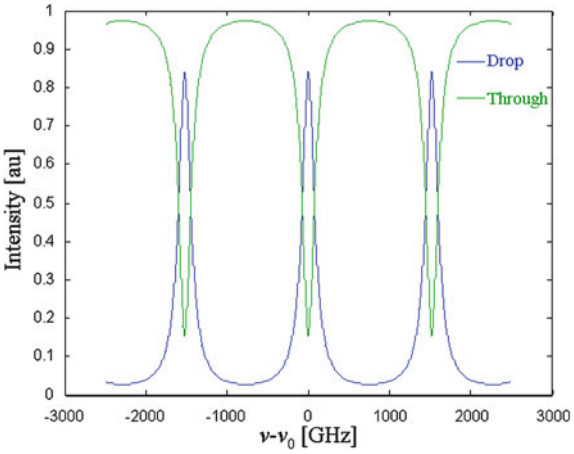


Fig. 8.10 Spectral transmission function of an add/drop multiplexer

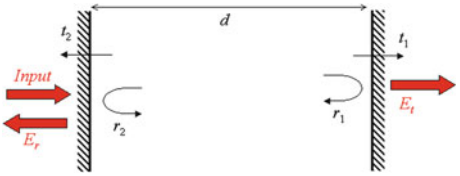


Fig. 8.11 Generic Fabry-Perot cavity

It is important to note that the spectral properties of the micro ring resonators are practically equivalent to those of the Fabry-Perot (FP) cavity as long as nonlinear effects are neglected. Figure 8.11 illustrates a generic FP resonator where the reflection and transmission coefficients at each interface are given, respectively, by $r_{1,2}$ and $t_{1,2}$ and the phase accumulation in a single roundtrip is given by $\phi = 2k_0nd$, where n is the refractive index and d is the thickness of the FP cavity.

The spectral responses of the transmitted and reflected signals are given by [31]:

$$E_r = \frac{-r_1 + r_2 \exp(i\phi)}{1 - r_1 r_2 \exp(i\phi)} E_{\text{in}}, \tag{8.39a}$$

$$E_t = \frac{t_1 t_2 \exp(i\phi/2)}{1 - r_1 r_2 \exp(i\phi)} E_{\text{in}} \tag{8.39b}$$

where E_r, E_t and E_{in} are, respectively, the reflected, transmitted and inserted fields. Note the complete analogy between r_i and $(1 - \kappa_t)^{1/2}$, t_i and κ_i, E_r and T, and E_t and D. Intuitively, one can consider the two half rings of the micro ring resonator as equivalent to the opposite propagation directions in the FP cavity, i.e. in the micro ring the optical paths of the counter-propagating waves are spatially separated. Consequently, the reflection coefficients in the interfaces of the FP cavity, which connect the amplitudes

of the counter-propagating waves in each roundtrip, are equivalent to the bar transmission coefficient of the directional couplers in the micro ring resonator.

Although it seems that the FP cavity and micro ring resonators are completely equivalent, there are two subtle differences which must be indicated. First, for high intensities where nonlinear effects are non-negligible, the FP and the micro ring differ because of the formation of a standing wave in the FP cavity (and not in the micro ring resonator). This standing wave generates an index modulation (because of nonlinear effects) which modifies the coupling and the interactions between the counter-propagating waves.

Another profound difference between the micro ring and the FP resonators is their response to rotation. A micro ring resonator circumvents a finite area and is, therefore, susceptible to Sagnac phase shift when subjected to rotation. As a result, when rotated, the resonance frequencies of the micro ring undergo a shift which depends on the magnitude and sign of the rotation. The FP cavity on the other hand does not circumvent an area and is, therefore, unaffected by rotation.

8.3.2 Waveguiding Properties of Micro Ring Resonators

In the previous section we have treated the properties of the curved waveguide comprising the micro ring as similar to those of a conventional (straight) waveguide. Although such analysis is fine for illustration and for extracting the main properties and features of the micro ring resonator, there are several subtle issues and differences that should be considered when an accurate analysis of such device is required.

In particular, there are two important effects which should be considered. The first is the transverse (or radial, to be precise) mode profile of a micro ring resonator. The lowest order mode of a conventional, straight waveguide with symmetric cladding (e.g. a symmetric slab waveguide in the 1-D case) is symmetric. In a curved waveguide, on the other hand, the reflection symmetry of the structure is removed because there is a clear difference between the inner and external radii of the waveguide boundaries.

The second effect is the emergence of radiation losses stemming from the curved geometry. This loss mechanism, often referred to as "bending losses" is inherent to the circular geometry and can be reduced by enlarging the bending radius and increasing the index contrast between the micro ring core and clad but cannot be eliminated completely.

Figure 8.12 shows the radial profile of silicon over insulator waveguide (250 nm × 400 nm) for various bending radii. As the bending radius is decreased the mode profile becomes more asymmetric and is shifted towards the external interface of the waveguide. An intuitive understanding of this phenomenon can be obtained by employing a conformal mapping to the radial and angular coordinates [32]:

$$\rho = R \cdot \exp\left(U/R\right), \tag{8.40a}$$

$$\theta = V/R \tag{8.40b}$$

where R is in arbitrary radius although it is convenient to set it as the radius of the micro ring.

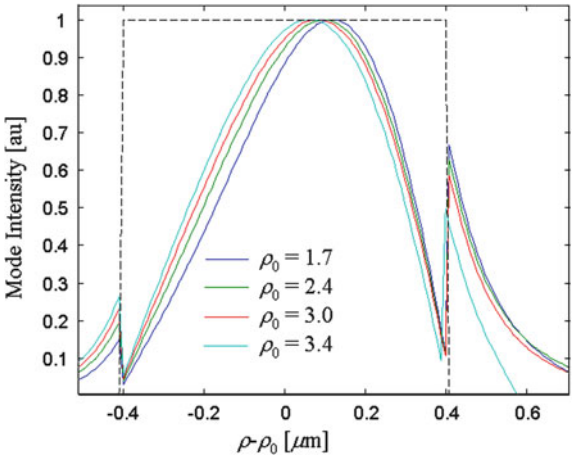


Fig. 8.12 Radial modal profile of SOI structure. The dashed line indicates the refractive index profile

The conformal transformation (8.40a) maps the wave equation in cylindrical coordinates into a Cartesian-like format (with U and V as the x and y coordinates) but with an equivalent index profile n_{eq}:

$$\frac{\partial^2 E}{\partial U^2} + \frac{\partial^2 E}{\partial V^2} + k_0^2 n_{\text{eq}}^2 (U)\, E = 0, \tag{8.41a}$$

$$n_{\text{eq}}(U) = n(U) \cdot \exp(U/R) \tag{8.41b}$$

The equivalent index is shown in Fig. 8.13 (the parameters are defined in the figure caption). Note, that the equivalent index increases exponentially as a function of U (i.e. it is effectively larger for larger radii). As a result, the mode profile is "pulled" towards the larger index resulting in an asymmetric profile.

The mechanism of the bending losses can also be explained by the effective index profile. Because the index effectively increases monotonically for larger radii there are no real confined modes for the curved waveguide, only leaky ones—a phenomenon which is manifested by the existence of bending losses.

8.3.3 Single and Multi-Micro Ring Configurations

The add/drop multiplexer scheme presented in Sect. 8.3.1 is probably one of the simplest micro ring configurations which have been studied. However, from the practical applications point of view, in particular for telecommunication applications, the single-micro ring add/drop configuration exhibits several deficiencies. The spectral response of a single add/drop multiplexer is periodic, thus as an optical filer it does not isolate a

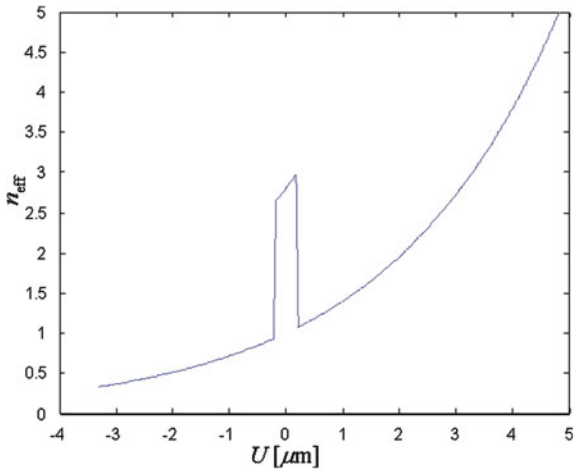


Fig. 8.13 Equivalent index profile of a 400 nm wide micro ring with $n_{\text{core}} = 2.8$, $n_{\text{clad}} = 1$, $R = 3$

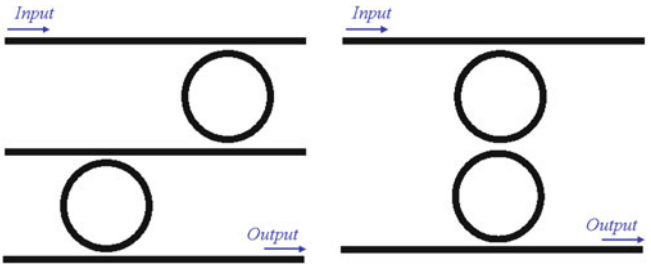


Fig. 8.14 Indirectly **a** and directly **b** coupled micro ring resonators

single frequency band but rather a set of bands which are separated by the FSR of the micro ring. For telecommunication applications, it is desired that the filter shape would exhibit a flat-top, sufficiently wide, profile and high extinction ratio. The single micro ring filter, on the other hand, possesses a Lorentzian line-shape which is inappropriate for data transmission. In addition, extending the filter bandwidth requires larger coupling coefficients which, in turn, significantly reduce the extinction ratio of the filter.

In order to resolve some of these deficiencies, multiple-micro ring configurations can be employed. Multiple micro ring filters can be realised by cascading single micro ring add/drop filters or by coupling the micro rings directly (see Fig. 8.14). It is interesting to note that despite the very different geometry, both configurations are equivalent (although not for the same coupling coefficients). In fact, both configurations can be employed in order to realise conventional Chebyshev or Butterworth optical filters [33].

By cascading several micro rings it is possible to achieve a flat-top profile, the desired bandwidth and the extinction ratio. The design parameters are the number of micro rings and the coupling coefficients which can be determined using the tools of digital signal

processing which allow to design the positions of the zeros and poles of the transmission function [33].

The periodic transfer function of the micro ring can also be amended by employing the Vernier effect and using multiple micro rings having different FSRs [34]. Consequently, only frequencies which are in resonance with all micro rings comprising the filter are passed instead of multiple bands. The employment of micro rings with different radii effectively increases the FSR of the device providing it with a single transmission range within the telecom band.

In addition to the enhanced FSR, the employment of the Vernier effect can facilitate the tunability of the device. In order to tune the transmission of micro rings-based filter composed of identical across the telecom band it is necessary to shift the resonance frequencies of the micro rings by at least 40 nm. However, if the Vernier effect is employed, it is necessary to shift these resonance frequencies by the largest FSR at most [34]. In this way, different resonances of the individual micro rings can be combined to achieve a transmission band across the complete band.

Coupled micro rings devices have interesting applications beyond filtering and telecommunications. Coupled resonator waveguides such as the CROW [35], SCISSOR [36] and related structures [37], exhibit slow group velocity and provide an attractive approach for the realisation of optical delay lines and optical storage devices [38], enhanced optical sensors [38] and more. In addition, the reduction in the group velocity is accompanied with an increase in the intensity of the field which can prove to be useful for exploiting optical nonlinearities at relatively modest power levels.

Coupled resonators slow-light structures have been studied extensively during the past decades. Figure 8.15 depicts schematics some of the most studied coupled micro ring structures. The CROW (Fig. 8.15(a)) consists of a series of directly coupled and identical micro resonators. The transmission properties of the CROW consist of passbands centred at the resonance frequencies of the individual resonators. The dispersion relations at the passband can be calculated either using a transfer matrix method [39] or a tight binding approach [40] and yield a cosine shaped relations:

$$\Delta\omega_K = \frac{1}{2}\Omega\Delta\alpha - \sqrt{\kappa}\cdot\Omega\cos(K\Lambda)/m\pi, \tag{8.42}$$

where $\Delta\omega_K = \omega_K - \Omega$ is the difference between the optical frequency and the resonance frequency of an individual micro ring. κ represents the coupling between the adjacent microdisks and $1/2\cdot\Omega\Delta\alpha$ is the self-frequency shift [40]. It should be noted that the most important parameter determining the dispersion relations is the coupling between adjacent coefficients. The stronger the coupling, the wider the passband which is formed around the resonance frequency. The coupling coefficient also determined the group velocity of the wave propagating along the CROW. Figure 8.16 depicts the dispersion relations of the CROW for various coupling coefficients. Decreasing the coupling coefficient results in a more shallow dispersion relations and, correspondingly, slower group velocity.

The group velocity at the center of the passband is given by:

$$v_g = \Omega\sqrt{\kappa}\Lambda/m\pi = 2\Lambda\cdot\Delta\nu_{\mathrm{FSR}}\cdot\sqrt{\kappa} \tag{8.43}$$

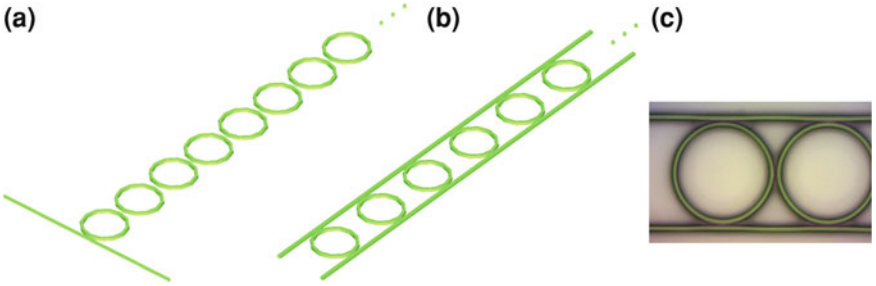


Fig. 8.15 Slow-light structures: **a** CROW; **b** SCISSOR; **c** SC-CROW

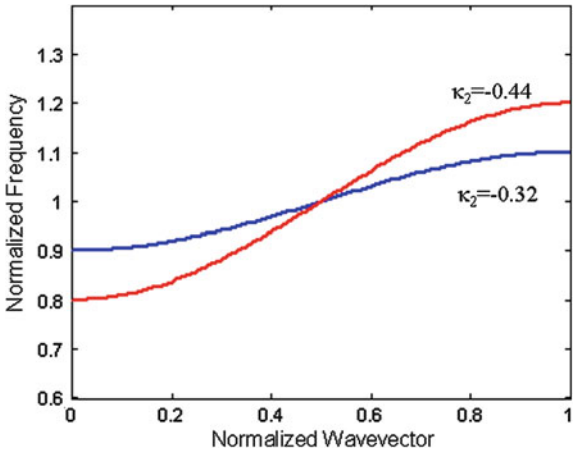


Fig. 8.16 Dispersion relations of a CROW

Although this is the fastest group velocity in the passband, the middle of the passband is the regime with the minimal group velocity dispersion which is crucial for reducing pulse distortions. Assuming a CROW consisting of N micro ring, the overall delay of the pulse can be approximated by:

$$\tau_d = N/2\Delta\nu_{\mathrm{FSR}}\sqrt{\kappa} \tag{8.44}$$

The corresponding bandwidth of the passband can be obtained directly from (8.42):

$$\Delta\omega_{\mathrm{band}} = 4\sqrt{\kappa}\cdot\Delta\nu_{\mathrm{FSR}} \tag{8.45}$$

Note, that the slower group velocities and longer delays accompanied to smaller coupling coefficients are achieved at the expense of narrower bandwidth. Taking the usable bandwidth of the CROW as the linear part of the dispersion relations (say, half the bandwidth), the delay-bandwidth product, which is one of the common figures of merit of delay lines, is:

$$\tau \cdot \Delta\omega_{\text{use}} = N \tag{8.46}$$

The direct trade off between the delay and the bandwidth is clearly seen because their product depends only on the number of resonators. (8.46) indicates an interesting and important feature of CROWs when it comes to performance, the specific details of the realisation are almost insignificant and the only parameter which matters is the number of micro rings. It should be noted, however, that this is not completely accurate because other parameters such as the FSR and loss are also important and although not directly related to the delay-bandwidth product, they do have a significant impact on the performance of the delay line. In particular, the overall loss of the signal propagating through the CROW can be estimated by:

$$\text{Loss} = aL_{\text{RT}}N/\sqrt{\kappa} \tag{8.47}$$

where L_{RT} is the physical roundtrip length of each resonator and a is the propagation loss coefficient of the waveguide comprising the micro ring. Thus, trying to improve the delay-bandwidth product results in larger overall transmission loss through the CROW, which is also a factor which must be taken into consideration.

The SCISSOR (Fig. 8.16(b)) consists of a series of *indirectly* coupled and identical micro-resonators. The micro rings are coupled to a mutual waveguide (or waveguides) which transfer the optical signal between the micro rings. There are several important differences between the transmission properties of the CROW and the SCISSOR. First, the SCISSOR exhibits to independent bands. The first set stems from the resonance frequencies of the individual micro rings and the second—from the periodicity of the structure determined by the length of the waveguide sections connecting the micro rings. Another important difference between the two structures is that in CROWs the transmission bands are centred on the resonance frequencies of the micro rings while in SCISSORs it is the bandgaps which are located at these frequencies.

Figure 8.17 depicts the dispersion relations of a SCISSOR. The parameters are defined in the figure captions. Despite the more complex structure, the dispersion relations of each transmission band of the SCISSOR are very similar to that of a CROW. The reason is that both bands stem from a similar mechanism—the periodicity of the micro ring and the SCISSOR structure. It should be noted, however, that other slow-light micro ring-based structures such as SC-CROW (see Fig. 8.15(c)) [37] exhibit different shapes of dispersion relation (i.e. not cosine shaped) with unique features such as mid-band zero group velocity points, etc.

8.3.4 Active Micro Ring Structures

The inherent feedback mechanism and resonance frequencies of micro ring resonators (as opposed to FP resonators which require the realisation of feedback mirrors) render them ideal for the realisation of integrated micro-lasers. In fact, active micro ring resonators, incorporating optical gain, can be realised in a manner similar to that of passive micro rings. Figure 8.18 shows a schematic of an active micro ring resonator with an I/O

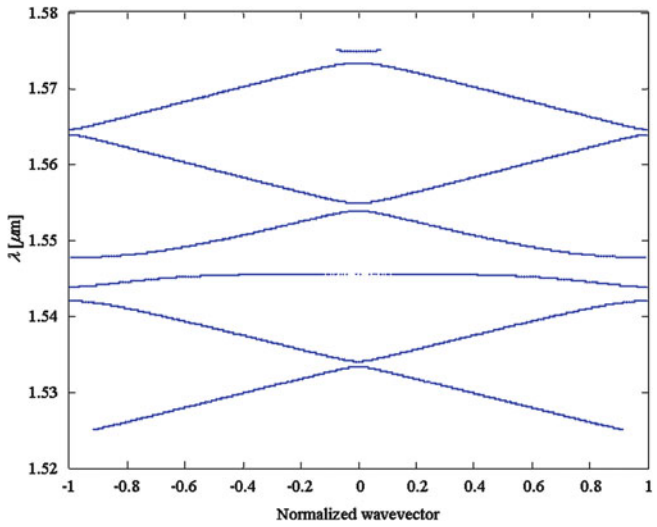


Fig. 8.17 Dispersion relations of a SCISSOR

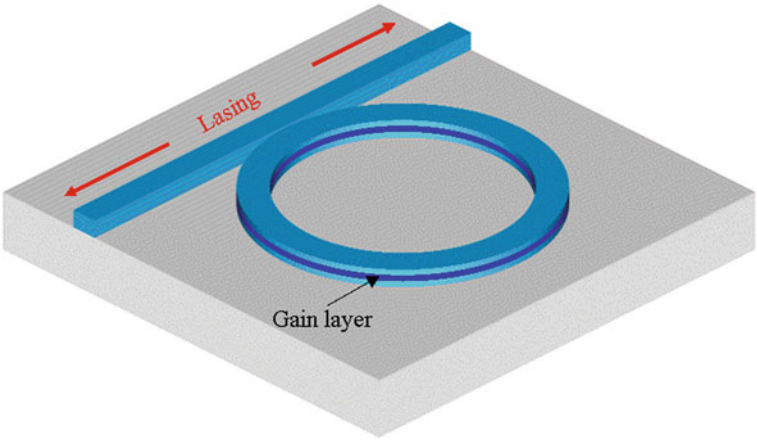


Fig. 8.18 Schematic of a micro ring based integrated laser

waveguide. Such device can be used as an amplifier or a laser depending on the cavity losses and the pumping level.

8.3.4.1 Microring Lasers

When sufficiently pumped, the structure illustrated in Fig. 8.18 can lase and emit radiation at the resonance frequencies of the micro ring. The lasing frequencies and the threshold lasing condition is given by:

$$\upsilon_{\mathrm{m}} = mc/2\pi R n_{\mathrm{eff}}, \tag{8.48a}$$

$$g_{\mathrm{th}} = \alpha + \ln(1-\kappa)/4\pi R \tag{8.48b}$$

where g_{th} is the threshold gain of the laser. The coupled waveguide serves as a natural output coupler which can extract light from the micro ring laser and channel it to any desired position on the chip.

Because of the clockwise–counterclockwise degeneracy of the structure it is expected that lasing will be established in both directions simultaneously and that a standing wave pattern will evolve in the micro ring laser.

Because of the multiple resonance wavelengths of the micro ring, lasing might occur in various wavelengths simultaneously. In addition, instability phenomena such as mode hopping may occur as well. This problem can be solved by reducing the radius of the micro ring to include a single, dominant, resonance in the gain bandwidth or by exploiting the Vernier effect and cascading rings with different radii.

8.3.4.2 Micro Ring Optical Amplifiers

A micro ring optical amplifier (MROA) is essentially a micro ring laser operated below threshold. An MROA can be realised using any of the configurations discussed above. The description of MROA relies on the FP modelling presented in Sect. 8.2.5 with the necessary modifications to account for the ring geometry. Applying the LATW approximation the static characteristics of the MROA can be analysed using using the modified carrier RE [41], where now the average photon densities for the signal photons and the amplified spontaneously emitted photons for MROA configurations of all-pass and add-drop (see Fig. 8.19 for relevant nomenclature) read [41] as:

All-Pass

$$S_{\mathrm{ASE,av}}^{\text{all-pass}} = \frac{2R_{\mathrm{sp}}n_{\mathrm{g}}}{gc}\left(\frac{\left(e^{gL}-1\right)}{gL}\frac{\left(1-\tau_1^2\right)}{1-\tau_1^2 e^{gL}} - 1\right), \tag{8.49a}$$

$$S_{\mathrm{signal,av}}^{\text{all-pass}} = \frac{\left(e^{gL}-1\right)}{gL}\frac{\kappa_1^2\left|E_{\mathrm{in}_1}\right|^2}{\left(1-\tau_1 e^{gL/2}\right)^2 + 4e^{gL/2}\tau_1\sin^2(\varphi/2)}, \tag{8.49b}$$

Add/Drop

$$S_{\mathrm{ASE,av}}^{\mathrm{add-drop}} = \frac{4R_{\mathrm{sp}}n_g}{gc} \times\left(\frac{\left(e^{gL/2}-1\right)}{gL}\frac{\left[\left(1-\tau_1^2\right)\left(1+\tau_2^2 e^{gL/2}\right)+\left(1-\tau_2^2\right)\left(1+\tau_1^2 e^{gL/2}\right)\right]}{1-\tau_1^2\tau_2^2 e^{gL}} - 1\right), \tag{8.50a}$$

$$S_{\mathrm{signal,av}}^{\mathrm{add-drop}} = \frac{\left(e^{gL}-1\right)}{gL}\frac{\kappa_1^2\left|E_{\mathrm{in}_2}\right|^2}{\left(1-\tau_2\tau_1 e^{gL/2}\right)^2 + 4e^{gL/2}\tau_2\tau_1\sin^2(\varphi/2)} \tag{8.50b}$$

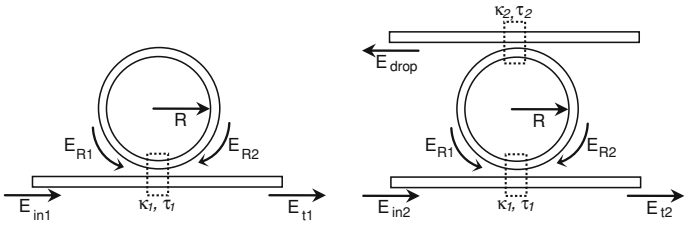


Fig. 8.19 Schematic of a micro-ring resonator in two configurations **a** the all-pass and **b** the add-drop

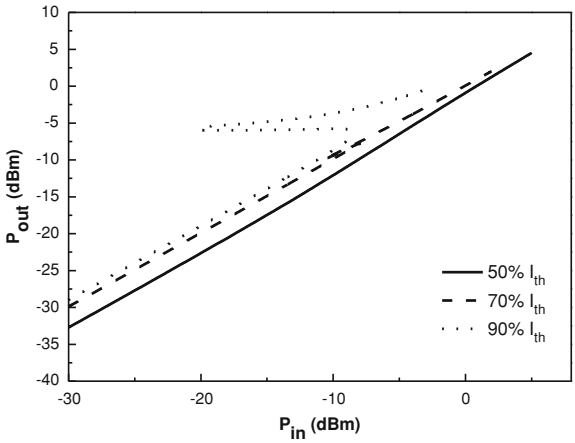


Fig. 8.20 Power Input–Output characteristics for MR with $\kappa = 0.4$ and $R = 20\,\mu\text{m}$ in all-pass configuration for various values of current. Detuning is $\Delta\lambda = 0.41\,\text{nm}$

The theoretical framework for MROA is completed with the expressions for the transmitted powers at the through and drop ports, for the all-pass and add-drop configurations (see also 8.39)

$$P_{t1} = P_{\text{in1}} \frac{\left(\tau_1 - e^{gL/2}\right)^2 + 4e^{gL/2}\tau_1 \sin^2(\varphi/2)}{\left(1 - \tau_1 e^{gL/2}\right)^2 + 4e^{gL/2}\tau_1 \sin^2(\varphi/2)}, \tag{8.51a}$$

$$P_{t2} = P_{\text{in1}} \frac{\left(\tau_1 - \tau_2 e^{gL/2}\right)^2 + 4e^{gL/2}\tau_1\tau_2 \sin^2(\varphi/2)}{\left(1 - \tau_1\tau_2 e^{gL/2}\right)^2 + 4e^{gL/2}\tau_1\tau_2 \sin^2(\varphi/2)}, \tag{8.51b}$$

$$P_{\text{drop}} = P_{\text{in1}} \frac{(\kappa_1\kappa_2)^2\, e^{gL/2}}{\left(1 - \tau_1\tau_2 e^{gL/2}\right)^2 + 4e^{gL/2}\tau_1\tau_2 \sin^2(\varphi/2)} \tag{8.51c}$$

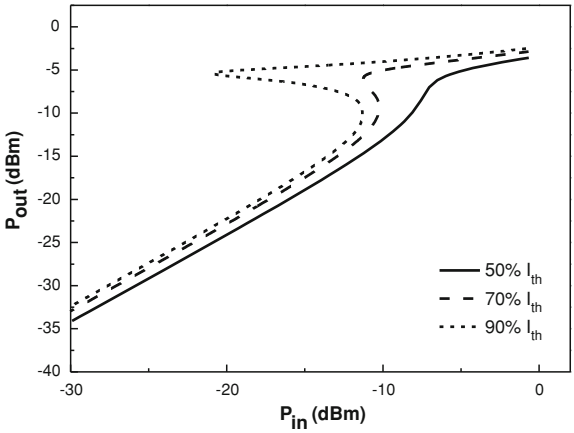


Fig. 8.21 Power Input–Output characteristics at the drop port for MR with $\kappa_1 = \kappa_2 = 0.5$ and $R = 20\,\mu$m in add–drop configuration for various values of current as indicated on the figure. Detuning is $\Delta\lambda = 0.33\,$nm

where P_{t1} is the transmitted power at the through port of the MR in the all-pass configuration, whereas P_{t2} and P_{drop} refer to the transmitted powers in the through and drop port, respectively, of the add-drop configuration. Equations (8.49–8.51) can be used to describe the nonlinear performance of MROAs. The input–output power characteristics for MROAs in all-pass and add-drop configuration are shown in Figs. 8.20, 8.21, respectively, for increasing values of current. In these figures the input signal is injected at a detuning of $\Delta\lambda = 0.41\,$nm and $\Delta\lambda = 0.33\,$nm with respect to the cavity resonance for Figs. 8.20, 8.21, respectively. The injected signal detuning along with the carrier-dependent refractive index (see 8.16) that varies with current, generate the bistable effects demonstrated in Figs. 8.20, 8.21.

8.4 Conclusion

This chapter aimed at presenting the basic elements of VC and micro ring thus underlining their potential for device applications.

The VC fundamental concepts were analysed and the design implications of these were explored. In particular, the importance of high reflective DBRs was outlined and the relevant theory presented. The analysis of VC was simplified with the help of the assumption of hard mirrors and the gain enhancement factor. The chapter focused on VCSOAs which can be studied in the context of RE and/or the FP method. Finally the SFM model was discussed for the polarisation dynamics of VCSELs.

The second part of the chapter dealt with MR resonators. The fundamental theory of MR was presented by means of analytic relations. The waveguiding properties are crucial for the design of MRs and hence were treated in some detail. Single and multi-ring configurations were analysed. Particular emphasis was put on MR structures in the

form of CROWs and SCISSORs as these provide efficient means for the tailoring of the spectral characteristics and dispersion. Finally the modelling and design of active MR structures (lasers and optical amplifiers) was presented.

Concluding it must be noted that this chapter serves as an introductory course, rather than a detailed account of all aspects of MRs and VCs as such an attempt would certainly extend the limits of a chapter. It is the authors' hope that this chapter will motivate and guide the interested reader to study further the exciting research fields of VC and MRs.

Acknowledgments Part of the work of DA was supported by the UK Engineering and Physical Science Research Council (EPSRC) under a visiting fellowship Grant No. EP/H00873X/1.

References

1. K. Iga, Surface emitting laser-its birth and generation of new optoelectronics field. IEEE J. Sel. Topics Quantum Electron. **6**(6), 1201–1215 (2000)
2. P.B. Hansen, G. Raybon, M.-D. Chien, U. Koren, B.I. Miller, M.G. Young, J.-M. Verdiell, C.A. Burrus, A 1.54 μm monolithic semiconductor ring laser: CW and mode-locked operation. IEEE Photon. Technol. Lett. **4**, 411–413 (1992)
3. C.W. Wilmsen, H. Temkin, L.A. Coldren (eds.), *Vertical-Cavity Surface-Emitting Lasers: Design, Fabrication, Characterization, and Applications, (Cambridge Studies in Modern Optics)* (Cambridge University Press, Cambridge), (New Ed edition), 12 Nov 2001
4. S.F. Yu, *Analysis and Design of Vertical Cavity Surface Emitting Lasers (Wiley Series in Lasers and Applications)* (Wiley-Blackwell, New York), 19 Sept 2003
5. J.S. Harris, T. O'Sullivan, T. Sarmiento, M.M. Lee, S. Vo, Emerging applications for vertical cavity surface emitting lasers. Semicond. Sci. Technol. **26**, 014010 (2011)
6. G.P. Agrawal, *Fiber-Optic Communication Systems*, 3rd edn. (Wiley-Interscience, New York, 2002
7. M. Silver, W.E. Booij, S. Malik, A. Galbraith, S. Uppal, P.F. McBrien, G.M. Berry, P.D. Ryder, S.J. Chandler, D.M. Atkin, R. Harding, R.M. Ash, Very wide temperature (−20 to 95°C) operation of an uncooled 2.5Gbit/s 1300 nm DFB laser. in *Proceedings of the 14th IEEE LEOS Annual Meeting, 1 and 2*, pp. 796–797, 2001
8. M. Kondow, T. Kitatani, S. Nakatsuka, M. Larson, K. Nakahara, Y. Yazawa, M. Okai, K. Uomi, GaInNAs: a novel material for long-wavelength semiconductor lasers. IEEE J. Sel. Topics Quantum Electron. **3**, 719–730 (1997)
9. V.M. Ustinov, A.E. Zhukov, GaAs-based long-wavelength lasers. Semicond. Sci. Technol. **15**, R41–R54 (2000)
10. N. Margalit, D. Babic, K. Streubel, R. Mirin, R. Naone, J. Bowers, and E. Hu, Submilliamp long wavelength vertical cavity lasers. Electron. Lett. **32**(18), 1675–1677 (1996)
11. A. Syrbu, V. Iakovlev, G. Suruceanu, A. Caliman, A. Rudra, A. Mircea, A. Mereuta, S.Tadeoni, C.-A. Berseth, M. Achtenhagen, J. Boucart, E. Kapon, 1.55 μm optically pumped wafer-fused tunable VCSELs with 32-nm tuning range. IEEE Photonics Technol. Lett. **16**(9), 1991–1993 (2004)
12. R. Shau, M. Ortsiefer, J. Rosskopf, G. Bohm, F. Kohler, and M.-C. Amann, Vertical-cavity surface-emitting laser diodes at 1.55 μm with large output power and high operation temperature. Electron. Lett. **37**, 1295 (2001)
13. M.-R. Park, O.-K. Kwon, W.-S. Han, K.-H. Lee, S.-J. Park, B.-S. Yoo, All-epitaxial InAlGaAs-InP VCSELs in the 1.3−1.6-μm wavelength range for CWDM band applications. IEEE Photonics Technol. Lett. **18**(16), 1717–1719 (2006)
14. P. Yeh, A. Yariv, *Optical Waves in Crystals: Propagation and Control of Laser Radiation* (Wiley, New York, 1984)

15. D.I. Babic, S.W. Corzine, Analytic expressions for the reflection delay, penetration depth, and absorptance of quarter-wave dielectric mirrors. IEEE J. Quantum Electron. **28**, 514–524 (1992)
16. S.W. Corzine, R.S. Geels, J.W. Scott, R.H. Yan, L.A. Coldren, Design of Fabry-Perot surface-emitting lasers with a periodic gain structure. IEEE J. Quantum Electron. **25**, 1513–1524 (1989)
17. J. Piprek, S. Bjorlin, J.E. Bowers, Design and analysis of vertical-cavity semiconductor optical amplifiers. IEEE J. Quantum Electron. **37**, 127–134 (2001)
18. M.J. Adams, J.V. Collins, I.D. Henning, Analysis of semiconductor laser optical amplifiers. IEE Proc. Pt J. **132**, 58–63 (1985)
19. P. Royo, R. Koda, L.A. Coldren, Vertical cavity semiconductor optical amplifiers: comparison of Fabry-Perot and rate equation approaches. IEEE J. Quantum Electron. **38**, 279–284 (2002)
20. B.R. Bennett, R.A. Soref, J.A. Del Alamo, Carrier induced change in refractive index of InP, GaAs and InGaAsP. IEEE J. Quantum Electron. **26**, 113–122 (1990)
21. M.J. Adams, H.J. Westalake, M.J. O' Mahony, I.D. Henning, A comparison of active and passive optical bistability in semiconductors. IEEE J. Quantum Electron. **21**, 1498–1504 (1985)
22. M.J. Adams, A. Hurtado, D. Labukhin, I.D. Henning, Nonlinear semiconductor lasers and amplifiers for all-optical information processing. Chaos, 20, 037102, (2010)
23. J. Martin-Regalado, F. Pratl, M. San Miguel, N.B. Abraham, Polarization properties of vertical-cavity surface-emitting lasers. IEEE J. Quantum Electron. **33**, 765–783 (1997)
24. M. San Miguel, Q. Feng, J.V. Moloney, Light-polarization dynamics in surface emitting semiconductor lasers. Phys. Rev. A **52**(2), 1728–1739 (1995)
25. A. Homayounfar, M.J. Adams, Locking bandwidth and birefringence effects for polarized optical injection in vertical-cavity surface-emitting lasers. Opt. Commun. **269**, 119–127 (2007)
26. R.K. Al-Seyab, D. Alexandropoulos, I.D. Henning and M.J. Adams, Instabilities in Spin-Polarized Vertical-Cavity Surface-Emitting Lasers. IEEE Photon. J. **3**, 799–809 (2011)
27. K.E. Chlouverakis, M.J. Adams, Stability maps of injection-locked laser diodes using the largest Lyapunov exponent. Opt. Commun. **216**, 405–412 (2003)
28. K.K. Lee, D.R. Lim, H.-C. Luan, A. Agarwal, J. Foresi, and L.C. Kimerling, Effect of size and roughness on light transmission in a SiOSiO2 waveguide: Experiments and model. Appl. Phys. Lett. **77**, 1617–1619 (2000)
29. K. Vahala, Optical microcavities. Nature **424**, 839–846 (2003)
30. J. Scheuer, A. Yariv, Fabrication and Characterization of low-loss polymeric waveguides and micro-resonators. J. Euro. Opt. Soc. Rapid Pub. **1**, 06007 (2006)
31. H.A. Haus, Waves and Fields in Optoelectronics (Prentice-Hall, Englewood Cliffs, 1983)
32. M. Heiblum, J.H. Harris, Analysis of curved optical waveguides by conformal transformation. IEEE J. Quantum Electron. **11**, 75–83 (1975)
33. C.K. Madsen, J.H. Zhao, *Optical Filter Design and Analysis: A Signal Processing Approach*, 1st edn. (Wiley-Interscience, New York, 1999)
34. J. Scheuer, G.T. Paloczi, A. Yariv, All-optically tunable wavelength-selective reflector consisting of coupled polymeric microring resonators. Appl. Phys. Lett. **87**, 251102 (2005)
35. J. Scheuer, G.T. Paloczi, J.K.S. Poon, A. Yariv, Coupled resonator optical waveguides: towards slowing and storing of light. Opt. Photon. News **16**, 36 (2005)
36. J. Heebner et al., Distributed and localized feedback in microresonator sequences for linear and nonlinear optics. J. Opt. Soc. Am. B. **21**, 1818–1832 (2004)
37. O. Weiss, J. Scheuer, Side coupled adjacent resonators CROW—formation of mid-band zero group velocity. Opt. Express **17**, 14817 (2009)
38. J.B. Khurgin, R.S. Tucker (ed.), *Slow Light: Science and Applications* (CRC Press, Boca Raton, 2008)
39. J.K.S. Poon, J. Scheuer, S. Mookherjea, G.T. Paloczi, Y. Huang, A. Yariv, Matrix analysis of coupled-resonator optical waveguides. Opt. Express **12**, 90 (2004)
40. A. Yariv, Y. Xu, R.K. Lee, and A. Scherer, Coupled-resonator optical waveguide: a proposal and analysis. Opt. Lett. **24**, 711–713 (1999)
41. D. Alexandropoulos, H. Simos, M.J. Adams and D.Syvridis, Optical bistability in active semiconductor micro-ring resonators. IEEE J. Sel. Top. Quantum Electron. **14**, 918–926 (2008)

Index

X. Marie and N. Balkan (eds.), *Semiconductor Modeling Techniques*,
Springer Series in Materials Science 159, DOI: 10.1007/978-3-642-27512-8,