High-Speed Electronics and Optoelectronics

Devices and Circuits

Sheila Prasad
Hermann Schumacher
and Anand Gopinath

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521862837

This page intentionally left blank

High-Speed Electronics and Optoelectronics

This authoritative account of electronic and optoelectronic devices operating at frequencies greater than 1 GHz covers the concepts and fundamental principles of operation, and, uniquely, their circuit applications too.

Key features include:

- a comprehensive coverage of electron devices, such as MESFET, HEMT, RF MOSFET, BJT and HBT, and their models;
- discussions of semiconductor devices fabricated in a variety of material systems, such as Si, III–V compound semiconductors and SiGe;
- a description of light-emitting diodes, semiconductor lasers and photodetectors;
- an executive summary at the beginning of each chapter;
- plentiful real-world examples; and
- end-of-chapter problems to test understanding of the material covered.

From crystal structure to atomic bonding, recombination and radiation in semiconductors to p–n junctions and heterojunctions, a wide range of critical topics is covered. Moreover, a chapter on analogue circuit applications provides an introduction to scattering parameter theory, followed by descriptions of different types of amplifier and oscillator utilising HBTs and HEMTs. Optimisation algorithms, such as simulated annealing and neural network applications, and parameter extraction of electronic device equivalent circuit models are also discussed. Graduate students in electrical engineering, industry professionals and researchers will all find this a valuable resource.

Sheila Prasad is Professor Emeritus in the Electrical and Computer Engineering Department at Northeastern University. Her current research interests include microwave and high-speed semiconductor devices and circuits, and optoelectronic circuits. She has co-authored the book *Fundamental Electromagnetic Theory and Applications* with Ronold W. P. King and has authored over 130 journal and conference publications.

Hermann Schumacher is Professor and Director of the Competence Center on Integrated Circuits in Communications, Institute of Electron Devices and Circuits, University of Ulm. He is also the Director of the International Master Program on Communications Technology at the University of Ulm, and has authored over 150 journal and conference publications.

Anand Gopinath is Professor in the Department of Electrical and Computer Engineering at the University of Minnesota. He is Life Fellow of the IEEE, Fellow of the OSA and Fellow of IET (UK). His research is in the areas of RF/microwave and optical semiconductor devices, integrated optics and metamaterials.

High-Speed Electronics and Optoelectronics: Devices and Circuits

SHEILA PRASAD
Northeastern University, Boston

HERMANN SCHUMACHER
University of Ulm, Germany

ANAND GOPINATH
University of Minnesota, Minneapolis

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521862837

First published in print format 2009

ISBN-13 978-0-511-57982-0 eBook (EBL)

ISBN-13 978-0-521-86283-7 Hardback

Contents

Preface

Starting from the development of transistor technology to laser technology, the field of solid state devices and their circuit applications has advanced rapidly. The silicon bipolar junction transistor was first applied to low frequency circuits. The subsequent advances in materials science made it possible to fabricate compound semiconductor transistors capable of operating at microwave frequencies and high speeds. This presented the capability of applications in both analogue and digital circuits. At the same time, the wide choice of high performance semiconductor materials also enabled the development of optoelectronic devices such as lasers and light-emitting diodes. The communications industry continues to grow and diversify, thus necessitating the design of circuits which will satisfy the requirements of mobile telephones which are becoming more and more sophisticated in their performance. Circuit design has applications in other areas such as optical communications.

This book focusses on high-speed electronics and optoelectronics where the devices operate at frequencies ≥ 1 GHz. It is presented in two parts with devices being discussed in the first part and the circuit applications in the second part. In Part One, semiconductor devices fabricated in a variety of material systems – Si, III–V compound semiconductors and SiGe – are presented. We discuss the concepts and the fundamental principles of operation. We do not attempt to present the latest results as they will already be obsolete by the time the book is published. It is assumed that the reader has had a course in fundamental solid state physics.

Chapter 1 reviews semiconductor materials and physics. For the reader who is familiar with the topics, this chapter will be a brief review. If not, the reader can go to the references section to get a detailed coverage of the topics. Semiconductor materials are described followed by brief discussions of crystal structure and bonding. The section on quantum mechanics is intended to present only the important concepts and is not a comprehensive treatment of the subject. Semiconductor properties are described followed by types of semiconductors. Semiconductor junctions are treated in detail as they are the basis of the devices to be treated in subsequent chapters.

Chapter 2 presents high-frequency/high-speed electronic devices starting with the MESFET, which was the first transistor to operate at microwave frequencies. The development of the high electron mobility transistor (HEMT) represented a major advance in technology and is presented here in detail. The recent application of MOSFETs to radio frequency has been successful and the properties are covered in detail. Finally, bipolar and heterojunction bipolar transistors (HBTs) are described. Models for the transistors are presented and their method of implementation is described.

Chapter 3 presents the optimisation and parameter extraction of the circuit models of the electronic devices. The simulated annealing algorithm is discussed followed by the application of neural networks to circuit modelling. The genetic algorithm is defined and its application to optimisation is shown. Parameter extraction methods are given for circuit models using semi-analytical methods and basic expressions are derived.

Chapter 4 deals with various optical sources such as light-emitting diodes and lasers, giving details of their physical properties and their modes of operation. The discussion of emitters is followed by an extensive coverage of a variety of photodetectors.

In Part Two of the book, we discuss analogue circuits at the gate level. We will assume that the reader has a background (at the undergraduate level) in fundamental analogue circuit theory. Chapter 5 (Part Two of the book) deals with the components of high-speed analogue circuits. After a review of scattering parameter theory, the power and noise relations for two-port networks are discussed. Transistor amplifiers are covered in detail, showing the application of the devices described in Chapter 2. This is followed by a discussion of oscillators and mixers for high-speed circuits. Important passive components of high-speed circuits complete this chapter.

We have a layered approach to each chapter in the book. There is an executive summary at the beginning of each chapter. This will make the book valuable also for technical managers who may not want to go through the chapter content in detail. We have extensive problems at the end of each chapter, which will give the student applications of the theory. This book should be useful to research engineers and graduate students. Results from various research papers are presented, many of which are only available in journals which are referenced extensively. However, the reader need not go to the original papers as the results are given in sufficient detail to give a good understanding of the material.

Acknowledgements

Sheila Prasad

I would like to express my gratitude to Professor Clifton G. Fonstad, Jr, at the Massachusetts Institute of Technology. My long collaboration with him started with the first sabbatical leave at MIT when I worked in his group. It initiated my work on HBTs at microwave frequencies and the continued support he provided to me and my students in his laboratory resulted in this successful research. I acknowledge my colleague at Northeastern University, Dr Michael Vai (now at MIT Lincoln Laboratory), with whom I performed research on optimisation and modelling techniques. Many students worked with me on various aspects of the research reported in this book. I would particularly like to acknowledge the work of Dr Bin Li whose research results continue to be cited in publications. I would also like to acknowledge my student Kofi Deh for his help with the figures and manuscript editing. Dr Henry Choy and Dr Wojtek Giziewicz, both of whom were students at MIT, gave me invaluable suggestions for the book material and also helped me with MATLAB, graphics programmes and Latex when needed. I acknowledge my colleague at Northeastern University, Professor Jeff Hopwood (now at Tufts University), with whom I had many useful discussions about the content of the book. It has been a great experience to work with both of my co-authors. Last but not least, I would like to thank my husband, Fred Hinchey, for his great patience and support in the course of this book project.

Hermann Schumacher

I gratefully acknowledge the valuable assistance of Dr Andreas Trasser, Dr Wolfgang Haag and Ms Ursula Winter in proofreading the original manuscript. Their helpful suggestions had a significant impact. Most importantly, I am eternally indebted to my wife Christiane. Without her patience and loving care, this book would never have materialised.

Anand Gopinath

I acknowledge the valuable discussions on lasers and photodiodes with my past and present graduate students including Ross Schermer, Prakash Koonath, William

Berglund, Jaesang Oh, Kang-Hyun Baek, Klein Johnson and others. I am grateful to my wife Marian for her patience and support while this book was being written.

Joint acknowledgements

Finally, we would like to thank our publisher Dr Julie Lancashire for her patience and her guidance throughout this book project. We also thank Dr Phil Meyler who encouraged us to initiate this project by submitting a book proposal which was accepted by Cambridge University Press. Thanks are also due to Ms Sarah Matthews who has been very helpful in the last and most difficult stages of the project.

Part One

Devices

1 Review of semiconductor materials and physics

1.1 Executive summary

Semiconductor devices are fabricated using specific materials that offer the desired physical properties. There are three classes of solid state materials: insulators, semiconductors and conductors. This distinction is based on the electrical conductivity of these materials with insulators having the lowest and conductors having the highest conductivity. Semiconductors fall in between and their conductivity is affected by several factors such as temperature, the incidence of light, the application of a magnetic field and impurities. This versatility makes semiconductors very important in electronics and optoelectronics applications.

Semiconductors themselves are divided into two classes: elemental and compound. Each type has distinctive physical properties which are exploited in device design. Typical elemental semiconductor device materials are silicon and germanium; examples of compound semiconductors are GaAs, InP, AlGaAs and SiGe. The single crystal structure of these materials is that of a periodic lattice and this determines the properties of the semiconductors. Silicon has the diamond crystal structure and the compound semiconductors have the zincblende lattice structure. The bonding between atoms in a crystal of the semiconductors is termed *covalent bonding*, where electrons are shared between atoms. Fundamental principles of quantum mechanics are applied to determine the energy band structure of the semiconductor.

The basic device physics involves the description of the energy band structure, the density of states, the carrier concentration and the definition of donors and acceptors. Semiconductors are categorised as direct or indirect depending on the bandgap. The absorption mechanism is described and radiation and recombination processes important to device performance are detailed. The two carrier transport processes are drift and diffusion. The currents due to these transport processes are expressed in terms of the applied electric field, the carrier mobility and the carrier concentration. The junction formed by p-type semiconductor (excess holes) and n-type semiconductor (excess electrons) is described and the characteristics of such a junction are given. The important Schottky diode, a junction formed by a metal and a semiconductor layer (n-doped in this case) is characterised.

Heterostructures formed by dissimilar semiconductors are important in device design. The properties of heterojunctions of semiconductor materials are presented. Silicon–germanium heterojunctions are of particular interest as high performance electronic

Table 1.1 Portion of the periodic table showing semiconductor material elements

Period	Group III	Group IV	Group V
2	B Boron	C Carbon	N Nitrogen
3	Al Aluminium	Si Silicon	P Phosphorus
4	Ga Gallium	Ge Germanium	As Arsenic
5	In Indium	Sn Tin	Sb Antimony

Table 1.2 Elemental and binary compound semiconductors

Elements	IV–IV Binary compounds	III–V Binary compounds
Si Silicon	SiC Silicon carbide	AlAs Aluminium arsenide
Ge Germanium	SiGe Silicon germanium	AlP Aluminium phosphide
		AlSb Aluminium antimonide
		BN Boron nitride
		GaAs Gallium arsenide
		GaN Gallium nitride
		GaSb Gallium antimonide
		InAs Indium arsenide
		InP Indium phosphide
		InSb Indium antimonide

Table 1.3 Ternary and quaternary semiconductors

Ternary compounds	Quaternary compounds
$Al_xGa_{1-x}As$ Aluminium gallium arsenide	$Al_xGa_{1-x}As_ySb_{1-y}$ Aluminium gallium arsenic antimonide
$GaAs_{1-x}P_x$ Gallium arsenic phosphide	$Ga_xIn_{1-x}As_{1-y}P_y$ Gallium indium arsenic phosphide

devices have been designed using this material alloy. This chapter gives a detailed discussion of these heterojunctions.

1.2 Semiconductor materials

Materials used for semiconductors fall into two categories: elemental semiconductors and compound semiconductors. Table 1.1 shows the section in the periodic table which has the semiconductor elements and Table 1.2 lists examples for elemental and binary compound semiconductors. Some ternary and quaternary semiconductors are listed in Table 1.3.

1.3 Types of solids

There are three types of solids: crystalline, polycrystalline and amorphous. The arrangement of atoms is periodic in three dimensions in a crystalline solid with forces binding the atoms together. This periodicity exists over the entire crystal and it will appear the same regardless of the region where the crystal is viewed. If the periodicity of the atoms occurs over a small region of the solid and changes in different regions of the solid, the solid is termed to be *polycrystalline*. Atoms in amorphous solids exhibit no periodicity. Figure 1.1 shows the three different types of solids.

1.4 Crystal structure

Semiconductor materials such as Si, Ge and GaAs that are to be used for devices are crystalline, that is, a single crystal. This periodic arrangement of atoms in a crystal is termed a *lattice* and the distance between the atoms is the *lattice constant*. The unit cell is a fundamental unit in the crystal and a repetition of the unit cell generates the entire lattice. The unit cell is not unique and can be chosen in various ways as shown in Figure 1.2(a). This is a two-dimensional representation of the crystal lattice. The entire lattice can be constructed by translations of any of the three unit cells in two coordinate directions. The primitive unit cell is the smallest unit cell. A generalised primitive three-dimensional unit cell is shown in Figure 1.2(b). The coordinate directions are **a,b,c**. In cubic structures, these would be the rectangular coordinates. The basic cubic crystal structures are (a) the simple cubic, (b) the body-centred cubic and

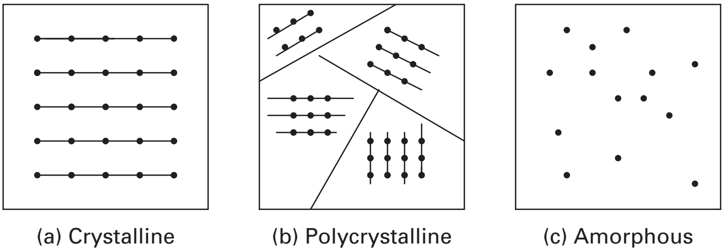
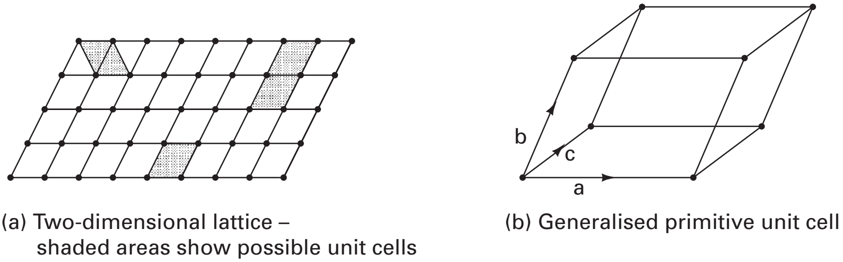

(a) Crystalline (b) Polycrystalline (c) Amorphous

Fig. 1.1 Schematic arrangement of atoms in solids.



(a) Two-dimensional lattice – shaded areas show possible unit cells (b) Generalised primitive unit cell

Fig. 1.2 Unit cells.

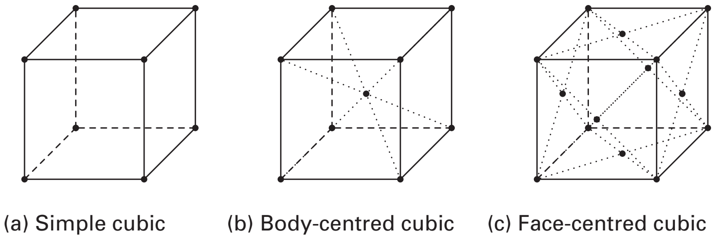


Fig. 1.3 Types of cubic lattices.

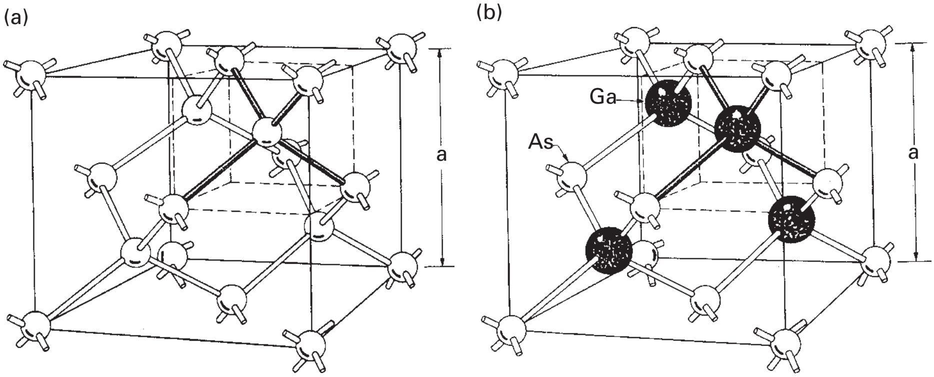


Fig. 1.4 (a) Diamond lattice and (b) Zincblende lattice. (S. M. Sze, *Semiconductor Devices: Physics and Technology*, John Wiley & Sons, 1985). Reprinted with permission of John Wiley & Sons, Inc.

(c) the face-centred cubic shown in Figure 1.3. The simple cubic lattice has an atom at each corner of the cube, where the length of a side of the cube is a, the lattice constant. The body-centred cubic lattice (BCC) has an additional atom in the centre of the cube and the face-centred cubic lattice (FCC) has an additional atom in the centre of each face of the cube. The two most important semiconductor crystal structures are the diamond lattice structure and the zincblende structure. Silicon and germanium have the diamond lattice structure and most of the binary compound semiconductors such as GaAs have the zincblende lattice structure. The only difference between the diamond and the zincblende structures is that the latter has two different types of atoms as seen in Figure 1.4. The diamond structure consists of two inter-penetrating FCC sublattices of atoms. The second FCC cube is shifted by one-fourth of the body diagonal, which is the longest diagonal. In the zincblende structure of GaAs, one sublattice has gallium atoms and the other has arsenic atoms.

1.5 Crystal directions and planes

Crystals are of finite size and hence have surfaces. It is necessary to define the planes at the crystal surfaces and the crystallographic directions, both of which determine the

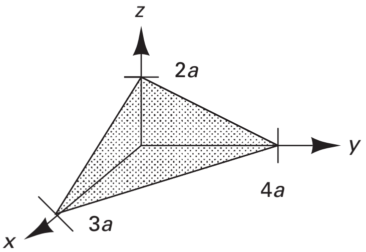


Fig. 1.5 Representation of plane with Miller indices [6, 5, 8].

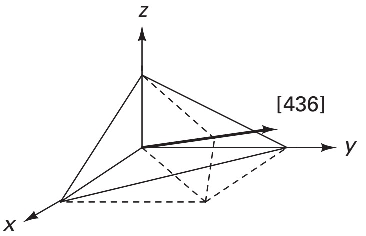


Fig. 1.6 Representation of direction with Miller indices [6, 5, 8].

properties of semiconductor devices. The rectangular coordinate system defines the cubic crystal and the plane surfaces and directions are described by a set of indices called the *Miller indices*. Planes are described by the indices (h,k,l) and the directions perpendicular to these planes are described by the same indices [hkl].

Example: Find the Miller indices of the plane which makes intercepts 3*a*,4*a*,2*a* along the coordinate axes in a cubic crystal, where *a* is the lattice constant. Draw the direction vector with the same Miller indices.

Solution: The intercepts are 3, 4 and 2. The reciprocals are 1/3, 1/4 and 1/2. Multiplication by the lowest common denominator, which is 12, yields (4,3,6). These are the Miller indices which define the plane shown in Figure 1.5. It can be shown that parallel planes are described by the same Miller indices.

The Miller indices of the direction are given as [436]. The intercepts on the three coordinate axes are 3, 4 and 2. The direction vector is drawn and seen to be perpendicular to the planes shown in Figure 1.6.

The basic planes in cubic crystals are shown in Figure 1.7. It is also important to describe specific directions in a crystal in addition to the planes. As in the case of the crystal plane, a crystal direction is also described by three integers which are the components of a vector drawn in the particular crystal direction. The crystal planes and directions of most interest are shown in Figure 1.8. The [hkl] direction is perpendicular to the (hkl) plane.

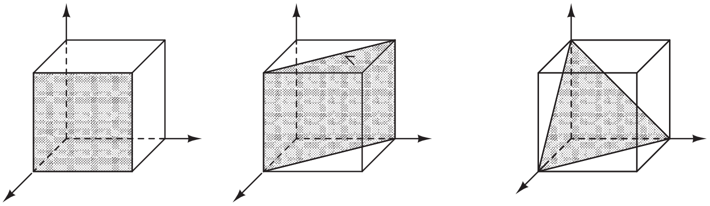
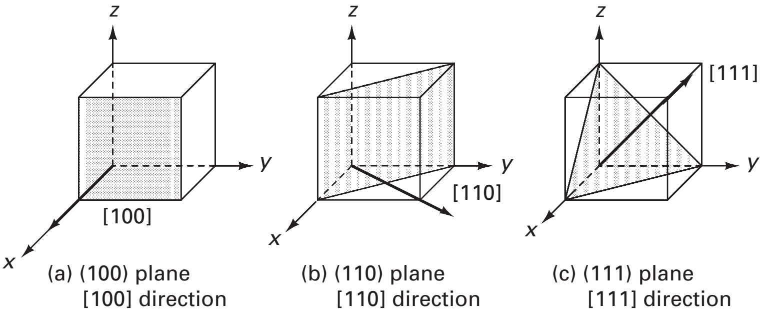


Fig. 1.7 Basic crystal planes.



(a) (100) plane [100] direction
(b) (110) plane [110] direction
(c) (111) plane [111] direction

Fig. 1.8 Important crystal planes and directions.

1.6 Atomic bonding

Atoms are held together by bonding forces to form solids. When the attractive and repulsive forces are equivalent, the atoms are in equilibrium and maintain the spacing characterised by the lattice constant, a. There are different bonding classifications which are described by the dominant force of attraction. When one of the atoms gives up an electron in the outer shell to another atom, positive and negative ions are produced. There is a Coulomb interaction force of attraction between them. This is termed *ionic bonding*. At equilibrium, the forces of attraction and repulsion are equivalent. Sodium chloride (NaCl) and Potassium chloride (KCl) are examples of ionic bonding after the formation of the Na^+ and Cl^- ions.

1.6.1 Covalent bonding

This type of bonding results when electrons are shared by neighbouring atoms. The hydrogen atom is the simplest example of covalent bonding. Each of the two electrons bonds with the other to complete the lowest energy shell as shown in Figure 1.9.

Each atom in a diamond or zincblende lattice has four nearest neighbours. Each atom has four electrons in the outer orbit. These are the valence electrons and each atom shares these valence electrons with its four neighbours. The interaction between the shared electrons results in bonding forces which are quantum mechanical in nature. In other words, each electron pair constitutes a covalent bond. Elements in group IV such as Si and Ge have four valence electrons as shown in the references [8, 14, 15]. These are available for bonding as seen in Figure 1.10. Compound semiconductors such as

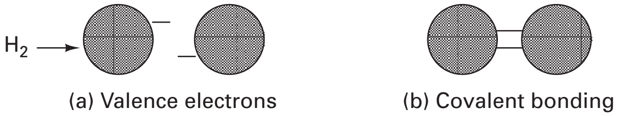


Fig. 1.9 Covalent bonding in hydrogen.

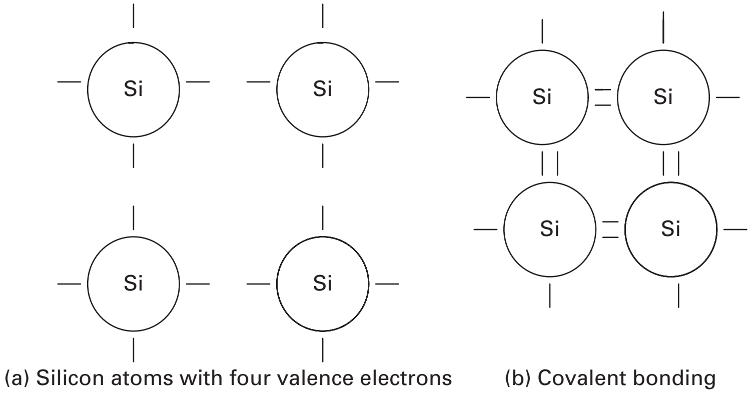


Fig. 1.10 Covalent bonding in silicon.

GaAs exhibit both covalent as well as ionic bonding. This is due to the fact that Ga and As occur in two different groups in the periodic table and hence there is a transfer of charge resulting in some ionic bonding.

1.7 Atomic physics

The theories of atomic physics were based on experimental observations. These theories subsequently explained the experiments and led to the understanding of atoms in matter.

1.7.1 The photoelectric effect

The measurements of Planck on a heated sample of material indicated that energy is radiated in discrete units called *quanta* as shown in Equation (1.1).

$$E = h\nu, \tag{1.1}$$

where h (Planck's constant) $= 6.63 \times 10^{-34}$ J · s and ν is the frequency of the radiation. Heinrich Hertz discovered the photoelectric effect in 1887. The experiments performed by Philipp Lenard, a former student of Hertz, showed that if light shines on a metal surface in vacuum, some of the electrons receive enough energy so that they are emitted from the surface into the vacuum. They were interpreted by Albert Einstein, who received the Nobel Prize for his work in 1921. This is termed the *photoelectric effect* and the maximum energy is a function of the frequency of the incident light. The quantised units of light energy are called *photons*.

Based on further experimental observations of Davisson and Germer (USA) and Thompson (UK) on the diffraction of electrons by the atoms in a crystal, de Broglie related the wavelength of a particle of momentum $p = mv$, where m is the mass of the particle as seen in Equation (1.2):

$$\lambda = \frac{h}{p} = \frac{h}{mv}. \tag{1.2}$$

1.7.2 The Bohr model of the atom

A model of the atom was first proposed by Bohr. In his model, the electrons move in stable circular orbits about the nucleus and the electron may move to an orbit of higher or lower energy. The electron would either gain energy or lose energy by the absorption or emission of a photon of energy $h\nu$. Bohr further proposed that the angular momentum of the electron moving in a circular orbit was an integral multiple of Planck's constant as seen in Equation (1.3).

$$p_\theta = \frac{nh}{2\pi} = n\hbar, \;\; n = 1, 2, 3, \ldots \tag{1.3}$$

The hydrogen atom with one electron and the nucleus illustrates this concept in a simple manner as seen in Figure 1.11.

Assuming that the electron of mass m rotates in a stable orbit of radius r with velocity v, the angular momentum is written in Equation (1.4):

$$p_\theta = mvr = n\hbar. \tag{1.4}$$

The electrostatic force between the charge on the nucleus and the charge on the electron must be equal to the centripetal force for the electron to remain in stable orbits. This yields the expression in Equation (1.5) for the energy of the electron [15]:

$$E_\mathrm{n} = -\frac{mq^4}{2(4\pi\epsilon_0)^2 n^2 \hbar^2}. \tag{1.5}$$

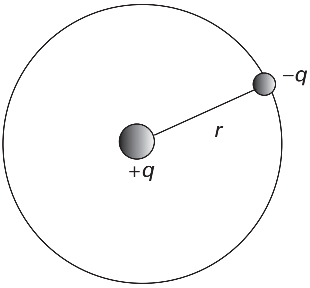


Fig. 1.11 Bohr model of the hydrogen atom.

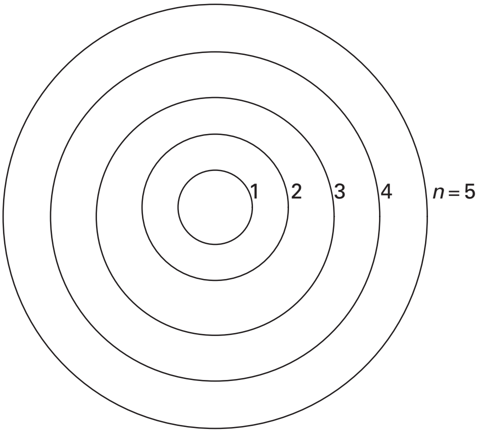


Fig. 1.12 Electron orbits in Bohr model (not to scale).

The electron orbits in the Bohr model are shown in Figure 1.12.

1.8 The de Broglie relation

The initial theoretical and experimental results of Planck, Einstein and Bohr laid the foundation for the development of quantum mechanics. It was de Broglie, however, who first postulated that if waves were seen to behave as particles then it could be that particles might behave like waves.

In the Bohr formulation, the electron which travels in a circular orbit of radius r is assumed to behave like a wave with a wavelength λ. It travels in a circular path equal in length to the circumference $2\pi r$, which will be an integral number of wavelengths so that

$$n\lambda = 2\pi r. \tag{1.6}$$

The Bohr formulation yielded the linear velocity of the electron to be

$$v = \frac{q^2}{4\pi\epsilon_0 n\hbar}. \tag{1.7}$$

Using this velocity relation, the wavelength can be written as

$$\lambda = \frac{h}{mv} = \frac{h}{p}, \tag{1.8}$$

where p is the linear momentum of the electron. Thus, de Broglie postulated that the relationship between the wavelength and the linear momentum p of a particle is given by Equation (1.2).

$$p = \frac{h}{\lambda} = \frac{h}{2\pi}\frac{2\pi}{\lambda} = \hbar k. \tag{1.9}$$

This is the *de Broglie relationship*. For free electrons, the energy–momentum relationship is as follows:

$$E = \frac{mv^2}{2} = \frac{p^2}{2m}; \; p = \sqrt{2mE}. \tag{1.10}$$

Hence, the experiments of Davisson and Germer and of Thompson were verified by the de Broglie relationship.

1.9 Quantum mechanics

Newtonian mechanics can be used to describe physical behaviour that is macroscopic. Typical examples of this are planetary motion, the classical electromagnetic fields and fluid motion. The motion of electrons and the interaction of electrons in atoms in semiconductor materials cannot, however, be described thus since we are dealing with microscopic behaviour. This physical behaviour on the atomic scale can only be described by quantum mechanics rather than Newtonian mechanics. Quantum or wave mechanics had as its basis the physical understanding developed by Planck and de Broglie. The classical laws of the conservation of energy, momentum and angular momentum are also assumed to be valid in quantum mechanics. Hence, the physics involved in the interaction between atoms can be described mathematically by quantum mechanics.

1.9.1 Probability and the uncertainty principle

When the motion of the particle is microscopic, the parameters cannot be described exactly but rather in terms of average (expectation) values. Hence we have, for example, the expectation values of position, momentum and energy of an electron. So, we have a probabilistic rather than an exact description of the particle behaviour. There is, thus, an inherent uncertainty in the position and momentum of the particle. This was formulated by Heisenberg and is termed the *Heisenberg uncertainty principle*. The uncertainty in the measurement of the position and momentum of particle motion is given as

$$(\Delta x)(\Delta p_{\mathrm{x}}) \geq \hbar. \tag{1.11}$$

The uncertainty in energy is related to the time at which the energy was measured and is given by

$$(\Delta E)(\Delta t) \geq \hbar. \tag{1.12}$$

These equations show that the simultaneous measurements of position and momentum on the one hand and energy and time on the other hand cannot be performed with arbitrary accuracy.

It follows that we can only determine the probability of finding an electron in a certain position or having a certain momentum. This leads to the definition of a probability density function. The probability of finding a particle in a range, say, from x to $x + dx$ is given by

Table 1.4 Classical variables and quantum operators

Classical variable	Quantum operator
x	x
$f(x)$	$f(x)$
Momentum $p(x)$	$\frac{\hbar}{j}\frac{\partial}{\partial x}$
Kinetic energy $\frac{p^2}{2m}$	$\frac{-\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$
Potential energy V	V
Total energy E	$\frac{-\hbar}{j}\frac{\partial}{\partial t}$

$$\int_{-\infty}^{\infty} P(x)dx = 1, \tag{1.13}$$

where $P(x)$ is a normalised function. The average value of a function x is defined as

$$\langle f(x)\rangle = \int_{-\infty}^{\infty} f(x)P(x)dx = 1. \tag{1.14}$$

The correspondence between classical and quantum mechanical quantities is shown in Table 1.4.

The basic principles of quantum mechanics will now be reviewed. Each particle in a physical system is described by a wave function $\Psi(x, y, z, t)$. The function and its space derivatives are continuous, finite and single-valued.

The probability of finding a particle with wave function Ψ in the volume *dxdydz* is $\Psi^*\Psi dxdydz$. Then we have the following definition for three-dimensional space:

$$\int_{-\infty}^{\infty} \Psi^*\Psi dxdydz = 1. \tag{1.15}$$

The expectation value of any physical quantity X can be written as

$$< X > = \int_{-\infty}^{\infty} \Psi^* X_{\text{oper}}\Psi dxdydz, \tag{1.16}$$

where X_{oper} is the operator corresponding to the variable X.

The classical equation for energy conservation is Kinetic energy + Potential energy = Total energy:

$$\frac{p^2}{2m} + V = E. \tag{1.17}$$

1.9.2 The wave equation

We obtain the quantum mechanical energy equation by substituting the corresponding operators which operate on the one-dimensional wave function $\Psi(x, t)$:

$$\frac{-\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + V(x)\Psi(x,t) = E\Psi(x,t) = \frac{-\hbar}{j}\frac{\partial\Psi(x,t)}{\partial t}. \tag{1.18}$$

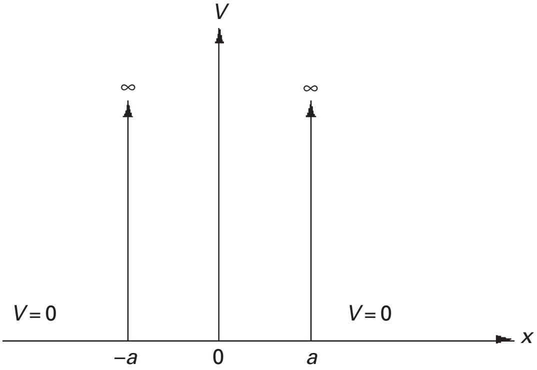


Fig. 1.13 Infinite potential well, width = $2a$.

This is the one-dimensional Schrödinger wave equation. The three-dimensional wave equation is

$$\frac{-\hbar^2}{2m}\nabla^2\Psi + V(x)\Psi = E\Psi = \frac{-\hbar}{j}\frac{\partial\Psi}{\partial t}. \tag{1.19}$$

The wave equation is applied to the solution of various physical problems. The problem of the infinite potential well provides an understanding of the method of solution and an insight into the discrete energies of a single electron [14, 15].

This basic physical concept is important since quantum wells can be fabricated using semiconductor structures for devices. A general solution of the one-dimensional wave equation can be written as follows:

$$\Psi(x,t) = \psi(x)\exp\left(\frac{-jEt}{\hbar}\right). \tag{1.20}$$

We consider the infinite quantum well of width $2a$ with zero potential outside the well as shown in Figure 1.13.

On solving the one-dimensional wave function, we obtain n solutions and the discrete energy levels are given by [14, 15],

$$E_\text{n} = \frac{\pi^2\hbar^2 n^2}{8m_0 a^2}, \tag{1.21}$$

where m_0 is the rest mass of the electron and a is the lattice constant of the crystal. The one-dimensional problem of a particle in a finite potential well can also be solved and the allowed energies of the particle determined [10]. The phenomenon of tunnelling wherein an electron with energy E tunnels through a potential barrier with barrier height V_0 greater than E is also explained by quantum mechanics. Classically, the electron would not be able to show this behaviour. If we have a potential barrier of width a, the one-dimensional Schrödinger equation can be solved in the three regions I, II and III as

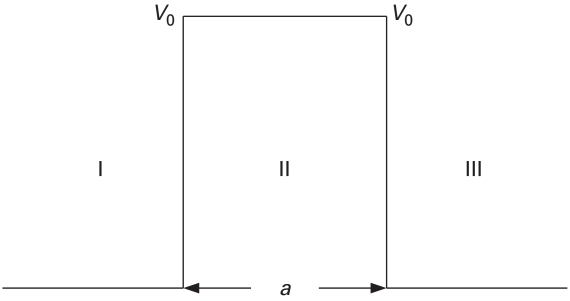


Fig. 1.14 Potential barrier.

shown in Figure 1.14. There are three regions for the problem. Regions I and III have zero potential. Say region II has a potential V_0, then the solutions in the three regions are given by:

$$Region\ \ I\colon \psi(x) = A\exp(jkx) + B\exp(-jkx);\, k^2 = \frac{2mE}{\hbar^2} \tag{1.22}$$

$$Region\ \ II\colon \psi(x) = C\exp(-\alpha x) + D\exp(+\alpha x);\ \ \alpha^2 = \frac{2m(V_0 - E)}{\hbar^2} \tag{1.23}$$

$$Region\ \ III\colon \psi(x) = F\exp(jkx);\ \ k^2 = \frac{2mE}{\hbar^2}. \tag{1.24}$$

Using the conditions that the wave function and its derivatives are continuous at the boundaries, $x = 0$ and $x = a$, the tunnelling probability is of the form:

$$T = \left|\frac{F}{A}\right|^2 = \frac{4}{4\cos h^2(\alpha d) + \left(\frac{\alpha}{k} - \frac{k}{\alpha}\right)^2 \sin h^2(\alpha d)}. \tag{1.25}$$

Boundary conditions are matched at the two boundaries and T, the tunnelling probability is determined.

The method of solution is the same regardless of the shape of the barrier. Triangular and trapezoidal barriers have a simple geometry and hence give us exact solutions. When the barriers are of arbitrary shape, the tunnelling probability is solved using the Wentzel–Kramers–Brillouin (WKB) approximation:

$$T \cong \exp\left[-2\int_{d_1}^{d_2} \mid f(x) \mid dx\right] \tag{1.26}$$

with

$$f(x) = \frac{2m_0}{\hbar^2}[V(x) - E], \tag{1.27}$$

where $V(x)$ is the arbitrary potential. The limits of the integral d_1 to d_2 represent the classically forbidden region, where the potential energy is larger than the total particle energy.

1.10 Statistical mechanics

1.10.1 The free electron

When the three-dimensional Schrödinger equation is solved, the general solution gives the wave function for the electron in motion in a region of zero potential. The behaviour of electrons in semiconductor crystals can be assumed to be like that of so-called *free electrons* under certain conditions, hence the importance of this result. The time-independent wave function solution is given by

$$\psi(\mathbf{r}) = \mathbf{A}\exp(\mathbf{k}\cdot\mathbf{r}), \tag{1.28}$$

where **A** is a complex quantity and is the amplitude, **k** is the wave vector and **r** is the three-dimensional space vector. This results in energies of the same form as Equation (1.21).

1.10.2 Fermi–Dirac distribution

The Fermi–Dirac distribution function $f(E)$ gives the probability that states with energy E are occupied by particles [10]:

$$f(E) = \frac{1}{1+\exp\left(\frac{E-E_{\mathrm{F}}}{kT}\right)}, \tag{1.29}$$

E_{F} represents the Fermi energy where f(E) becomes equal to 1/2.

1.11 Electrons in a semiconductor

Since semiconductors have periodic lattice structures, the electrons are subjected to a periodic potential. Hence the Schrödinger equation must be solved for a periodic potential [10]. The *Bloch theorem* states that the one-dimensional wave function for an electron in a periodic potential is given by

$$\psi(x) = V_{\mathrm{k}}(x)\exp(jkx), \tag{1.30}$$

where $V_{\mathrm{k}}(x)$ is a periodic potential with the same periodicity as the semiconductor crystal with lattice constant a such that

$$V_{\mathrm{k}}(x) = V_{\mathrm{k}}(x+na), \tag{1.31}$$

where n is an integer.

1.12 The Kronig–Penney model

An important model for the band structure is the Kronig–Penney model (Figure 1.15).

The one-dimensional periodic potential is given by

$$V(x) = 0,\ 0 \le x \le a \tag{1.32}$$

$$V(x) = V_0,\ -b \le x \le 0 \tag{1.33}$$

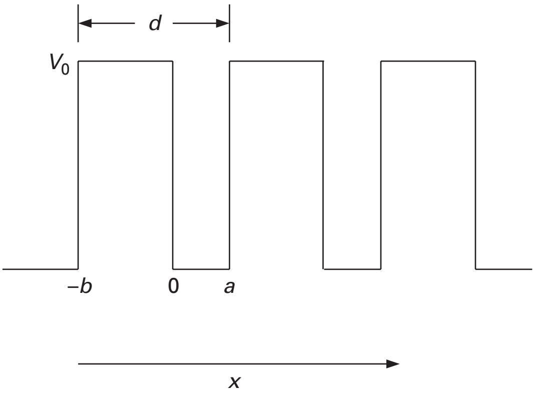


Fig. 1.15 Periodic potential for Kronig–Penney model.

The periodicity distance is $d = a + b$. The wave equation is solved in the three regions and the continuity conditions for the wave function and its derivatives are applied. The non-trivial solutions are obtained when the electron energy is less than and greater than the potential V_0 [10, 14]. The transcendental equation to be solved is

$$\cos k_x d = \cos a\alpha \cosh\ b\delta - \frac{\alpha^2 - \delta^2}{2\alpha\delta} \sin a\alpha \sinh\ b\delta,\ 0 < E < V_0 \tag{1.34}$$

$$\cos k_x d = \cos a\alpha \cos b\delta - \frac{\alpha^2 + \delta^2}{2\alpha\delta} \sin a\alpha \sin b\delta,\ E > V_0 \tag{1.35}$$

with

$$\alpha = \sqrt{\frac{2m_0 E}{\hbar^2}},\ \beta = \sqrt{\frac{2m_0(E - V_0)}{\hbar^2}}, \delta = \sqrt{\frac{2m_0(V_0 - E)}{\hbar^2}}. \tag{1.36}$$

The solution of the equation gives the energy E. The allowed energy bands are separated by band gaps with no allowed energies. It follows that there are forbidden energy regions for an electron which is subjected to a periodic potential in a semiconductor crystal.

1.12.1 Effective mass

When the centre of mass of a classical particle moves with a velocity v, we define a phase velocity. If we have a packet of travelling waves with a centre frequency ω and a wavenumber k, we have the classical dispersion relation for the group velocity:

$$v_g = \frac{d\omega}{dk}. \tag{1.37}$$

In the quantum mechanical formulation, the wavepacket is the analogue of the classical particle in a given region of space. This wavepacket consists of constant-energy wave function solutions and a centre energy is defined. Hence the wavepacket group velocity in the quantum-mechanical formulation can be written as in Equation (1.20):

$$v_g = \frac{1}{\hbar}\frac{dE}{dk}. \tag{1.38}$$

Using the force–momentum relations, we define the effective mass of an electron in a crystal as

$$m^* = \left(\frac{1}{\hbar^2}\frac{d^2E}{dk^2}\right)^{-1}. \tag{1.39}$$

Section 1.14.1 defines heavy and light holes corresponding to wide and narrow bands respectively.

1.12.2 Carriers in semiconductors

The two types of carriers in semiconductors are the conduction band electrons and the valence band holes. The electrons occupy the conduction band when the temperature is raised above 0 K. The unoccupied states in the valence band are holes and are defined to have a positive charge with the same magnitude as the electronic charge. Hence, we consider electrons in determining the conduction band properties and holes in determining the valence band properties. The band structures of several semiconductors are given by Pierret, and Streetman and Banerjee [10, 15] and others.

1.13 Semiconductors in equilibrium

1.13.1 Intrinsic semiconductors

A semiconductor is described as being intrinsic when there are no impurities and no defects in the crystal. The concentration of electrons in the conduction band is equal to the concentration of holes in the valence band. At 0 K, the electrons occupy all the available energy states in the valence band and all the states in the conduction band are empty. This follows from the fact that at 0 K, each electron is in the lowest possible energy state. As the temperature is increased the electrons are excited due to the acquired thermal energy and move into the conduction band leaving behind holes in the valence band. Therefore, the equilibrium concentration of electrons in the conduction band n_0 is equal to the equilibrium concentration of holes in the valence band p_0 in intrinsic semiconductors [2, 15]:

$$n_0 = p_0 = n_i, \tag{1.40}$$

where n_i is simply referred to as the intrinsic concentration of holes and electrons.

1.13.2 Extrinsic semiconductors

When impurity atoms are added to the intrinsic semiconductor such that the electron concentration is no longer equal to the hole concentration, it becomes an extrinsic

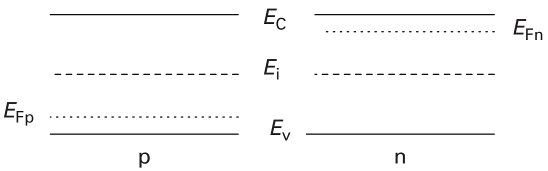


Fig. 1.16 Band diagram.

semiconductor and $n_0 \neq p_0$. Thus the doping of a semiconductor with impurities can produce excess electrons or holes. These atoms can be either donors or acceptors. If the dopant produces an excess of electrons, the dopant is referred to as a *donor*, the semiconductor becomes n-type material with $n > p$ and the current is predominantly due to the negatively charged electrons. If, on the other hand, the dopant generates holes, the dopant is referred to as an *acceptor*, the result is a p-type semiconductor with $p > n$ and the current is predominantly due to the positively charged holes. Note that the hole charge has the same magnitude as the electronic charge [2, 8, 15].

1.13.3 Semiconductor band diagrams

The band diagrams for p- and n-type semiconductors at thermal equilibrium are given in Figure 1.16. The bottom of the conduction band is E_c, the top of the valence band is E_v, the intrinsic energy level is at mid-band and is denoted by E_i and the Fermi level is E_F.

1.13.4 Electron and hole distribution

The distribution of electrons in the conduction band and holes in the valence band is obtained using the Fermi–Dirac probability function. The electron distribution in the conduction band is written as

$$n(E) = g_c(E) f(E), \tag{1.41}$$

where $g_c(E)$ is the density of quantum states in the conduction band and $f(E)$ is the Fermi–Dirac probability function given in Equation (1.29). The hole distribution in the valence band can be written in a similar way:

$$p(E) = g_v(E)[1 - f(E)]. \tag{1.42}$$

The density of states functions are written as

$$g_c(E) = \frac{m_n^* \sqrt{2m_n^*(E - E_c)}}{\pi^2 \hbar^3}, \quad E \geq E_c \tag{1.43}$$

$$g_v(E) = \frac{m_p^* \sqrt{2m_p^*(E_v - E)}}{\pi^2 \hbar^3}, \quad E \leq E_v. \tag{1.44}$$

The equilibrium concentration of electrons can now be written as

$$n_0 = \int_{E_c}^{\infty} n(E) \mathrm{d}E, \tag{1.45}$$

where $n(E)$ is given by Equation (1.41). Similarly, the equilibrium hole concentration is written as

$$p_0 = \int_{-\infty}^{E_\mathrm{v}} p(E)\mathrm{d}E, \tag{1.46}$$

where $p(E)$ is given by Equation (1.42). The equilibrium electron and hole concentrations in the conduction and valence bands respectively are written as

$$n_0 = N_\mathrm{c} \exp\left(\frac{-(E_\mathrm{c} - E_\mathrm{F})}{kT}\right) \tag{1.47}$$

$$p_0 = N_\mathrm{v} \exp\left(\frac{-(E_\mathrm{F} - E_\mathrm{v})}{kT}\right), \tag{1.48}$$

where N_c and N_v are the effective density of states functions in the conduction and valence bands respectively.

$$N_\mathrm{c} = 2\left(\frac{2\pi m_\mathrm{n}^* kT}{h^2}\right)^{3/2} \tag{1.49}$$

$$N_\mathrm{v} = 2\left(\frac{2\pi m_\mathrm{p}^* kT}{h^2}\right)^{3/2}. \tag{1.50}$$

The intrinsic carrier concentration n_i is given by

$$n_\mathrm{i}^2 = n_0 p_0. \tag{1.51}$$

By substitution of Equations (1.47) and (1.48), we can write the intrinsic concentration as

$$n_\mathrm{i}^2 = N_\mathrm{c} N_\mathrm{v} \exp\left(\frac{-(E_\mathrm{c} - E_\mathrm{v})}{kT}\right) \tag{1.52}$$

$$= N_\mathrm{c} N_\mathrm{v} \exp\frac{-E_\mathrm{g}}{kT}, \tag{1.53}$$

where E_g is the bandgap energy.

1.14 Direct and indirect semiconductors

When light illuminates a semiconductor, and the photon energy is equal to or larger than the band gap, the light is absorbed, and creates hole–electron pairs. These holes and electrons are equal in number to maintain charge neutrality, and since they are not in equilibrium, in due course they recombine; this recombination may be radiative or non-radiative. Radiative recombination, when a photon is emitted usually at the bandgap energy, only occurs in direct bandgap material, whereas non-radiative recombination may occur in both direct and indirect bandgap semiconductors. In indirect semiconductors, this non-radiative recombination requires a phonon to mediate the process. Non-radiative processes in direct bandgap material are usually through traps or due to surface recombination. The direct or indirect band gap defines whether the lowest position of the conduction band aligns with the maximum of the valence band along momentum space, where the effective momentum value k is equal to zero.

Direct bandgap semiconductors are capable of photon emission, by radiative recombination, but indirect semiconductors have a low probability of radiative recombination. However, indirect bandgap semiconductors may have isoelectronic impurity states which are direct, and therefore the recombination from these states may also be radiative. GaP, which is an indirect gap semiconductor, may be doped with zinc oxide or nitrogen to produce these states, and the widely used green or red light–emitting diodes are examples of this emission.

1.14.1 Absorption processes

Considering Figure 1.17, we note that three different valence bands are shown. Equation (1.39) had linked the carrier mobility the second derivative of energy with respect to k. It is then easy to understand why the band with less curvature at $k = 0$ (V_1) is called the “heavy hole band”, while the one which is more strongly bent (V_2) is called “light hole band”. In most bulk semiconductors, the light and heavy hole bands concide at $k = 0$ – they are degenerate. The band V_3 is called the “split-off band”. Note

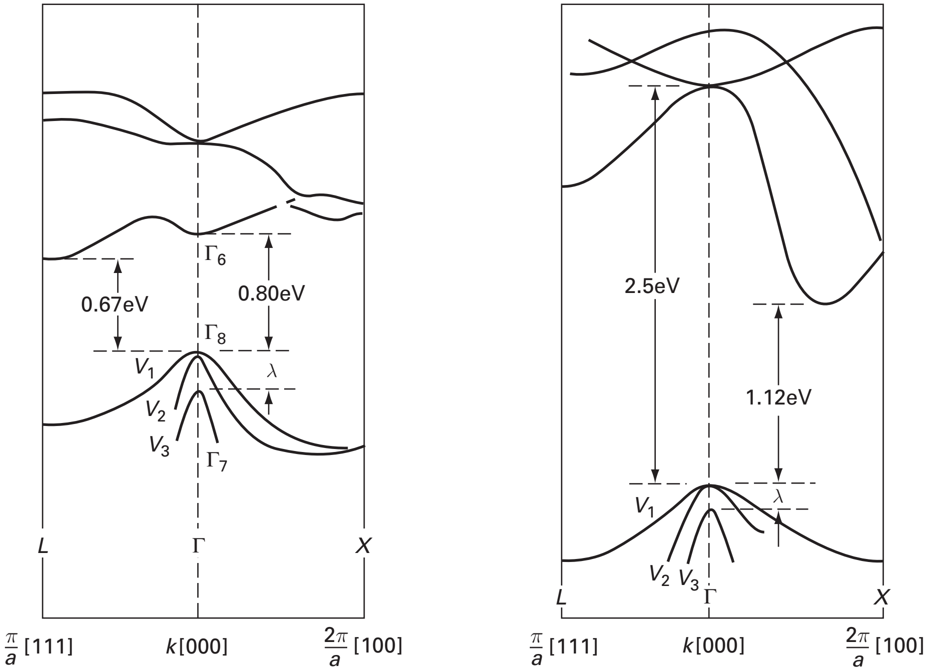


Fig. 1.17 Band diagram of bulk Ge (left) and Si (right). Note that the minimum of the conduction band is not aligned with maximum of the valence band at $k = 0$, indicating that these are indirect semiconductors. S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, 1st Edition, pp. 233–234, ©1989. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.

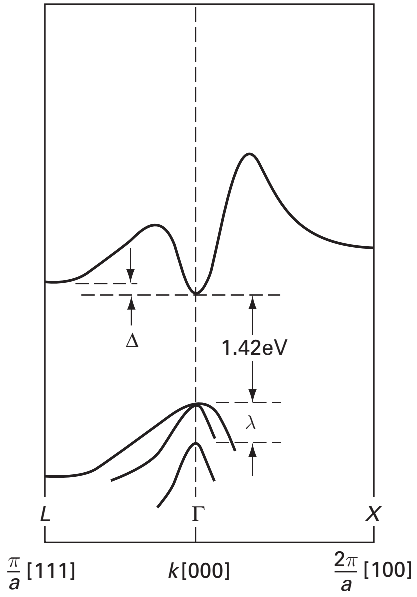


Fig. 1.18 Band diagram of bulk GaAs. Note that the minimum of the conduction band is aligned with maximum of the valence band at $k = 0$, indicating that this is a direct gap semiconductor. S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, 1st Edition, pp. 233–234, ©1989. Reprinted by permission of Pearson Education Inc., Upper Saddle River, NJ.

that in each of these cases, the minimum of the conduction band is not aligned with the maximum of the valence band where $k = 0$, which implies that these are indirect semiconductors. In Figure 1.18, the band diagram of the direct semiconductor GaAs is shown; here the conduction band minimum is along the $k = 0$ axis and aligned with the maximum of the valence band. Note the degenerate valence band of heavy holes and light holes and the split-off band.

Photo-excitation of semiconductors with photons energies equal to or greater than the bandgap energy of the material results in absorption, which in turn causes the creation of hole–electron pairs for each photon. The major source of absorption in semiconductors is the valence band to conduction band transition. In the case of direct semiconductors, the transition occurs when the photon energy is at the bandgap value or larger and results in the transition of an electron in the valence band to the conduction band. In indirect semiconductors, the absorption has to be mediated by phonons. In addition to the band to band absorption, transitions take place from acceptor to donor levels, from acceptor to conduction band, valence to donor level, all of which result in absorption below the bandgap energy. Figure 1.19 shows schematically the band to band transition for the direct gap semiconductor, and in Figure 1.20, the phonon-mediated transition. The conduction band to valence band and impurity band to impurity band transitions are shown schematically in Figure 1.21. Not shown in this figure are the impurity band to conduction and valence band transitions, all of which lead to absorption and emission of the appropriate photon energies.

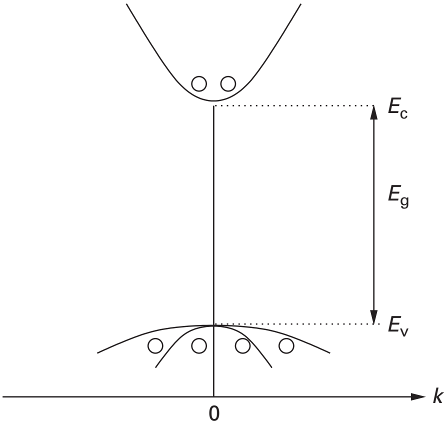


Fig. 1.19 Schematic diagram of the conduction and valence bands of a direct semiconductor and the transitions.

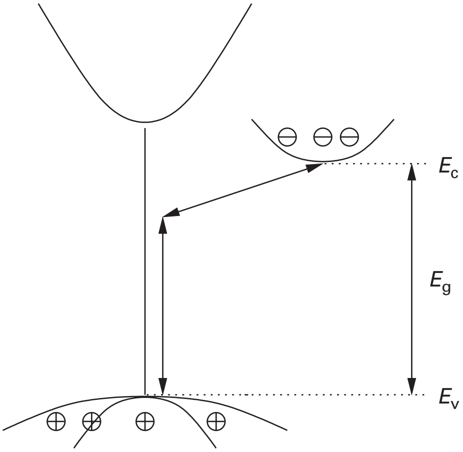


Fig. 1.20 Schematic diagram of the indirect semiconductor and the phonon-mediated transitions.

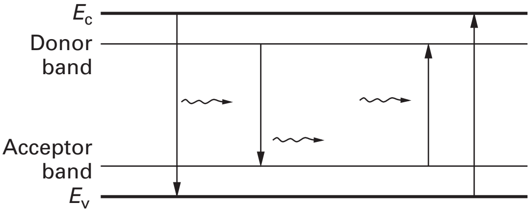


Fig. 1.21 Transitions possible with a semiconductor with impurity donor and acceptor bands: conduction band to valence band and impurity band to impurity band are illustrated. Others, conduction band to impurity band and valence band to impurity band have not been shown.

The absorption rate of the band to band transitions, for both the direct and indirect transitions may be calculated using quantum theory, but is not included here.

Free carrier absorption also occurs in most semiconductors as the carrier density is always non-zero. The absorption of a photon by a carrier within a band results in the carrier having a larger energy. The absorption coefficient is proportional to the carrier density [3]. This effect is important in the design of waveguide devices, where typically this may result in absorption of the order of 1 dB cm^{-1} when the carrier densities are high in the 10^{18} cm^{-3} region.

1.14.2 Exciton absorption

In pure semiconductors, the absorbed photon with bandgap energy or larger may create excitons, which are electron–hole pairs that are bound, and in the binding process give up the binding energy. The binding energy of these excitons is of the order of about 4.5 meV, and at low temperatures, an excitonic absorption peak is seen a little below the band to band absorption energy. At room temperature, this peak is not seen in bulk material, because the thermal broadening due to optical phonons is comparable, and the excitons that are created dissociate very rapidly. In quantum wells, however, the excitons remain extant at room temperature due to enhanced binding energies, which are typically two or three times that of the thermal broadening energy. Thus, the absorption characteristics of the material with quantum wells also show the excitonic absorption in addition to the usual band to band absorption. When a transverse electric field is applied to the quantum well, the absorption edge shifts to a longer wavelength. A simple explanation of this phenomenon is shown in Figure 1.22, where the schematic wave functions of the electron and hole in the quantum well are shown. When the transverse field is applied, then the quantum well bands tilt, and the resulting gap between the electron–hole wave functions decreases, which results in the absorption edge moving to a smaller energy and thus a longer wavelength.

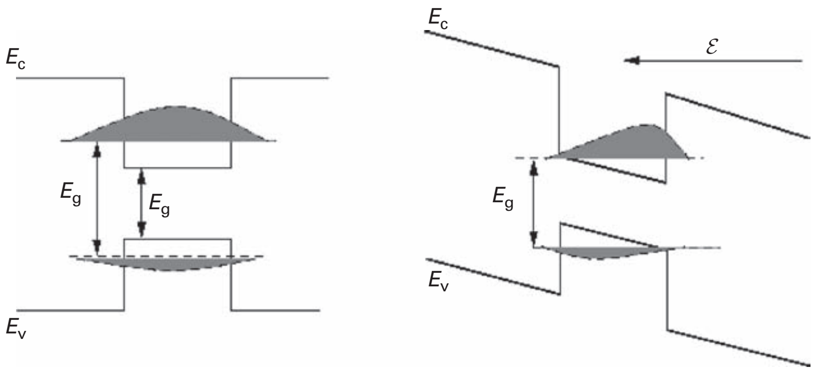


Fig. 1.22 A schematic diagram of the wave functions in a quantum well, and the effect of applying a field across the well, resulting in tilting of the wells. This so-called Quantum Confined Stark Effect reduces the effective band gap of the material.

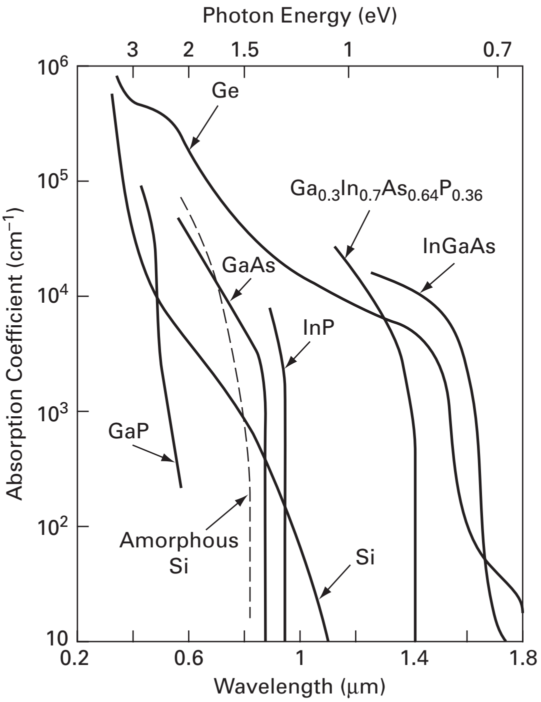


Fig. 1.23 Absorption coefficient for various semiconductors (M. Shur, *Physics of Semiconductor Devices*, Prentice Hall, 1990 ©Prentice Hall).

Other absorption mechanisms are due to valence to impurity band, impurity band to other impurity band or impurity band to conduction band transition, intraband absorption between different levels in the same band, and free carrier absorption.

The absorption spectra of different semiconductors is summarised in Figure 1.23.

1.15 Recombination and radiation in semiconductors

The absorption of photons by the semiconductor results in the generation of electrons and holes, which disturbs the equilibrium status of the semiconductor. Electrical injection also results in this non-equilibrium of an excess of electrons in the conduction band and an equal number of holes in the valence band. These recombine, both non-radiatively and radiatively, the latter in direct gap semiconductors. In general, the radiative transitions are dominated by the conduction band to valence band emission and therefore define the energy of the emitted photons. Other recombination processes include exciton recombination, donor to acceptor and other impurity recombinations. The radiation spectrum from recombination is generally shifted to lower energy from the absorption spectrum, and this is termed the *Stokes* or the *Franck–Condon shift* due to imperfections in materials or interfaces.

In general the excess electrons and holes decay at some rate, resulting in the density varying as $\exp(-t/\tau)$, where τ is defined as the lifetime of the carriers. The decay of these carriers results in transfer of energy to the lattice in the form of phonons for the non-radiative decay and transfer of energy to photons for radiative decay.

The corresponding lifetimes are labelled as τ_{nr} and τ_{r} for the non-radiative and radiative decay, respectively, and the corresponding non-radiative and radiative rates are R_{nr} and R_{r}, respectively. Thus, the total lifetime constant τ is given as:

$$\frac{1}{\tau} = \frac{1}{\tau_{\mathrm{nr}}} + \frac{1}{\tau_{\mathrm{r}}}. \tag{1.54}$$

The corresponding total spontaneous rate of recombination is given by

$$R_{\mathrm{spon}} = R_{\mathrm{nr}} + R_{\mathrm{r}}. \tag{1.55}$$

Devices such as the light-emitting diode (LED) largely depend on spontaneous emission, and in this case the internal quantum efficiency is given by

$$\eta_{\mathrm{internal}} = \frac{R_{\mathrm{r}}}{R_{\mathrm{nr}} + R_{\mathrm{r}}}. \tag{1.56}$$

The exponential decay rate of the excess carriers is inversely proportional to recombination rate, and if the excess of electron is Δn, then the recombination rates R_{r} and R_{nr} are given by the expressions: $R_{\mathrm{r}} = \Delta n/\tau_{\mathrm{r}}$ and $R_{\mathrm{nr}} = \Delta n/\tau_{\mathrm{nr}}$. Then internal quantum efficiency may also be written as

$$\eta_{\mathrm{internal}} = \frac{\frac{1}{\tau_{\mathrm{r}}}}{\frac{1}{\tau_{\mathrm{r}}} + \frac{1}{\tau_{\mathrm{nr}}}} = \frac{1}{1 + \frac{\tau_{\mathrm{r}}}{\tau_{\mathrm{nr}}}} = \frac{\tau_{\mathrm{nr}}}{\tau_{\mathrm{r}} + \tau_{\mathrm{nr}}}. \tag{1.57}$$

The total spontaneous recombination rate is given by the equation:

$$R_{\mathrm{total}} = A\Delta n + B\Delta n^2 + C\Delta n^3. \tag{1.58}$$

The first term is the Shockley–Read–Hall recombination due to defects and traps, the second is the spontaneous emission due to radiative transition, and the third is the Auger recombination term. Auger recombination is non-radiative, occurs at high injection levels, and is a three-particle process. It becomes important in ternary and quaternary compounds of InP-based materials, and is evident in the long wavelength laser structures.

1.15.1 Spontaneous and stimulated emission

The radiative recombination process discussed above occurs spontaneously, and this is used in traditional LED structures. In lasers, stimulated emission is the source of light, and in this section the relationship between absorption, spontaneous emission and stimulated emission, first outlined by Einstein in 1917, is discussed. The derivations given here follow the approach outlined by Casey and Panish [4] and Agrawal [1].

It can be shown that the blackbody radiation law is given by

$$P(E) = \frac{8\pi\overline{n}^3 E^2}{h^3 c^3 (e^{E/kT} - 1)}, \tag{1.59}$$

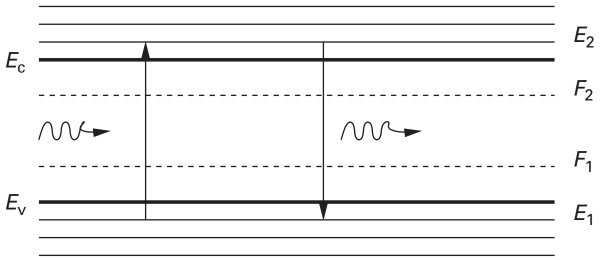


Fig. 1.24 Transitions from level E_1 to E_2 for absorption and from E_2 to E_1 for emission, F_1 and F_2 are the electron and hole quasi-Fermi levels respectively.

where $\overline{n}$ is the index of the material under consideration, h is Planck's constant, c is the speed of light in vacuum, E is the energy given by $h\nu$, ν is the frequency and k is Boltzmann's constant. This is the expression for the energy density blackbody radiation $P(E)$, and is in thermal equilibrium, when the input radiation is equal to the outgoing radiation.

For a semiconductor, consider the transitions from the conduction band to the valence band and also the reverse. The energy levels in each of these bands have to obey the Pauli exclusion principle, which implies only two carriers at each level. Thus, the band is a series of levels, as shown in Figure 1.24, and the transition energy for an electron from a level E_1 in the valence band to a level E_2 in the conduction band requires that an incident photon has energy given by $h\nu = E_2 - E_1$. Let the probability of this transition taking place be given by B_{12}, and let f_1 be the probability that an electron exists at level E_1 and $(1 - f_2)$ be the probability that a vacancy occurs at level E_2. Also assume that the radiation density of photon energy incident on the semiconductor is given by $P(E_{21})$. Then the upward transition rate is given by

$$r_{12} = B_{12} f_1 (1 - f_2) P(E_{21}). \tag{1.60}$$

Note that f_1 and f_2 take the form of the Fermi–Dirac distribution

$$f_\mathrm{i} = \frac{1}{e^{(E_\mathrm{i} - F_\mathrm{i})/kT} + 1}, \tag{1.61}$$

where F_i is the corresponding quasi-Fermi level, k is Boltzmann's constant and T the temperature in Kelvin.

Similarly, the downward transition rate, now called the *stimulated transition*, is given as

$$r_{21}(\text{stim}) = B_{21} f_2 (1 - f_1) P(E_{21}), \tag{1.62}$$

where B_{21} is the transition probability, f_2 is the probability that an electron is present at E_2 and $(1 - f_1)$ is the probability that there is vacancy at E_1.

Finally, there is the spontaneous transition from E_2 to E_1, without any incident radiation involved, given by

$$r_{21}(\text{spon}) = A_{21} f_2 (1 - f_1). \tag{1.63}$$

In thermal equilibrium, the input radiation is equal to the output, the Fermi levels $F_1 = F_2$, and hence

$$r_{12} = r_{21}(\text{spon}) + r_{21}(\text{stim}). \tag{1.64}$$

Equating, simplifying and noting that $P(E_{21})$ is the blackbody radiation term,

$$\begin{aligned} P(E_{21}) &= \frac{8\pi \overline{n}^3 E^2}{h^3 c^3 (e^{E_{21}/kT} - 1)} &(1.65)\\ &= \frac{A_{21} f_2(1-f_1)}{B_{12} f_1(1-f_2) - B_{21} f_2(1-f_1)} &(1.66)\\ &= \frac{A_{21}}{B_{12} e^{E_{21}/kT} - B_{21}}. &(1.67)\end{aligned}$$

Equating Equations (1.65) and (1.67) and separating them into temperature-dependent and temperature-independent terms give the following results:

$$A_{21} = \frac{8\pi \overline{n}^3 E^2}{h^3 c^3} B_{21} \tag{1.68}$$

and

$$B_{21} = B_{12}. \tag{1.69}$$

These are Einstein's coefficients and their relationships with each other.

The condition under which stimulated emission dominates is an interesting one. This requires a non-equilibrium condition in which the presence of incident radiation is required. This results in the population densities in the conduction and valence bands to be different from the equilibrium condition. For stimulated emission to dominate, the stimulated emission rate, $r_{21}(\text{stim})$, needs to exceed the absorption rate r_{12}. Substituting from Equations (1.60) and (1.62),

$$B_{21} f_2(1-f_1) P(E_{21}) > B_{12} f_1(1-f_2) P(E_{21}). \tag{1.70}$$

Since $B_{21} = B_{12}$, this equation becomes

$$f_2(1-f_1) > f_1(1-f_2). \tag{1.71}$$

Substituting for f_1 and f_2 from Equation 1.61, this equation becomes

$$e^{(F_2-F_1)/kT} > e^{(E_2-E_1)/kT} \tag{1.72}$$

or

$$F_2 - F_1 > E_2 - E_1. \tag{1.73}$$

This implies that the difference in the quasi-Fermi levels is greater than the emission energy of the photon. If the emission is at bandgap energy, then the difference between quasi-Fermi levels needs to be greater than the bandgap energy E_g.

1.16 Carrier transport in semiconductors

Drift and diffusion are the two mechanisms whereby carriers are transported in semiconductors such that there is current flow. It will be assumed that thermal equilibrium will not be disturbed during these processes [2, 8, 10, 15].

1.16.1 Drift current

When an external electric field is applied to a semiconductor, it produces a force that will accelerate the electrons and holes in opposite directions as long as there are available energy states in the conduction and valence bands. The net drift of charge will produce a current which is the drift current. If the electric field is denoted as $\mathcal{E}$, the drift current densities for electrons and holes are written as

$$J_{\mathrm{n(drift)}} = qnv_{\mathrm{ndr}} = q\mu_{\mathrm{n}}n\mathcal{E} \tag{1.74}$$

$$J_{\mathrm{p(drift)}} = qpv_{\mathrm{pdr}} = q\mu_{\mathrm{p}}n\mathcal{E}, \tag{1.75}$$

where q is the charge on a particle (electron or hole), J is the surface density of current, v_{ndr} and v_{pdr} are the drift velocities of electrons and holes, respectively, and μ is the mobility.

1.16.2 Diffusion current

Electrons flow from a region of higher concentration to a region of lower concentration, producing a flux of electrons and an electron diffusion current which is in the opposite direction to the flux. The hole flow is such that the hole flux and the hole diffusion current are in the same direction since the holes are positively charged. The diffusion current densities for electrons and holes are given by

$$J_{\mathrm{ndiff}} = qD_{\mathrm{n}}\frac{dn}{dx} \tag{1.76}$$

$$J_{\mathrm{pdiff}} = -qD_{\mathrm{p}}\frac{dp}{dx}, \tag{1.77}$$

where D_{n} and D_{p} are the electron and hole diffusion coefficients respectively. The diffusion coefficient is related to the mobility μ by the Einstein relation:

$$D = \frac{\mu kT}{q}. \tag{1.78}$$

Hence,

$$\frac{D_{\mathrm{n}}}{\mu_{\mathrm{n}}} = \frac{D_{\mathrm{p}}}{\mu_{\mathrm{p}}} = \frac{kT}{q}. \tag{1.79}$$

Adding Equations (1.74)–(1.77),

$$J = (q\mu_{\mathrm{n}}n + q\mu_{\mathrm{p}}p)\mathcal{E} + qD_{\mathrm{n}}\frac{dn}{dx} - qD_{\mathrm{p}}\frac{dp}{dx}. \tag{1.80}$$

1.17 p–n junction

When a junction is formed by a p-type and an n-type semiconductor, holes move from the p to the n side across the metallurgical junction and electrons move in the opposite direction. There are concentration gradients of electrons and holes giving rise to diffusion. Furthermore, when the electrons leave the n region, positively ionised donor atoms remain behind and, similarly, negatively ionised acceptor atoms remain in the p region. These ionised donors and acceptors reside on both sides of the metallurgical junction and are not mobile. The length of this region increases as the diffusion continues. The resultant electric field is directed from the positive charge to the negative charge. This field builds up in such a way as to oppose the diffusion of the carriers in both directions across the junction. An equilibrium condition is reached and there is no net flow of current across the junction. The ionised region on both sides of the metallurgical junction is called the *depletion* or *space charge region*. The p–n junction is seen in Figure 1.25. The band diagram and space charge distribution of a p–n homojunction are shown in Figure 1.26. The general form of Poisson's equation for an abrupt junction where there is an abrupt change in the doping concentration is

$$\frac{d\mathcal{E}}{dx} = \frac{q}{\epsilon_\mathrm{s}}(p - n + N_\mathrm{D} - N_\mathrm{A}), \tag{1.81}$$

where p is the hole concentration, n is the electron concentration and N_D and N_A are the ionised donor and acceptor concentrations respectively. If the metallurgical junction is the origin, the depletion region on the p side extends to $-x_\mathrm{p}$ and on the n side to x_n. Poisson's equation is written for the depletion regions on either side of the metallurgical junction as

$$\frac{d\mathcal{E}}{dx} = -\frac{q}{\epsilon_\mathrm{s}}N_\mathrm{A}, \quad -x_\mathrm{p} < x \le 0 \tag{1.82}$$

$$\frac{d\mathcal{E}}{dx} = \frac{q}{\epsilon_\mathrm{s}}N_\mathrm{D}, \quad 0 < x \le x_\mathrm{n}, \tag{1.83}$$

where ϵ_s is the permittivity of the semiconductor material. The regions outside the depletion region are neutral regions and hence the electric field is zero. Equations (1.82) and (1.83) are solved using the boundary condition on the electric field to get

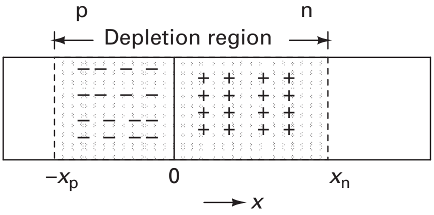


Fig. 1.25 p–n junction at equilibrium.

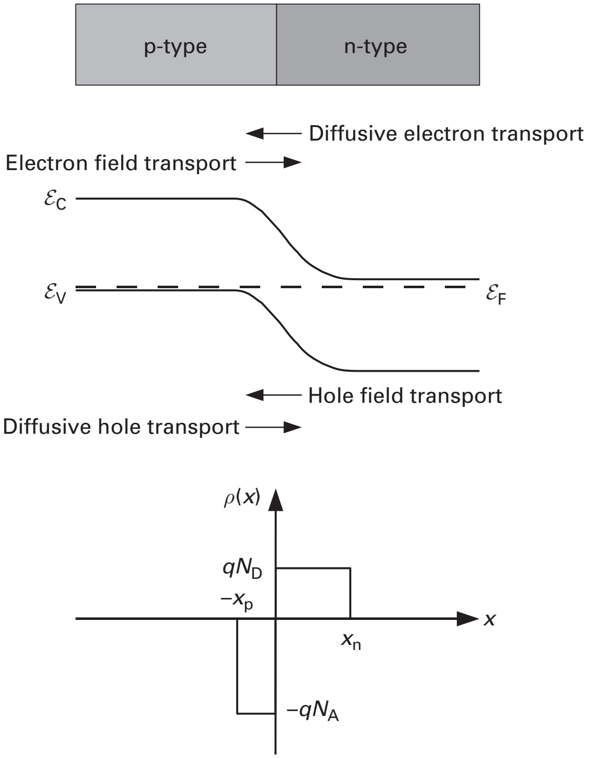


Fig. 1.26 p–n junction without externally applied voltage – band diagram (centre) and space charge distribution (bottom).

$$\mathcal{E} = -\frac{qN_A}{\epsilon_s}(x + x_p), \quad -x_p < x \leq 0 \tag{1.84}$$

$$\mathcal{E} = \frac{qN_D}{\epsilon_s}(x - x_n), \quad 0 < x \leq x_n. \tag{1.85}$$

The potential is related to the electric field by the equation

$$\mathcal{E} = -\frac{dV}{dx}. \tag{1.86}$$

Integrating Equations (1.84) and (1.85)

$$V(x) = \frac{qN_A}{2\epsilon_s}(x + x_p)^2, \quad -x_p < x \leq 0 \tag{1.87}$$

$$V(x) = -\frac{qN_D}{2\epsilon_s}(x - x_n)^2, \quad 0 < x \leq x_n. \tag{1.88}$$

1.17.1 The built-in potential

The built-in potential on the p side of the junction is the potential difference across the depletion region. It is determined similarly on the n side.

$$V_{\text{bip}} = \frac{qN_{\text{A}}}{2\epsilon_{\text{s}}}x_{\text{p}}^2 \tag{1.89}$$

$$V_{\text{bin}} = \frac{qN_{\text{D}}}{2\epsilon_{\text{s}}}x_{\text{n}}^2. \tag{1.90}$$

The total built-in potential across the junction is

$$V_{\text{bi}} = (V_{\text{bip}} + V_{\text{bin}}) \tag{1.91}$$

$$= \frac{q}{2\epsilon_{\text{s}}}[N_{\text{A}}x_{\text{p}}^2 + N_{\text{D}}x_{\text{n}}^2]. \tag{1.92}$$

The continuity of the electric field across the junction at $x = 0$ requires that

$$N_{\text{A}}x_{\text{p}} = N_{\text{D}}x_{\text{n}}. \tag{1.93}$$

It is assumed that the dopants are fully ionised and the total ionised positive charge per unit area on the n side is equal to the total ionised negative charge per unit area on the p side. At thermal equilibrium, there is no net current flow and hence the drift and diffusion currents are equal. The electron current is

$$J_{\text{n}} = 0 \tag{1.94}$$

$$= J_{\text{ndrift}} + J_{\text{ndiff}} \tag{1.95}$$

$$= q\mu_{\text{n}}n\mathcal{E} + qD_{\text{n}}\frac{dn}{dx}. \tag{1.96}$$

The hole current is written as

$$J_{\text{p}} = 0 \tag{1.97}$$

$$= J_{\text{pdrift}} + J_{\text{pdiff}} \tag{1.98}$$

$$= q\mu_{\text{p}}p\mathcal{E} - qD_{\text{p}}\frac{dp}{dx}. \tag{1.99}$$

When the net hole current is zero, and with the electric field equal to the gradient of the potential, it may be shown that

$$V_{\text{bi}} = \frac{kT}{q}\ln\frac{N_{\text{A}}N_{\text{D}}}{n_{\text{i}}^2}. \tag{1.100}$$

It has been assumed that there is full ionisation of the dopant impurity levels such that the majority carrier concentrations are the doping concentrations and the equilibrium concentrations are related by

$$n_0 p_0 = n_{\text{i}}^2. \tag{1.101}$$

1.17.2 The depletion layer width

The widths of the depletion layer in the p- and n-type semiconductors may be calculated. The maximum electric field occurs at the metallurgical junction, $x = 0$. This is given by

$$\epsilon_{\text{s}}\mathcal{E}_{\text{max}} = qN_{\text{D}}x_{\text{n}} = qN_{\text{A}}x_{\text{p}}. \tag{1.102}$$

Using the Equations (1.92) and (1.93) with

$$| V_{\mathrm{bi}} | = \frac{\mathcal{E}_{\max}}{2}[x_{\mathrm{n}} + x_{\mathrm{p}}] \tag{1.103}$$

$$W = x_{\mathrm{n}} + x_{\mathrm{p}} \tag{1.104}$$

$$= \sqrt{\frac{2\epsilon_{\mathrm{s}}}{q}\left(\frac{1}{N_{\mathrm{D}}} + \frac{1}{N_{\mathrm{A}}}\right) | V_{\mathrm{bi}} |}. \tag{1.105}$$

1.17.3 The depletion capacitance

The depletion capacitance is the capacitance at the p–n junction. The depletion layer is modelled as a parallel plate capacitor. The capacitance is written as

$$C_{\mathrm{j}} = \frac{\epsilon_{\mathrm{s}} A}{W}, \tag{1.106}$$

where A is the area of the p–n junction and W is the depletion layer width given by Equation (1.105). The junction capacitance is given by

$$C = A\sqrt{\frac{q\epsilon_{\mathrm{s}} N_{\mathrm{A}} N_{\mathrm{D}}}{2V_{\mathrm{bi}}(N_{\mathrm{A}} + N_{\mathrm{D}})}}. \tag{1.107}$$

1.17.4 p–n junction under bias

At thermal equilibrium, the total electrostatic potential across the p–n junction is the built-in potential,V_{bi}, and the potential difference between the p and n regions is qV_{bi}. If now a voltage V_{A} is applied with the positive terminal connected to the p side and the negative to the n side, the junction is forward-biased and the total electrostatic potential across the junction is $V_{\mathrm{bi}} - V_{\mathrm{A}}$, resulting in a reduction of the depletion layer width. A potential barrier was formed at thermal equilibrium restricting the motion of the majority carriers. The application of the forward bias reduces the height of the barrier. If, on the other hand, a voltage is applied with the positive terminal connected to the n side and the negative terminal to the p side, the electrostatic potential across the junction is $V_{\mathrm{bi}} - (-V_{\mathrm{A}})$ and the height of the barrier is increased with the reverse bias. The depletion widths and the energy band diagrams are shown in the figure. The width of the depletion layer is given by

$$W = \sqrt{\frac{2\epsilon_{\mathrm{s}}}{q}\left(\frac{N_{\mathrm{D}} + N_{\mathrm{A}}}{N_{\mathrm{A}} N_{\mathrm{D}}}\right)(V_{\mathrm{bi}} \mp V_{\mathrm{A}})}. \tag{1.108}$$

1.17.5 Current–voltage characteristics

The total current density in a p–n junction is given as:

$$J = q\left[\frac{D_{\mathrm{n}} n_{\mathrm{p0}}}{L_{\mathrm{n}}} + \frac{D_{\mathrm{p}} p_{\mathrm{n0}}}{L_{\mathrm{p}}}\right]\left[\exp\left(\frac{qV_{\mathrm{A}}}{kT}\right) - 1\right], \tag{1.109}$$

where

$$L_\mathrm{n} = \sqrt{D_\mathrm{n}\tau_\mathrm{n}} \tag{1.110}$$

$$L_\mathrm{p} = \sqrt{D_\mathrm{p}\tau_\mathrm{p}} \tag{1.111}$$

$$n_\mathrm{p0} = \frac{n_\mathrm{i}^2}{N_\mathrm{A}} \tag{1.112}$$

$$p_\mathrm{n0} = \frac{n_\mathrm{i}^2}{N_\mathrm{D}}. \tag{1.113}$$

The current density expression now reduces to

$$J = J_0\left[\exp\left(\frac{qV_\mathrm{A}}{kT}\right) - 1\right] \tag{1.114}$$

where

$$J_0 = qn_\mathrm{i}^2\left[\sqrt{\frac{D_\mathrm{n}}{\tau_\mathrm{n}}}\frac{1}{N_\mathrm{A}} + \sqrt{\frac{D_\mathrm{p}}{\tau_\mathrm{p}}}\frac{1}{N_\mathrm{D}}\right], \tag{1.115}$$

where D_n and D_p are the Einstein coefficients for electrons and holes and τ_n and τ_p are the electron and hole lifetimes.

1.18 Schottky diode

Another important type of semiconductor junction is the *Schottky diode*, a metal–semiconductor junction.

Its band diagram is shown in Figure 1.27, assuming that the semiconductor layer is n-doped, that the doping concentration is constant throughout the layer (homogeneous doping), and that the Boltzmann approximation for the Fermi distribution is

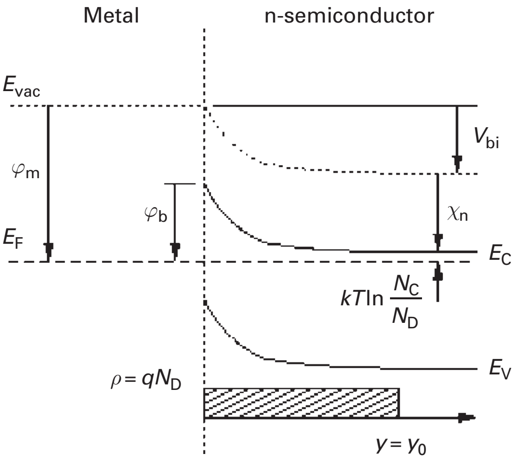


Fig. 1.27 Band diagram of a Schottky contact on a homogeneously n-doped semiconductor layer, without an externally applied voltage. The bottom graph shows schematically the extension of the space charge region.

valid. We see that under ideal circumstances, the Schottky barrier height φ_{b} is the difference between the metal work function φ_{m} and the electron affinity χ_{n}. In practice, however, the effective Schottky barrier height φ_{b} is also influenced by states at the metal–semiconductor interface and shows only a small dependence on the metal work function. On GaAs, $\varphi_{\mathrm{b}} \approx 0.8\,\mathrm{eV}$.

The built-in voltage can be easily calculated:

$$V_{\mathrm{bi}} = \varphi_{\mathrm{b}} - kT \ln \frac{N_{\mathrm{C}}}{N_{\mathrm{D}}}, \tag{1.116}$$

where N_{C} is the density of states in the conduction band and N_{D} the semiconductor doping concentration.

The Schottky barrier φ_{b} causes a depletion of the semiconductor layer immediately adjacent to the metal–semiconductor interface. Devoid of electrons, the positively charged ionised donors remain and form a *space charge region* with a space charge density $\rho = qN_{\mathrm{D}}$. It is indicated in the bottom graph of Figure 1.27 – we make the usual 'box shape' assumption for the space charge region, i.e. we assume the space charge concentration to be constant throughout until $y = y_0$, then ending abruptly.

For a homogeneously doped semiconductor with a permittivity of ϵ_{s} and a doping concentration of N_{D},

$$y_0 = \sqrt{\frac{2\epsilon_{\mathrm{s}} V_{\mathrm{bi}}}{qN_{\mathrm{D}}}} \tag{1.117}$$

without any externally applied voltage.

1.19 Heterostructures

The ability to mix semiconductors of different chemical composition in a single crystal gives an important degree of freedom in device design. The combination of semiconductor materials of different stoichiometry in a single crystal is called a *heterostructure*.

Of foremost interest is the ability to change the energetic width of the forbidden gap – the bandgap energy, or band gap in short. This is shown in Figure 1.28 for some popular atomic and binary semiconductor materials. We note that changing the stoichiometry not only modifies the bandgap energy, but also the lattice constant, which can be understood as an average distance between the atoms in the crystal. This change in lattice constant is a major complication when designing devices, but we will exclude it for now by considering the material system (Al,Ga)As, where the lattice constant is almost independent of stoichiometry.

The ability to change bandgap through stoichiometry opens up several interesting design options.

- We may, for example, introduce a built-in electric field for one carrier species, but not for the other – schematically shown in Figure 1.29. This is a p-type semiconductor which has a smaller bandgap on the left-hand side ($E_{\mathrm{g,I}}$) than on the right-hand side.

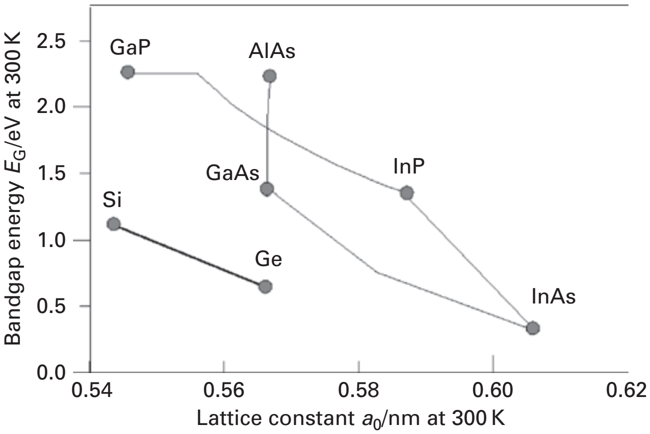


Fig. 1.28 Bandgap energy and lattice constant for several popular semiconductor materials.

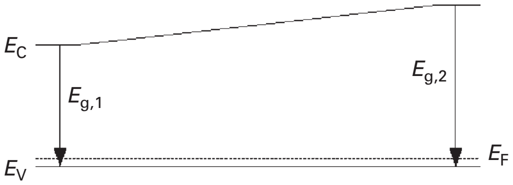


Fig. 1.29 Hypothetical band diagram of a p-doped graded heterostructure.

This could be done by starting with GaAs and gradually increasing the aluminium content while progressing to the right. Due to the p-type doping, the valence band stays approximately equidistant to the Fermi energy E_F (neglecting the change in the valence band density of states N_V), which is constant in thermodynamic equilibrium.[1] Due to the change in bandgap energy, the conduction band energy will change strongly and provide a built-in drift field for electrons, which in this schematic will be accelerated from right to left.

We will later use such a structure to accelerate the electrons in the base of a heterostructure bipolar transistor.

- Or we may abruptly change the bandgap by an abrupt modification of the stoichiometry (see Figure 1.30). In this case, we use an n-type semiconductor material, so the conduction bands remain approximately lined up horizontal (save for the stoichiometry-induced change in the conduction band density of states), but the change in bandgap results in a significant additional energy barrier for holes, which will keep them from moving from region 1 into region 2. This *carrier confinement* is at the heart of any semiconductor laser structure, and will also be used in heterostructure bipolar transistors. Note that a comparable energy barrier does not exist for electrons – an energy barrier has been created for one carrier species only, a feat possible only through the introduction of semiconductor heterostructures.

[1] The picture is a simplification because changing the stoichiometry also changes the density of states in the valence band, so there will be some variation in the E_F to E_V distance, which has been omitted.

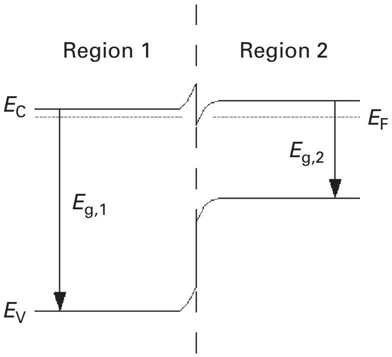


Fig. 1.30 Energy band diagram of an abrupt transition between two materials in a semiconductor heterostructure.

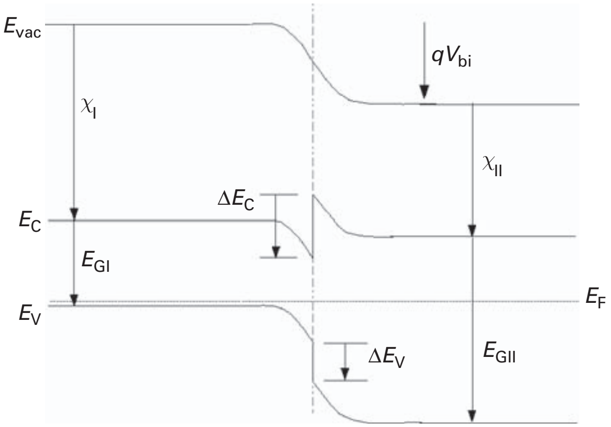


Fig. 1.31 Constructing heterostructure band diagrams using Anderson's rule.

1.19.1 Constructing heterostructure band diagrams

To efficiently use heterostructure band diagrams in the understanding of high-speed electronic and optoelectronic devices, we have to be able to construct their band diagrams. A simple procedure shall be described here.

It uses a modification of *Anderson's rule*, which has been a proven technique to draw the band diagram of the important SiO_2–Si interface. Anderson's rule postulates that the vacuum energy level is continuous and uses the electron affinities, i.e. the energetic spacing between the vacuum level and the conduction band, to calculate the band alignment in an abrupt heterostructure. The electron affinity χ is material-dependent.

When constructing the band diagrams, see Figure 1.31; we start out with the Fermi energy E_F, which in thermodynamic equilibrium is constant throughout the structure. We next draw conduction and valence bands *far away from the interface* between the two materials, which requires knowledge of the doping type and concentration, the bandgap energy and the densities of state as described in the introductory review section.

Next, we use knowledge of the electron affinities χ in the two materials to draw the vacuum level, again far away from the interface. Poisson's equation can be used to calculate the exact potential in the interfacial region, which is used to draw the continuous vacuum energy level. The material properties remain constant right up to the interface, so we can draw the conduction and valence bands to be perfectly parallel to the vacuum level.

The resulting conduction and valence bands show *discontinuities* at the interface between the two materials, in direct consequence of the the different electron affinities and the continuity of the vacuum energy level.

The built-in potential V_{bi} can be calculated as follows (refer again to Figure 1.31). We assume here for simplicity's sake that the Boltzmann equation can be used in lieu of the Fermi–Dirac function:

$$q \cdot V_{\mathrm{bi}} = \chi_{\mathrm{I}} - \chi_{\mathrm{II}} + E_{\mathrm{G,I}} + kT \cdot \ln\left(\frac{N_{\mathrm{V,I}} N_{\mathrm{C,II}}}{N_{\mathrm{A,I}} N_{\mathrm{D,II}}}\right). \tag{1.118}$$

Assume that you externally apply a voltage $-V_{\mathrm{bi}}$ to the structure, which compensates the built-in potential – the energy bands would then be constant within the two regions of the heterostructure (flatband condition); now you easily recognise that

$$\Delta E_{\mathrm{C}} = \chi_{\mathrm{I}} - \chi_{\mathrm{II}} \tag{1.119}$$

and

$$\Delta E_{\mathrm{V}} = E_{\mathrm{G,I}} - E_{\mathrm{G,II}} - \Delta E_{\mathrm{C}} = \Delta E_{\mathrm{G}} - \Delta E_{\mathrm{C}}. \tag{1.120}$$

However, in reality, the experimentally determined band discontinuities of heterostructures do not agree well with the values predicted using the electron affinities. This has to do with the different interface conditions between a free surface (used to determine the electron affinities) and a semiconductor heterostructure.

Anderson's rule can still be used, however. Note that only the difference of the electron affinities matters, not their absolute values. Hence, you can place the vacuum level at an arbitrary distance from the conduction band when drawing the band diagram, provided that the difference in the hypothetical electron affinities agrees with the *experimentally determined* ΔE_{C}.

1.19.2 Band line-up

Abrupt heterostructures can have three fundamental band alignments – refer to Figure 1.32 for a schematic representation:

(i) In a *type 1* heterojunction, the conduction band of the material with the lower bandgap is below the conduction band of the material with the higher bandgap, but the valence band of the lower-bandgap material is above the valence band of the higher-bandgap material. The smaller bandgap is hence fully within the larger bandgap.

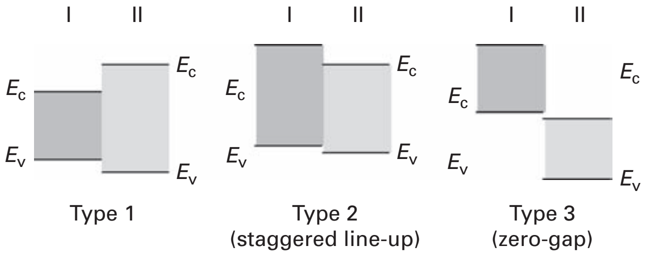


Fig. 1.32 Schematic representation of heterostructure line-ups.

(ii) In a *type 2* heterojunction, the conduction band and the valence band of one material are below their counterparts in the other material. This is sometimes also referred to as a *staggered line-up*.

(iii) Finally, in a *type 3* heterojunction, the valence band in one material is above the conduction band in the other. This is called a *zero-gap* configuration.

In current practical devices, the type 1 line-up is by far the most common.

1.19.3 Lattice mismatch

As already observed in Figure 1.28, changing the stoichiometry will generally also modify the lattice constant. When combining materials with different lattice constants, the mismatch will create strong mechanical strain at the interface (experienced as tensile strain by the material with the smaller lattice constant and as compressive strain by other materials).

Please refer to Figure 1.33. We shall visualise in a schematic fashion the problem of lattice mismatch, taking the important Si/SiGe heterostructure as an example.

In this example, a thin SiGe layer shall be sandwiched between two thick Si layers.[2] Provided that the thickness of the SiGe layer remains below a certain *critical thickness*, all of the lattice mismatch can be compensated for by *elastically* straining the thin layer (and a small fraction of the neighbouring layers). This case is also called the *pseudomorphic* case, a term which we will encounter later on, e.g. in the discussion of modern field effect transistor (FET) structures.

If, however, the strained layer thickness is significantly increased beyond the critical thickness, the mechanical forces at the interface become so large that they are able to break crystalline bonds – a crystal defect is created which will have detrimental effects in both optoelectronic and electronic structures and is hence to be avoided (but for a very few special cases, where this *plastic* strain relaxation is used deliberately in areas of the device devoid of mobile charge carriers).

Hence, the critical thickness is a parameter which must be carefully obeyed in heterostructure device design. It depends on both the lattice mismatch and the elements forming the heterostructure.

[2] 'Thin' here means that the whole layer can be deformed by the mismatch-generated strain, while 'thick' means that the largest part of the layer remains unstrained.

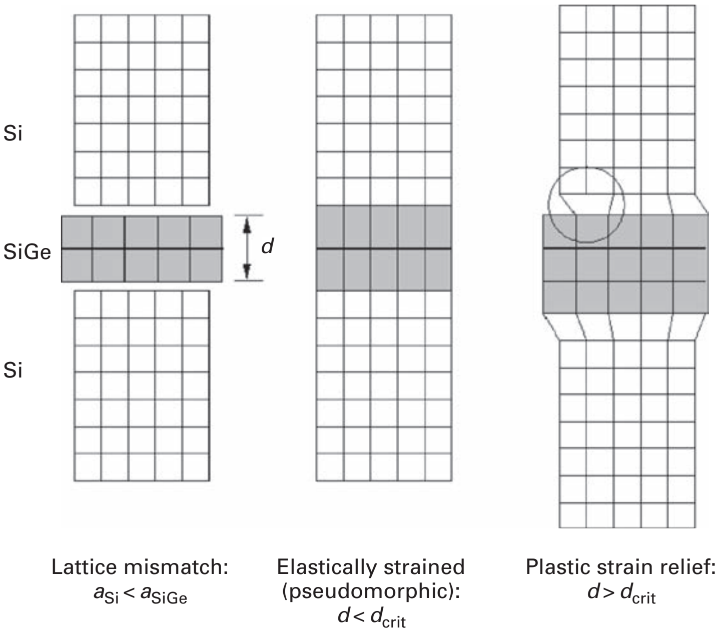


Fig. 1.33 Schematic representation of lattice mismatch in heterostructures.

This introductory section on heterostructures concludes with what 2000 Nobel laureate Herbert Kroemer has called the *Central Design Principle* of semiconductor heterostructures:

> Heterostructures use energy gap variations in addition to electric fields as forces acting on holes and electrons to control their distribution and flow [7].

Our future treatment of high-speed electronic and optoelectronic devices will only exemplify this fundamental observation.

1.20 Silicon–germanium heterostructures

Until the late 1980s, practical electronic and optoelectronic devices, if they included compound semiconductor materials, were dominated by materials combining elements from the third and fourth columns of the periodic table of elements, such as GaAs or InP. However, compound semiconductors may also be formed from elements in the fourth column – elements which already are semiconductors. One example is SiC, which has been investigated extensively as a blue-light emitter. As an aside, SiC under its old trademark of 'carborundum' was the first semiconductor material for which light emission was ever observed and is finding renewed interest for high-power transistors [11].

The currently most important group IV–IV semiconductor material, however, is the silicon–germanium alloy which shall receive special treatment here commensurate with its importance.

Table 1.5 Basic properties of silicon and germanium

Property	Silicon	Germanium	Unit
Lattice type	Diamond	Diamond	
Lattice constant	0.5431	0.5657	nm
Direct band gap	3.40	0.80	eV
Indirect band gap	1.11	0.664	eV
Indirect bandgap direction	{100}	{111}	
Relative dielectric constant	11.9	16.2	

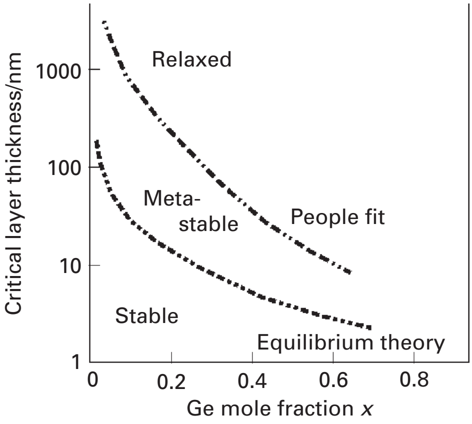


Fig. 1.34 Critical thickness of a strained SiGe layer on Si.

Table 1.5 [13] summarises some fundamental properties of silicon and germanium. As we see, both are indirect semiconductors. Of special interest is the difference in band gap, which will allow to build efficient heterostructure devices. However, the significant difference of the relative dielectric constant (and consequently the refractive index) has also been used in photonic waveguiding devices.

We also note that the lattice constant is quite different – the ratio of lattice constants is

$$\frac{a_{\mathrm{Ge}}}{a_{\mathrm{Si}}} = 1.042$$

for a mismatch of 4.2%.

Therefore, any SiGe heterostructure will be strained, and the strain has significant consequences. In most cases, a pseudomorphic heterostructure free from crystal defects will be desired, and hence the critical layer thickness has to be obeyed.

There is, however, already some discussion about the value of the critical layer thickness. Refer to Figure 1.34 adapted from [5], which plots the critical thickness of a strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ layer on relaxed silicon. This is the case found in Si/SiGe heterostructure bipolar transistors (HBTs), where the base is formed from a SiGe alloy, while the emitter and collector layers are silicon.

There are two substantially different curves for the critical thickness as a function of the germanium mole fraction x.

(i) The equilibrium theory limit constitutes a safe upper limit in that strained layers grown with combinations of Ge content and thickness will remain stable even under high-temperature processing.

(ii) The *People fit* [9] takes experimental data into account. Here, depending on the growth temperature, much higher products of layer thickness and Ge mole fraction will still result in defect-free layers, provided that they do not encounter high-temperature processing after layer deposition.

The region between the equilibrium-theory and People-fit curves is called the *metastable region*. Many practical SiGe devices possess Ge fraction and layer thickness combinations in this region.

The strain also very substantially affects the band structure of SiGe heterojunctions. This is important to understand the concepts of SiGe heterostructure devices further down. In the above example of an elastically strained SiGe layer on a relaxed silicon layer, the heterojunction forms a type 1 interface, i.e. the SiGe band gap is located energetically in the forbidden gap of silicon. Most of the bandgap difference is reflected in the valence band discontinuity on an abrupt heterostructure. If, however, we place a thin strained Si layer on a relaxed SiGe layer, the heterojunction will be of type 2, the staggered line-up configuration, with the Si conduction band below the SiGe conduction band. The resulting conduction band discontinuity will further down be used to construct n-channel Si/SiGe heterostructure field effect transistors (HFETs).

The effect of layer composition has been compiled by Schäffler [12] into a convenient chart, shown in Figure 1.35.

Its use shall be demonstrated using the following example. Suppose you are looking for the band alignment of a strained $\mathrm{Si}_{0.1}\ \mathrm{Ge}_{0.9}$ layer on relaxed $\mathrm{Si}_{0.4}\ \mathrm{Ge}_{0.6}$. In this case, $x_{\mathrm{relaxed}} = 0.9$ and $x_{\mathrm{strained}} = 0.6$. In this case, we read $\Delta E_{\mathrm{C}} = 0$ and $\Delta E_{\mathrm{V}} = +220\,\mathrm{meV}$.

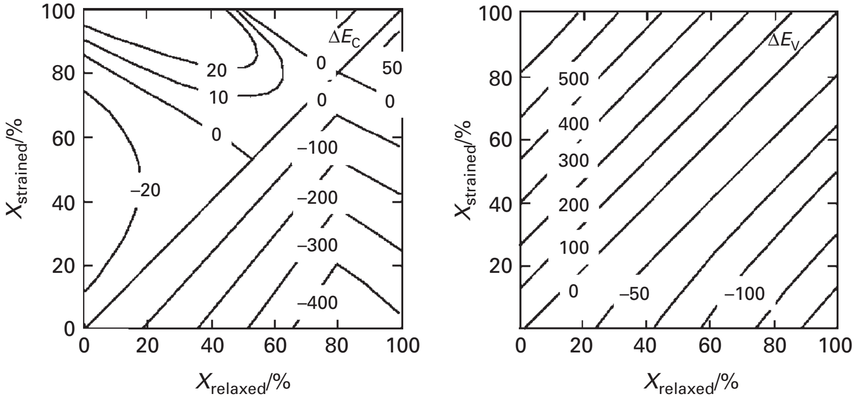


Fig. 1.35 Conduction band (ΔE_{C}) and valence band (ΔE_{V}) discontinuities of $\mathrm{Si}_{1-\mathrm{x}}\mathrm{Ge}_{\mathrm{x}}$ heterostructures, depending on the relaxed-layer and strained-layer Ge mole fraction x. The discontinuity values are listed in meV. Graphs adopted from Schäffler (1995), Ref. 12.

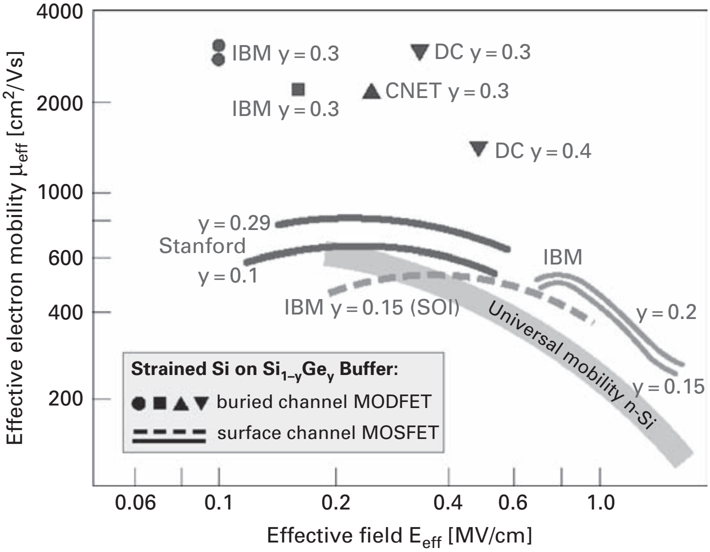


Fig. 1.36 Electron mobility enhancement in strained-Si MOSFET channels [6]. Used with permission from Oerlikon Systems.

Another interesting effect of strain in Si/SiGe heterostructures is that it modifies mobility, an effect investigated for performance enhancement in Si MOSFETs. Here, a thin silicon layer is deposited on top of a relaxed SiGe buffer. The SiGe buffer does not carry any current, but provides a means of introducing a tensile lateral strain into the silicon channel layer above. Depending on the electric field in the channel, a mobility enhancement of 30% or more may result, compared to the Si universal mobility curve (see Figure 1.36). [6]

1.21 Problems

(1) Consider a three-dimensional potential well which is infinite such that the following conditions are satisfied:

$$V(x, y, z) = 0, -a < x < a, -a < y < a, -a < z < a, \tag{1.121}$$

Outside the well

$$V(x, y, z) = \infty. \tag{1.122}$$

Show that the energy is quantised and determine its form.

(2) (a) Determine the amplitudes of the transmitted and reflected waves in the finite potential barrier. Use these results to determine the transmission and reflection coefficients and hence the tunnelling probability (also called the *transmissivity*) given in Equation (1.25).

(b) A potential barrier has width $a = 30$ Å and height $V_0 = 0.5$ eV. An electron is incident on this barrier. Sketch the transmissivity function.

(c) The barrier width is reduced to 5 Å. Sketch the transmissivity function. Comment on the difference as the barrier width is reduced.

(3) For the Kronig–Penney model, obtain the solution of the transcendental equation and show the allowed energy bands (MATLAB is useful in this problem).

(4) A sample of GaAs is doped with a background concentration of $N_A = 3 \times 10^{15}\,\text{cm}^{-3}$ acceptors. Then, $10^{19}\,\text{cm}^{-3}$ donors are added. Find the room temperature concentrations of electrons and holes in the original material and the material that results after the addition of the donors. Draw the band diagram for each of the materials.

(5) A silicon ingot is doped with $10^{17}\,\text{cm}^{-3}$ arsenic atoms. Find the carrier concentrations and the Fermi level at 300° K. Assume complete ionisation of impurity atoms. Draw the band diagram. Show the Fermi level and use the intrinsic Fermi level as the energy reference. Repeat for a temperature of 77° K.

(6) (a) The energy band gap for GaAs at 300 K is 1.42 eV. Assume that the Fermi energy level is at mid-band. Calculate the probability that an energy state at $E = E_c + kT/2$ is occupied by an electron.

(b) What is the probability that the energy state at $E = E_v - kT/2$ is empty?

(c) Using the density of states function $g(E)$, determine the volume density of states between 0 and 2 eV. (Integrate $g(e)dE$.)

(7) A sample of InP is doped with 5×10^{16} tellurium atoms/cm^3 and $2 \times 10^{15}\,\text{cm}^{-3}$ zinc atoms. Calculate the electron and hole concentrations.

(8) A sample of GaAs is illuminated with light of intensity 1.5 mW cm^{-2} at a wavelength of $\lambda = 470$ nm. The area of the illuminated surface is 15 mm^2. If the carrier lifetimes are 10 ns, how many photons are incident on the surface of the GaAs sample? If there is uniform absorption of the photons within 1 μm of the sample surface, determine the equilibrium concentrations of electrons and holes in the dark and the excess carrier concentrations when the sample is illuminated.

(9) (a) An n$^+$–p diode has a long p region. The current I is measured at a constant forward bias of 500 mV. Derive an expression for the fractional change in current $\Delta I/I$ resulting from a temperature change of ΔT. Evaluate this expression at $T = 300$ K. Assume that the diode current is dominated by recombination of electrons in the neutral p region and consider the temperature dependence to be due solely to the factors that have an exponential dependence on temperature.

(b) Suppose that part of the diode current results from recombination in the depletion region. How will this affect the sensitivity of the temperature measurement?

(10) A Silicon p–n junction has $5\times10^{17}\,\text{cm}^{-3}$ donor atoms and $4\times10^{19}\,\text{cm}^{-3}$ acceptor atoms. The junction area is 150 μm^2. Determine the junction capacitance given that the reverse bias is 5 V. How does this change for the same forward bias?

(11) It is found experimentally that a Schottky barrier is formed when Al is evaporated on n-type GaAs with a donor concentration of $4 \times 10^{15}\,\text{cm}^{-3}$. Al has a work function of 4.36 V and GaAs has an electron affinity of 4.07 V.

(a) Determine the built-in voltage V$_{\text{bi}}$, the depletion layer width W and the capacitance at zero bias, assuming no surface states.

(b) If the Fermi level at the surface is pinned so that

$$\begin{aligned} q\phi_0 &= \frac{1}{3}E_\mathrm{g} \\ q\phi_\mathrm{B} &= (E_\mathrm{g} - q\phi_0), \end{aligned}$$

determine the same quantities as in part (a) with a reverse bias of 1 V.

(12) An n^+-GaAs/p-AlGaAs heterojunction is formed with $N_\mathrm{D} = 10^{19}\,\mathrm{cm}^{-3}$ and $N_\mathrm{A} = 5 \times 10^{17}\,\mathrm{cm}^{-3}$. Draw the band diagram and determine the following quantities:

(a) The electric field $\mathcal{E}(x)$ and the potential difference $V(x)$ at the junction.
(b) The built-in potentials on either side of the junction and the total built-in potential, V_bi.
(c) The depletion widths in the two regions.
(d) If a forward bias of 3 V is applied, calculate the depletion widths.
(e) Calculate the capacitance of this diode.
(f) Calculate the current in the diode.

References

[1] Agrawal G. P., Dutta N. K. (1993). *Semiconductor Lasers*, Van Nostrand Reinhold.

[2] Anderson B. L., Anderson R. L. (2005). *Fundamentals of Semiconductor Devices*. McGraw-Hill.

[3] Bhattacharya P. (1997). *Semiconductor Optoelectronic Devices*, 2nd edn. Prentice Hall.

[4] Casey H. C., Panish M. B. (1978). *Heterostructure Lasers Part A: Fundamental Principles*. Quantum electronics – principles and applications, Academic Press.

[5] Gruhle A. (1994). *SiGe Heterojunction Bipolar Transistors of Silicon-based Millimeter Wave Devices*. Springer Verlag, 149–189.

[6] Hackbarth T., Zeuner M., König U. (2002). The future of SiGe beyond HBT applications. *Uniaxis chip Magazine*, July 10–12.

[7] Kroemer H. (1982). Heterostructure bipolar transistors and integrated circuits. *Proceedings of the IEEE 70*. 1 (January), 13–25.

[8] Neaman D. A. (2003). *Semiconductor Physics and Devices: Basic Principles*. McGraw-Hill.

[9] People R., Bean J. C. (1985). Calculation of critical layer thickness versus lattice mismatch for $\mathrm{Ge}_x\mathrm{Si}_{1-x}$/Si strained-layer heterostructures. *Appl. Phys. Lett. 47*, 322–324.

[10] Pierret R. F. (2003). *Advanced Semiconductor Fundamentals*. Modular Series on Solid State Devices, Vol. VI. Prentice Hall.

[11] Round H. J. (1907). A note on Carborundum. *Elect.World 49,* 10, 309.

[12] Schäffler F. (1995). *Properties of Strained and Relaxed Silicon Germanium* of EMIS Data Reviews. 12. IEE INSPEC, 10–12.

[13] Schäffler F. (1997). High mobility Si and Ge structures. *Semicond. Sci. Technol. 12,* 10, 1515–1549.

[14] Singh J. (1994). *Semiconductor Devices – An Introduction*. McGraw-Hill.

[15] Streetman B., Banerjee S. (2000). *Solid State Elecronic Devices*, 5th edn. Prentice Hall.

2 Electronic devices

2.1 Executive summary

This chapter introduces the active devices commonly used in high-speed electronics. It starts with a discussion of the metal–semiconductor field effect transistor, or MESFET – historically the oldest FET concept, which for decades was the most prominent device in microwave electronics. Its pitfalls led to the development of an advanced transistor structure, the high electron mobility transistor (HEMT). It incorporates heterostructures to gain additional freedom in device design. HEMTs mostly replaced MESFETs in micro- and millimetre-wave applications.

Metal-oxide-semiconductor field effect transistors (MOSFETs), which dominate digital electronics, are rapidly making inroads at microwave and even millimetre-wave frequencies. They will be discussed as well, and we will recognise similarities between HEMTs and MOSFETs in the physics of the intrinsic transistor.

Finally, bipolar junction transistors (BJTs) will be introduced, showing how a dilemma in the optimum design of the base layer led to the invention of the heterojunction bipolar transistor (HBT) – again, heterostructures come to the rescue.

For all these components, the chapter will discuss their fundamental physical operation, non-ideal and parasitic effects, and linear and non-linear models, as well as examples in several material systems.

2.2 MESFET

2.2.1 Introduction and current control mechanism

The **m**etal–**s**emiconductor **f**ield **e**ffect **t**ransistor (MESFET) is conceptually the simplest of the commonly used transistor structures and shall therefore be discussed here first. The fundamental idea is quite straightforward: the current flowing through a slab of semiconductor material (from now on called the *channel*) depends on three fundamental parameters for a given externally applied voltage:

(i) velocity of charge carriers,
(ii) density of charge carriers,
(iii) the geometric cross-section the carriers flow through.

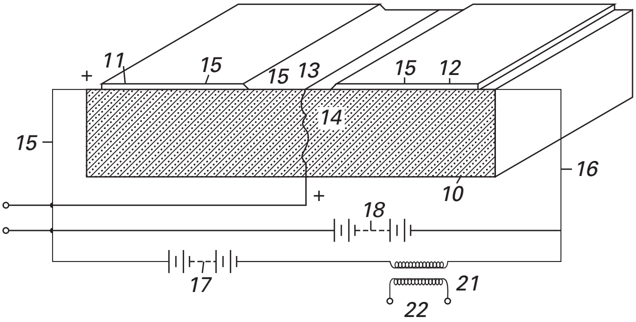


Fig. 2.1 Lilienfeld’s FET concept, from his US patent application in 1926.

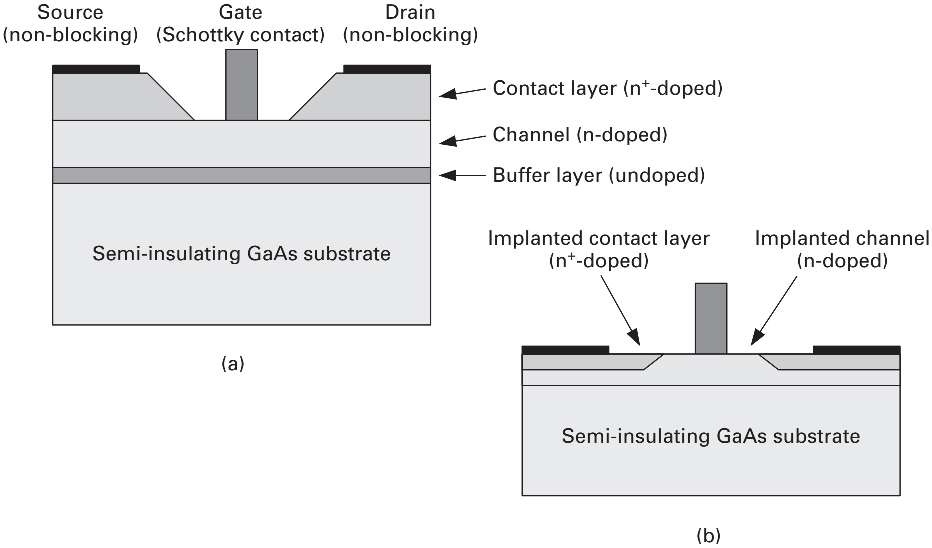


Fig. 2.2 Simplified cross-section of a MESFET with (a) an epitaxially grown channel, (b) fabricated using ion implantation.

While the carrier velocity will depend on the local electric field, in the simplest case the density of charge carriers is given by the doping concentration. The channel cross-section can be influenced externally, if we constrict the current flow using the depletion region of a diode. This method was recognised very early and is the object of a patent filed in 1926 by Julius Edgar Lilienfeld [35]. Lilienfeld’s concept (see Figure 2.1), already used a metal–semiconductor junction to control the current flow, but was never realised. The practical realisation of the MESFET is predated by the silicon junction field effect transistor (JFET), which uses a p–n diode as the controlling element and was first described by Shockley [57].

Figure 2.2 shows two somewhat simplified cross-sections of what a MESFET looks like. The layer structure in Figure 2.2(a) is defined by epitaxial growth. Above a semi-insulating substrate, a thin undoped buffer layer is grown to improve the interface quality, then the channel layer follows whose doping concentration and thickness are very important design parameters, as we will shortly see. Above it, we find a highly doped contact layer, intended to improve the formation of non-blocking contacts and to reduce series resistances between the source and drain contacts and the channel region, but whose exact thickness and doping concentration have no bearing on the fundamental properties of the transistor. Below the gate contact, the contact layer is etched away to allow the blocking Schottky contact to contact the channel layer directly.

Figure 2.2(b) shows a very similar structure; only now the differently doped semiconductor regions are formed by ion implantation. This results in lower cost, however; the lattice damage caused by the ion bombardment will negatively impact carrier velocity and also lead to an increase in low-frequency noise. This will not be discussed in detail here.

In both cases, it is assumed that the carrier species in the channel are electrons (n-channel), as this is the more common variant; however, p-channel devices can be fabricated with equal ease.

While a MESFET can be structured on many different semiconductor materials, only devices fabricated in GaAs and in SiC are commercially relevant. The GaAs MESFET was, for many years, the mainstay of microwave solid-state electronics and shall be discussed here, while the SiC MESFET with its excellent thermal properties and high breakdown voltages is used predominantly in power amplifiers for mobile phone base stations.

For the benefit of clarity and to obtain analytic expressions, we will simplify the structure even further. Figure 2.3 shows the three-dimensional view of the simplified structure. First of all, note the coordinate system which will be used similarly throughout. The x axis is parallel to the 'long' extension of the gate stripe. The y axis is

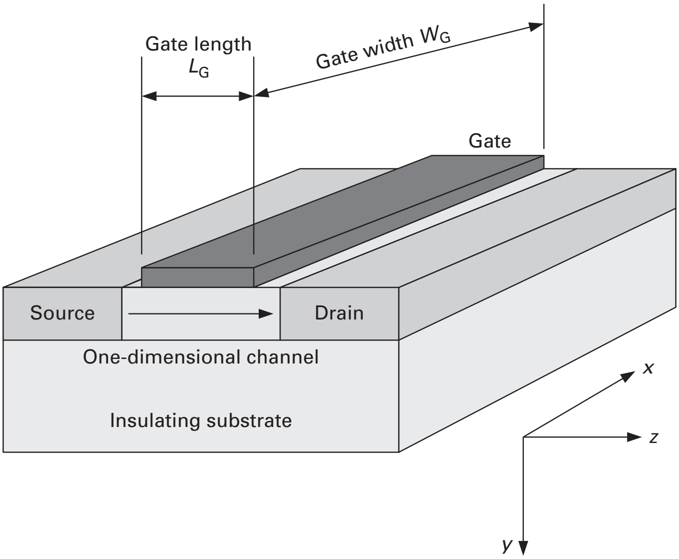


Fig. 2.3 Simplified MESFET structure used in the analytic calculations.

perpendicular to the semiconductor surface, while the z axis is parallel to the surface in the direction of the 'short' extension of the gate. The 'long' gate dimension in x-direction is called the *gate width* W_G, while the *gate length* L_G is the extension in the z direction.

We assume now that the channel is one-dimensional – the electric field in the channel has only a z component. To neglect the electric field in the x direction is generally justified as $W_G \gg L_G$, but to neglect the electric field in the channel in the y direction is a simplification.

Another important simplification in the channel is the *gradual channel approximation*. In general, current flow in semiconductor devices can be driven by the electric field (this is the drift current) or by concentration gradients – this is the diffusion current or a combination of both. Here, we assume that the drift current entirely dominates and the diffusion current can be neglected.

The gate electrode forms a blocking contact with the semiconductor layer under the channel, a *Schottky diode*, which was discussed already in Chapter 1.

Figure 2.4 shows only the cross-section of the device. The source electrode shall be the reference potential, hence $V_S = 0$.

The gate electrode potential with respect to the source is the gate-source voltage V_{GS}. In an n-channel device, where the channel layer is n-doped, it will generally be negative to maintain the gate-channel diode in a blocking state. The drain-source voltage V_{DS} in an n-channel device will be positive.[1]

The extension of the space charge region, shown schematically in Figure 2.4, depends on the local gate-channel voltage V_G. We find for $V_G(z)$:

$$V_G(z) = V_{GS} - V(z), \tag{2.1}$$

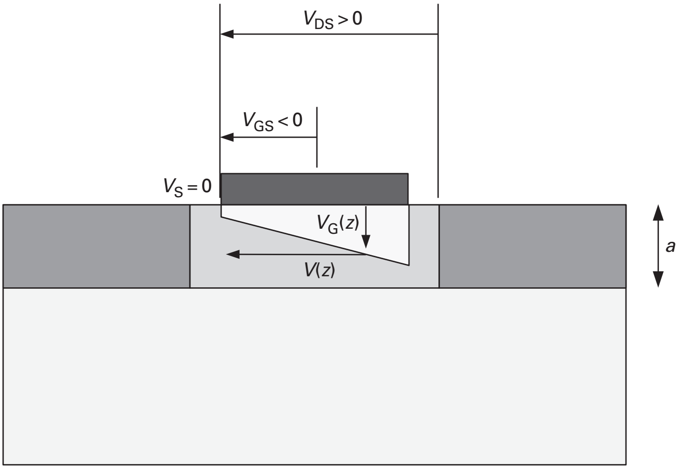


Fig. 2.4 MESFET channel with space charge region for small drain-source voltages.

[1] Because our structure is symmetric with respect to the non-blocking contacts, this defines the drain – in an n-channel device, the drain is the contact with the higher potential.

where $V(z)$ is the voltage drop in the channel between point z and source. As the drain is at a higher potential than the source, $V(z) > 0$, the gate-channel voltage becomes more negative as z increases.

At point z, the extension of the space charge region is

$$h(z) = \sqrt{\frac{2\epsilon_\mathrm{s}[V_\mathrm{bi} - V_\mathrm{G}(z)]}{qN_\mathrm{D}}} = \sqrt{\frac{2\epsilon_\mathrm{s}[V_\mathrm{bi} - V_\mathrm{GS} - V(z)]}{qN_\mathrm{D}}}. \tag{2.2}$$

with N_D the channel doping concentration, assumed to be constant throughout the channel.

2.2.2 Drain current using a constant-mobility assumption

Let us first consider small V_DS, such that $h(z) < a$, with a the channel thickness for all $0 < z < L_\mathrm{G}$ – there is always an undepleted part of the channel remaining. We will now calculate the channel current.

The channel current is always calculated in the same fashion: by multiplying the moved charge density (here, qN_D), the cross-section through which it is moved (here $W_\mathrm{G}[a - h(z)]$) and the charge velocity. For low fields, the charge velocity can be calculated from the electron mobility μ_n and the local electric field, here $dV(z)/dz$. Therefore, we find an expression for the channel current as a function of the z coordinate:

$$I(z) = qN_\mathrm{D}W_\mathrm{G}[a - h(z)]\mu_\mathrm{n}\frac{dV(z)}{dz}. \tag{2.3}$$

As a consequence of Kirchhoff's law, we know that the charge current entering at the source will be equal to that leaving at the drain – this is called *current continuity*. It allows us to eliminate the z dependence of the current through a simple mathematical trick.

As $I(z) = \mathrm{const} = I_\mathrm{D}$, obviously $\int_0^{L_\mathrm{G}} I(z)dz = I_\mathrm{D}L_\mathrm{G}$.

Consider further that

$$h^2(z) = \frac{2\epsilon_\mathrm{s}}{qN_\mathrm{D}}[V_\mathrm{bi} - V_\mathrm{GS} + V(z)].$$

Differentiating both sides with respect to z leads to

$$2h(z)\frac{dh(z)}{dz} = \frac{2\epsilon_\mathrm{s}}{qN_\mathrm{D}}\frac{dV(z)}{dz},$$

and finally

$$\frac{dV(z)}{dz} = \frac{qN_\mathrm{D}}{\epsilon_\mathrm{s}}h(z)\frac{dh(z)}{dz}.$$

Through parameter substitution, we find

$$I_\mathrm{D} = \frac{1}{L_\mathrm{G}}\int_{z=0}^{z=L_\mathrm{G}} I(z)dz = \frac{q^2N_\mathrm{D}^2W_\mathrm{G}\mu_\mathrm{n}}{\epsilon_\mathrm{s}L_\mathrm{G}}\int_{h(0)}^{h(L_\mathrm{G})} h(z)[a - h(z)]dh.$$

As $V(0) = 0$, $h(0) = \sqrt{\frac{2\epsilon_\mathrm{s}}{qN_\mathrm{D}}(V_\mathrm{bi} - V_\mathrm{GS})}$. Incidentally, the necessary gate-source voltage to fully close the channel at the source end is the *pinch-off voltage* V_P:

$$V_\mathrm{P} = V_\mathrm{bi} - a^2\frac{qN_\mathrm{D}}{2\epsilon_\mathrm{s}}. \tag{2.4}$$

Using V_P, we can write Equation (2.2) in the following form:

$$h(z) = a\sqrt{\frac{V(z) - V_\mathrm{GS} - V_\mathrm{bi}}{V_\mathrm{bi} - V_\mathrm{P}}}. \tag{2.5}$$

The required value $h(L_\mathrm{G})$ is now found very easily – we know that $V(z = L_\mathrm{G}) = V_\mathrm{DS}$ and therefore

$$h(z = L_\mathrm{G}) = a\sqrt{\frac{V_\mathrm{DS} - V_\mathrm{GS} - V_\mathrm{bi}}{V_\mathrm{bi} - V_\mathrm{P}}}. \tag{2.6}$$

We can now finally solve the current integral using the constant-mobility assumption, and find for the drain current:

$$I_\mathrm{D}(V_\mathrm{GS}, V_\mathrm{DS}) = \frac{q^2 N_\mathrm{D}^2 \mu_\mathrm{n} a^3 W_\mathrm{G}}{6\epsilon_\mathrm{s} L_\mathrm{G}} \tag{2.7}$$
$$\left\{\frac{3V_\mathrm{DS}}{V_\mathrm{bi} - V_\mathrm{P}} - 2\frac{(V_\mathrm{DS} - V_\mathrm{GS} + V_\mathrm{bi})^{3/2} - (V_\mathrm{bi} - V_\mathrm{GS})^{3/2}}{(V_\mathrm{bi} - V_\mathrm{P})^{3/2}}\right\}.$$

We had so far assumed that the channel would remain at least partially open. This requires that $h(L_\mathrm{G}) \leq a$. From Equation (2.6) we find that this translates into

$$V_\mathrm{DS} \leq V_\mathrm{GS} - V_\mathrm{P} \equiv V_\mathrm{k}, \tag{2.8}$$

where V_k is the *knee voltage*.

For

- $V_\mathrm{DS} \leq V_\mathrm{k}$ the MESFET is in the *linear regime*, while for
- $V_\mathrm{DS} > V_\mathrm{k}$ it is in the *saturated regime*.

Figure 2.5 shows simulated output current–voltage characteristics for a hypothetical MESFET with a pinch-off voltage $V_\mathrm{P} = -2\,\mathrm{V}$ and a built-in voltage of $V_\mathrm{bi} = 0.7\,\mathrm{V}$, calculated using Equation (2.7). The drain current has been normalised to $q^2 N_\mathrm{D}^2 \mu_\mathrm{n} a^3/(6\epsilon_\mathrm{s} L_\mathrm{G})$.

Note that for very small V_DS, the dependence of I_D on V_DS is almost linear. MESFETs can be used as electronically controllable resistors, e.g. in variable microwave attenuators or in transmit/receive switches.

For $V_\mathrm{DS} \rightarrow V_\mathrm{k}$, the drain current saturates. A common assumption in simple FET models is that for $V_\mathrm{DS} > V_\mathrm{k}$, $I_\mathrm{D}(V_\mathrm{DS} > V_\mathrm{k}) = I_\mathrm{D}(V_\mathrm{DS} = V_\mathrm{k}) = \mathrm{const}$.

Conceptually, current continuity in the constant-mobility model requires that for increasing z, the electric field increases to compensate for the decrease in undepleted channel height. Near $V_\mathrm{DS} = V_\mathrm{k}$, $a - h(z) \rightarrow 0$ would imply $dV(z)/dz \rightarrow \infty$ at the drain end, which is a fundamental flaw of this model. It is still valuable to investigate MESFET behaviour at low V_DS.

2.2.3 Constant-velocity approximation

In all semiconductor materials, the assumption that the carrier velocity increases linearly with increasing electric field, i.e. that mobility is a constant, only holds for small electric fields. For large electric fields, the carrier velocity becomes independent of the electric

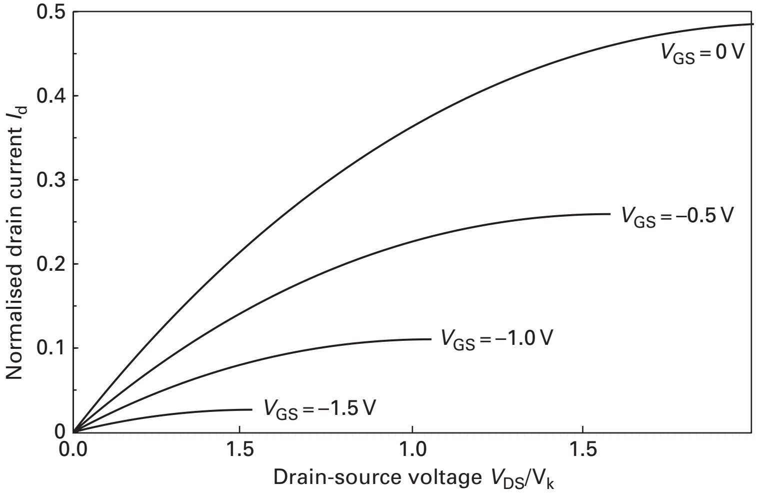


Fig. 2.5 Simulated output I–V characteristics of a MESFET using the constant-mobility assumption ($V_P = -2\,V$, $V_{bi} = 0.7\,V$).

field (in good approximation); its value is then called the *drift saturation velocity*, $v_{sat,n}$ for electrons or $v_{sat,p}$ for holes, respectively.

Let us now boldly assume that the charge carriers in the channel reach their drift saturation velocity immediately after entering the channel at the source side.

The channel current in our n-channel MESFET now becomes

$$I(z) = qN_D v_{sat,n} W_G[a - h(z)] = \text{const} = I_D,$$

due to the current continuity requirement. As the carrier velocity is now constant, this implies that the channel height must also be constant: $h(z) = \text{const} = h$. Hence,

$$I_D = qN_D v_{sat,n} W_G(a - h). \tag{2.9}$$

The extension of the space charge region can be easily calculated at $z = 0$ for a homogeneous channel doping profile:

$$h = h(0) = a\sqrt{\frac{V_{bi} - V_{GS}}{V_{bi} - V_P}}. \tag{2.10}$$

Inserting this into Equation (2.9), we find

$$I_D = qN_D v_{sat,n} W_G a\left(1 - \sqrt{\frac{V_{bi} - V_{GS}}{V_{bi} - V_P}}\right). \tag{2.11}$$

The value for $V_{GS} = 0$ is referred to as I_{DSS}:

$$I_{DSS} = qN_D v_{sat,n} W_G a\left(1 - \sqrt{\frac{1}{1 - \frac{V_P}{V_{bi}}}}\right). \tag{2.12}$$

The constant-velocity model is a priori only valid for high electric fields in the channel, in the saturation region of MESFET operation ($V_{DS} > V_k$). For very small

V_{DS}, Equation (2.7) still holds, and the electric field will not ‘jump’ close to source, in every case there will be a region close to source where the constant mobility approximation is more appropriate. More realistic models of MESFET operation will therefore have to combine the constant-velocity and constant-mobility approaches, as was first pointed out by Pucel, Haus and Statz [46]. This is, however, beyond the scope of this introduction.

In a technical MESFET with short gate length and in saturation region, the constant-mobility regime is restricted to an area very close to source, and the majority of the channel is velocity saturated. Equation (2.7) is, therefore, a good approximation for $I_{\mathrm{D}}(V_{\mathrm{GS}} > V_{\mathrm{P}})$ in saturation.

In practical cases, the onset of saturation is not due to $h(L_{\mathrm{G}}) \rightarrow a$, the channel pinching off at the drain end, but due to the onset of velocity saturation in the channel. This occurs much earlier, and hence the FET will pass from the linear to the saturated regime at significantly lower V_{DS} than predicted by the constant-mobility model.

The discussion so far was restricted to MESFETs with constant doping concentration in the channel. Often, however, N_{D} varies in the channel in the y direction. Two important examples are as follows:

(i) The *ion-implanted MESFET* (see Figure 2.2(b)). Here, the doping concentration varies according to

$$N_{\mathrm{D}}(y) = \frac{Q}{\sqrt{2\pi}\sigma} \exp\left[-\left(\frac{y - R_{\mathrm{P}}}{\sqrt{2\pi}\sigma}\right)^2\right],$$

where Q is the implanted dose, σ the standard deviation and R_{P} the projected range.

(ii) The *pulse-doped MESFET*, where only a fraction of the channel is highly doped (see Figure 2.6). The discussion of the pulse-doped MESFET is interesting because it has a distribution of mobile carriers similar to that of the HEMT which will be discussed further down.

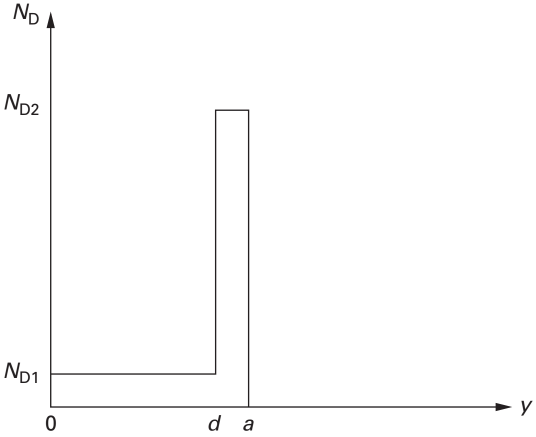


Fig. 2.6 Channel doping profile of a pulse-doped MESFET.

The general procedure to handle these non-uniform doping profiles is as follows:

Poisson's equation is used to obtain a relationship between the potential in the y direction and the space charge distribution $N_D(y)$:

$$\frac{d^2V(y)}{dy^2} = -\frac{q}{\epsilon_s}N_D(y).$$

Integrating Poisson's equation twice yields the required relationship between the gate-channel voltage and the extension of the space charge region. The current is now found by integrating the free carrier concentration over the undepleted channel cross-section:

$$I_D = q\ v_{sat,n}W_G \int_{h(z)}^{a} N_D(y)dy.$$

In the case of a pulse-doped MESFET [40] and $N_{D2} \gg N_{D1}$, the calculation yields

$$I_D = I_{DSS}\left[1 - \frac{\sqrt{1+\left(\frac{a^2}{d^2}-1\right)\frac{(V_{bi}-V_{GS})}{(V_{bi}-V_P)}}-1}{\frac{a}{d}-1}\right]. \tag{2.13}$$

Figure 2.7 compares the transfer characteristics $I_D = f(V_{GS})$ for a homogeneously doped MESFET and a pulse-doped MESFET with $a/d = 1.1$, i.e. where only 10% of the channel is highly doped. The drain current is normalised to the respective I_{DSS}, which will be different in both cases. Figure 2.7 should not be read suggesting that the pulse-doped MESFET has a lower transconductance than the homogeneously doped MESFET.

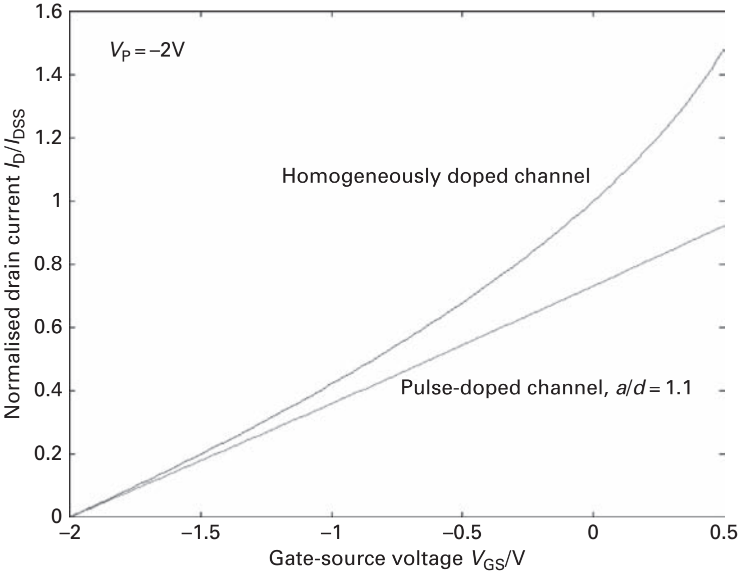


Fig. 2.7 Transfer characteristics $I_D = f(V_{GS})$ for homogeneously and pulse-doped MESFETs.

A striking difference is that $I_D = f(V_{GS})$ is very linear for the pulse-doped MESFET, which is an important advantage from the circuit designer's point of view.

All of the above models consider the drain current in the saturation regime to be independent of the drain-source voltage. In reality, however, I_D depends weakly on V_{DS} there.

The predominant reason for the $I_D = f(V_{DS})$ behaviour in the saturated regime is the scattering of charge carriers into the buffer/substrate layer under the channel. For a semi-insulating substrate, where the Fermi level is near mid-gap, the potential barrier between the channel and the substrate is approximately $E_G/2$ and may be overcome by electrons with sufficient kinetic energy. These electrons may produce two different effects, both of which lead to an $I_D = f(V_{DS})$ dependence:

(i) They may lead to a parasitic conduction current through the buffer or the substrate, adding to the channel current.
(ii) They may be captured by crystal faults or impurities in the buffer or the substrate, which act as charge traps. The resulting modification of charge below the channel influences the channel cross-section, just as a gate electrode would ('backgating').

The latter effect leads to a pronounced frequency dependence of the $I_D = f(V_{DS})$ behaviour.

2.2.4 Large-signal CAD model

For circuit design applications, the physical models considered so far are not convenient. For once, it would be useful to have a model which describes the full range of operation in one closed formula. More importantly, the physical design parameters such as channel thickness and doping concentration are often not accessible to the circuit designer.

Large-signal CAD models, therefore, are empirical in nature and have extractable parameters which can be determined from measurements on the final device.

An early empirical model which may be used for MESFETs in the saturation regime is the 'square-law' JFET model implemented in SPICE:

$$I_D(V_{GS}) = \beta(V_{GS} - V_P)^2, \tag{2.14}$$

where

$$\beta = \frac{I_{DSS}}{V_P^2}.$$

It fits the transfer characteristics of the constant-mobility model quite well at $V_{DS} = V_k$.

The model can be made to fit to non-parabolic transfer characteristics through a modification introduced by Statz *et al.* [58]:

$$I_D(V_{GS}) = \frac{\beta(V_{GS} - V_P)^2}{1 + \alpha(V_{GS} - V_P)}. \tag{2.15}$$

The linear region can be elegantly included in a closed form through the use of a hyperbolic tangent function, as was first pointed out by Curtice [10]:

$$I_D(V_{GS}, V_{DS}) = I_S(V_{GS}) \tanh(\gamma V_{DS}), \tag{2.16}$$

where $I_S(V_{GS})$ is the drain current according to Equation (2.15).

The dependence of the drain current on the drain-source voltage in the saturation regime can be introduced through an additional $1 + \lambda V_{DS}$ term. We arrive finally at the common CAD model expression for the MESFET:

$$I_D(V_{GS}, V_{DS}) = \frac{\beta(V_{GS} - V_P)^2}{1 + \alpha(V_{GS} - V_P)} \tanh(\gamma V_{DS})(1 + \lambda V_{DS}). \tag{2.17}$$

Capacitance model

We will now leave the quasi-static realm and introduce capacitances. The discussion will be restricted first to the intrinsic FET structure, while parasitic capacitances will be introduced in context with the small-signal equivalent circuit.

Assume a MESFET with a homogeneously doped channel region with $V_{DS} = 0$, which implies a constant extension of the gate space charge region, h. The charge on the gate electrode counter-balances the charge in the channel. In this case (n-channel) the gate charge is positive:

$$Q_{G0} = qN_D W_G L_G (a - h) = -qN_D W_G L_G a \left(1 - \sqrt{\frac{V_{bi} - V_{GS}}{V_{bi} - V_P}}\right), \tag{2.18}$$

using Equation (2.5) and $V(z) = 0$ due to $V_{DS} = 0$.

The gate-channel capacitance for $V_{DS} = 0$ can now be calculated as the first derivative of the gate charge with respect to the gate-channel voltage (which is identical to V_{GS} as $V_{DS} = 0$):

$$C_{GC} = \frac{\delta Q_{G0}}{\delta V_{GS}} = C_0 \sqrt{\frac{V_{bi} - V_P}{V_{bi} - V_{GS}}}, \tag{2.19}$$

where

$$C_0 = q\frac{N_D W_G L_G a}{2(V_{bi} - V_P)}.$$

For $V_{DS} > 0$, the Meyer capacitance approach originally developed for MOSFETs [38] is often used, which distinguishes between the linear ($V_{DS} < V_k$) and the saturated ($V_{DS} > V_k$) regimes:

- For $V_{DS} < V_k$,

$$C_{GS} = \frac{2}{3} C_{GC} \left[1 - \left(\frac{V_k - V_{DS}}{2V_k - V_{DS}}\right)^2\right]$$

$$C_{GD} = \frac{2}{3} C_{GC} \left[1 - \left(\frac{V_k}{2V_k - V_{DS}} \right)^2 \right].$$

- For $V_{DS} > V_k$,

$$C_{GS} = \frac{2}{3} C_{GC}$$

$$C_{GD} = 0.$$

The intrinsic $C_{GD} = 0$ in the saturated regime means that the gate-drain voltage has no influence on the channel charge.

Parasitic circuit elements

Our discussion so far was restricted to the intrinsic transistor, more precisely to the channel region. A realistic transistor model will also have to take extrinsic circuit elements into account (see Figure 2.8). The most important ones are:

(i) The source resistance R_S and the drain resistance R_D. They contain contributions from the semiconductor–metal contact at the source, and the semiconductor regions between the channel and the source and drain contacts, respectively. The source contact is most important, because it has a direct impact on the controlling gate-source voltage. Because the gate current can be generally neglected, the intrinsic gate-source voltage V_{GS} is related to the externally applied $V_{GS,e}$ as follows:

$$V_{GS} = V_{GS,e} - R_S I_D.$$

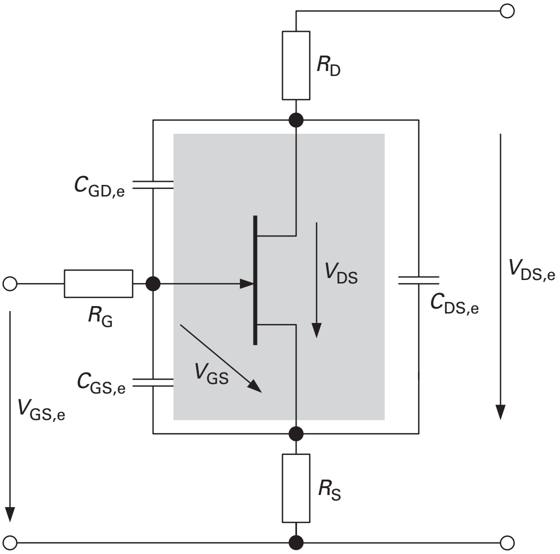


Fig. 2.8 Extrinsic circuit elements in the MESFET. The transistor symbol in the shaded box is the intrinsic transistor discussed so far.

(ii) The gate resistance, which is due to the series resistance of the gate electrode (in x direction in Figure 2.3). This can be a problem especially in modern devices with very small gate length L_G, typically $\leq 0.25\,\mu\text{m}$.

(iii) The parasitic capacitances are $C_{GS,e}$, $C_{GD,e}$ and $C_{DS,e}$. They are mostly due to the contact and interconnect metallisations within the transistor structure. $C_{GD,e}$ is of particular importance because it is in a feedback path in the frequently used common-source transistor configuration where it will give rise to the so-called *Miller capacitance*, and also may lead to amplifier instability.

2.2.5 Small-signal equivalent circuit

Introduction: small-signal versus large-signal model

The physical behaviour of electronic devices is generally non-linear, as has been seen above. However, in many cases, we only deal with very small perturbations around a given bias point, so that the non-linear functions can be approximated by linear ones, dramatically simplifying the calculation effort.

For example, the non-linear dependence of the drain current on the gate-source and drain-source voltages, $I_D(V_{GS}, V_{DS})$, can be approximated for small perturbations around a bias point $(I_{D,0}, V_{GS,0}, V_{DS,0})$ by a two-dimensional Taylor series, which is aborted after the linear term:

$$i_d = \frac{\delta I_D}{\delta V_{GS}} v_{gs} + \frac{\delta I_D}{\delta V_{DS}} v_{ds} + \cdots . \tag{2.20}$$

The lower-case symbols i_d, v_{gs} and v_{ds} denote small deviations from the bias point:

$$i_d = (I_D - I_{D,0});\; v_{gs} = (V_{GS} - V_{GS,0});\; v_{ds} = (V_{DS} - V_{DS,0}).$$

MESFET small-signal equivalent circuit

Refer again to Equation (2.20).

The first partial derivative is the *transconductance* g_m:

$$\frac{\delta I_D}{\delta V_{GS}} |_{V_{GS,0}, V_{DS,0}} \equiv g_m.$$

In saturation, we can use Equation (2.17) to calculate its bias-dependent value:

$$\begin{aligned} g_m &= \frac{\delta}{\delta V_{GS}} \left[\frac{\beta (V_{GS} - V_P)^2}{1 + \alpha (V_{GS} - V_P)} \tanh\,(\gamma V_{DS})(1 + \lambda V_{DS}) \right] \\ &\approx \beta \frac{2(V_{GS,0} - V_P)[1 + \alpha(V_{GS,0} - V_P)] - \alpha(V_{GS,0} - V_P)^2}{\left[1 + \alpha(V_{GS,0} - V_P)\right]^2} \\ &= \beta \frac{\alpha(V_{GS,0} - V_P)^2 + 2(V_{GS,0} - V_P)}{\left[1 + \alpha(V_{GS,0} - V_P)\right]^2}, \end{aligned} \tag{2.21}$$

assuming that $\lambda V_{DS} \ll 1$, and that when sufficiently in saturation, $\tanh\,(\gamma V_{DS}) \rightarrow 1$.

The second partial derivative in Equation (2.20) is the *output conductance* g_{ds}:

$$\frac{\delta I_{\mathrm{D}}}{\delta V_{\mathrm{DS}}}|_{V_{\mathrm{GS},0},V_{\mathrm{DS},0}} \equiv g_{\mathrm{ds}}.$$

In saturation, assuming again that $\tanh(\gamma V_{\mathrm{DS}}) \rightarrow 1$:

$$\begin{aligned} g_{\mathrm{ds}} &= \lambda \frac{\beta(V_{\mathrm{GS},0} - V_{\mathrm{P}})^2}{1+\alpha(V_{\mathrm{GS},0} - V_{\mathrm{P}})} \\ &\approx \lambda I_{\mathrm{D},0}, \end{aligned} \tag{2.22}$$

if we assume also $\lambda V_{\mathrm{DS},0} \ll 1$.

In saturation, the bias-dependent intrinsic gate-source capacitance is (see p. 57):

$$C_{\mathrm{GS,i}} = \frac{2}{3} C_{\mathrm{GC}} = q \frac{N_{\mathrm{D}} W_{\mathrm{G}} L_{\mathrm{G}} a}{3 \cdot \sqrt{(V_{\mathrm{bi}} - V_{\mathrm{P}})(V_{\mathrm{bi}} - V_{\mathrm{GS},0})}}. \tag{2.23}$$

To this, we have to add the extrinsic gate-source capacitance, so that

$$C_{\mathrm{GS}} = C_{\mathrm{GS,i}} + C_{\mathrm{GS,e}}.$$

The gate-drain capacitance has only an extrinsic component (refer again to p. 57):

$$C_{\mathrm{GD}} = C_{\mathrm{GD,e}}.$$

Likewise, C_{DS} is purely extrinsic:

$$C_{\mathrm{DS}} = C_{\mathrm{DS,e}}.$$

Adding the series resistances R_{G}, R_{S} and R_{D}, we arrive at a first small-signal equivalent circuit for the MESFET (see Figure 2.9).

A more complete small-signal equivalent circuit will add two more elements:

(i) the resistance R_{i} which improves the modelling of the non-velocity-saturated part of the channel near the source;
(ii) the domain capacitance C_{DC}.

The domain capacitance accounts for a charge dipole forming at the drain end of the channel. Provided that $R_{\mathrm{i}} \ll 1/(\omega C_{\mathrm{GS}})$, it can safely be omitted as it is absorbed in C_{DS}.

It is important to note that in Figure 2.10, V_{GS} drops over C_{GS} only.

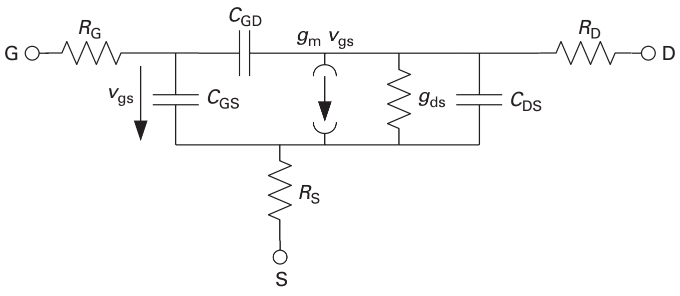


Fig. 2.9 Simple small-signal equivalent circuit of a MESFET.

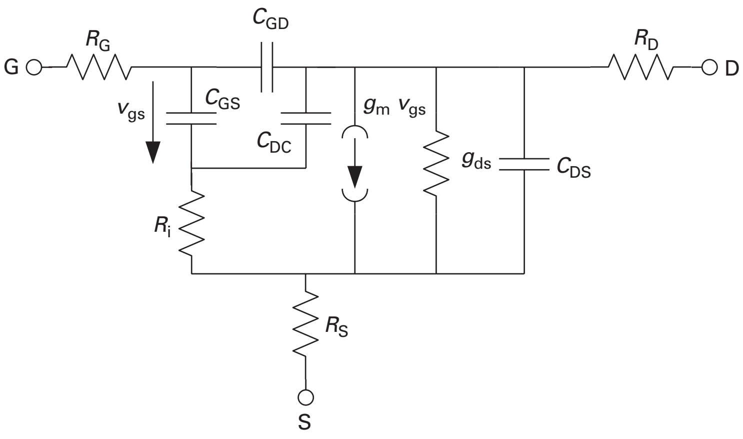


Fig. 2.10 Small-signal equivalent circuit of a MESFET including R_i and C_{DC}.

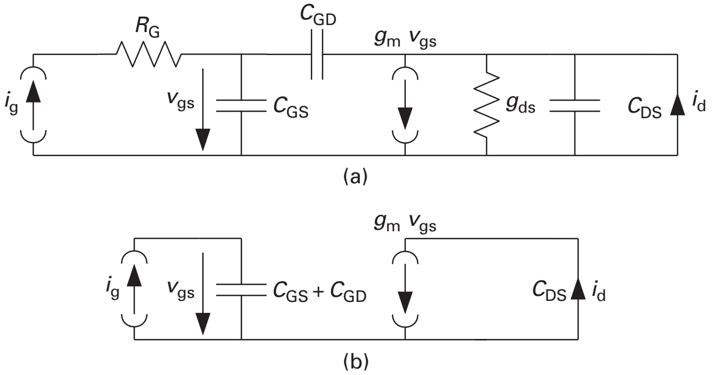


Fig. 2.11 (a) Simplified MESFET small-signal equivalent circuit connected for measuring f_T.
(b) Collapsed equivalent circuit due to current source at input and short circuit at output.

Transit frequency

The transit frequency of a two-port is defined as the frequency where the magnitude of the *short-circuit current gain* h_{21} becomes one:

$$|h_{21}(f = f_T)| = \frac{i_d}{i_g}\,|_{v_{ds}=0} = 1. \tag{2.24}$$

To calculate f_T from the small-signal parameters, we refer back to the simple MESFET equivalent circuit (Figure 2.9), and further simplify it by omitting the series resistances R_S and R_G, which in technical MESFETs are quite small.

Equation (2.24) can be interpreted as forcing a current i_g into the gate terminal, while measuring a short-circuit current i_d between the source and the drain terminals.

The appropriate connections are indicated in Figure 2.11(a). Because R_G is in series with an ideal current source, it has no effect here and can be omitted. Elements g_{ds} and C_{DS} are short-circuited and can be omitted also. C_{GD} is in parallel to C_{GS}.

Figure 2.11(b) shows the extremely simple equivalent circuit after taking these findings into consideration. The current transfer function is now:

$$i_{\mathrm{d}} = g_{\mathrm{m}} v_{\mathrm{gs}} = i_{\mathrm{g}} \frac{g_{\mathrm{m}}}{j\omega (C_{\mathrm{GS}} + C_{\mathrm{GD}})},$$

which means that

$$h_{21}(\omega) = \frac{g_{\mathrm{m}}}{j\omega (C_{\mathrm{GS}} + C_{\mathrm{GD}})}.$$

The magnitude of h_{21} becomes unity at

$$\omega_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{C_{\mathrm{GS}} + C_{\mathrm{GD}}},$$

or

$$f_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{2\pi (C_{\mathrm{GS}} + C_{\mathrm{GD}})}. \tag{2.25}$$

We will now relate the transit frequency to physical parameters. Let us go back to the simple velocity-saturated MESFET model (Section 2.2.3). In the simple model, we do not use the Meyer capacitance approach, but attribute the full gate-channel capacitance Equation (2.19) to the gate-source capacitance. Parasitic capacitances are neglected:

$$C_{\mathrm{GS}} = \frac{q N_{\mathrm{D}} W_{\mathrm{G}} L_{\mathrm{G}} a}{2\sqrt{(V_{\mathrm{bi}} - V_{\mathrm{P}})(V_{\mathrm{bi}} - V_{\mathrm{GS}})}}.$$

For the transconductance, we derive Equation (2.11) with respect to V_{GS} and find

$$g_{\mathrm{m}} = \frac{q N_{\mathrm{D}} v_{\mathrm{sat}} W_{\mathrm{G}} a}{2\sqrt{(V_{\mathrm{bi}} - V_{\mathrm{P}})(V_{\mathrm{bi}} - V_{\mathrm{GS}})}}.$$

Therefore,

$$f_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{2\pi C_{\mathrm{GS}}} = \frac{1}{2\pi} \frac{v_{\mathrm{sat}}}{L_{\mathrm{G}}}. \tag{2.26}$$

The transit frequency can be directly deduced from the carrier transit time through the channel.

Maximum frequency of oscillation

The maximum frequency of oscillation $f_{\max}$ is a measure of the power gain of a two-port (see Section 5.2.4). A common formulation [32] quoted in [36] for $f_{\max}$ from the small-signal parameters is

$$f_{\max} = \frac{f_{\mathrm{T}}}{2\sqrt{(R_{\mathrm{G}} + R_{\mathrm{i}} + R_{\mathrm{S}}) g_{\mathrm{ds}} + 2\pi f_{\mathrm{T}} R_{\mathrm{G}} C_{\mathrm{GD}}}}. \tag{2.27}$$

The expression refers to the equivalent circuit in Figure 2.10, but neglecting C_{DC}.

Note the importance of the series resistances, which did not factor into the calculation of f_{T} at all. $f_{\max}$ is much more useful to benchmark FETs for power amplification at microwave frequencies.

2.2.6 Noise performance

When discussing the noise performance of semiconductor devices, we have to distinguish between microwave noise, where the spectral power density of the contributing noise sources is frequency-independent (white noise), and low-frequency noise phenomena, where the spectral power density of the contributing noise sources increases with decreasing frequency.

Microwave noise

To assess the microwave noise performance of a FET, in principle three different noise sources need to be included, each due to the stochastic movement of charge carriers in different parts of the device. The simplified equivalent circuit in Figure 2.12 contains the MESFET's main noise sources:

(i) Areas where mobility is constant, i.e. the region behaves like an ohmic resistor, give rise to *thermal* or *Johnson* noise. In a realistic MESFET, we need to include Johnson noise for the gate resistance R_G and the source resistance R_S. The mean-squared value of a Johnson noise source can be expressed as: $\langle |e|^2 \rangle = 8kTR\Delta f$, where R is the resistance and Δf is the measurement bandwidth. Hence,

$$\left\langle |e_G|^2 \right\rangle = 8kTR_G\Delta f$$

$$\left\langle |e_S|^2 \right\rangle = 8kTR_S\Delta f.$$

(ii) Current flowing across an energy barrier gives rise to *shot noise*, which is proportional to the current. Here, a potential gate leakage current flows across the gate-channel Schottky diode, resulting in a shot noise component of

$$\left\langle |i_{glc}|^2 \right\rangle = 8qI_{GLC},$$

where I_{GLC} is the DC value of the gate leakage current.

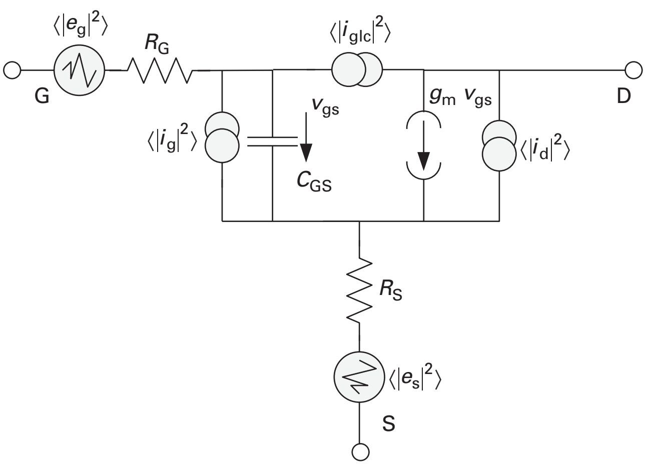


Fig. 2.12 Simplified noise equivalent circuit of a MESFET.

(iii) In the channel, the carrier velocity experiences fluctuations due to phonon and impurity scattering – this kind of noise is commonly called *channel noise*, first predicted by van der Ziel [62]. According to van der Ziel,

$$\left\langle |i_{\mathrm{d}}|^2 \right\rangle = 8kTg_{\mathrm{m}}P\Delta f,$$

where P is a fitting parameter equal to 1… 3.[2]

Note that van der Ziel did not yet include velocity saturation effects. In fact, analytic FET noise models are strictly valid only below the onset of saturation. However, deviating behaviour in the saturated region and in the presence of velocity saturation can be accommodated by a bias dependence in the parameter P [20].

(iv) Another effect must be taken into account. Due to the close proximity of the gate electrode to the channel, any charge fluctuation in the channel will lead to a phase fluctuation with the opposite sign on the gate electrode. This effect is the *induced gate noise* and was again pointed out by van der Ziel [63]:

$$\left\langle |i_{\mathrm{g}}|^2 \right\rangle = 8kT\Delta f(\omega C_{\mathrm{GS}})^2 R/g_{\mathrm{m}},$$

where R is a fitting parameter, which accommodates different geometries and bias points.

Because of their linked physical origin, $\langle |i_{\mathrm{d}}|^2 \rangle$ and $\langle |i_{\mathrm{g}}|^2 \rangle$ are not statistically independent, but show a strong correlation. The correlation coefficient is imaginary (due to the capacitive coupling) and strongly bias-dependent.

The gate leakage noise contribution $\langle |i_{\mathrm{glc}}|^2 \rangle$ is commonly neglected, because the gate diode is reverse biased. Using this assumption, Cappy [6] expressed the minimum noise figure as

$$F_{\min} = 1 + 2\sqrt{P + R - 2C\sqrt{PR}}\,\frac{f}{f_{\mathrm{T}}} \sqrt{g_{\mathrm{m}}(R_{\mathrm{S}} + R_{\mathrm{G}}) + \frac{PR(1 - C^2)}{R + P - 2C\sqrt{RP}}}, \tag{2.28}$$

where C is the magnitude of the correlation coefficient. For $C = 1$, Equation (2.28) is equivalent to the famous *Fukui equation* [21] for the minimum noise figure of FETs:

$$F_{\min} = 1 + k_{\mathrm{F}}\frac{f}{f_{\mathrm{T}}}\sqrt{g_{\mathrm{m}}(R_{\mathrm{G}} + R_{\mathrm{S}})}, \tag{2.29}$$

where k_{F} is a fitting factor.

Both Equations (2.28) and (2.29) calculate f_{T} using the approximation in Equation (2.25).

The noise equations contain an implicit bias dependence, which cannot be discussed in detail. Delagebeaudeuf *et al.* [12] showed for the bias dependence of parameter P,

$$P = \frac{I_{\mathrm{D}}}{\mathcal{E}_{\mathrm{crit}}L_{\mathrm{G}}g_{\mathrm{m}}}, \tag{2.30}$$

[2] van der Ziel discusses this in terms of the channel conductance g_{d0}, which is identical to the transconductance g_{m} at the very low V_{DS}.

which points towards a $\sqrt{I_D}$ dependence for $F_{\min}$, at least where $g_m \approx$ const. At very low I_D, however, g_m also decreases and $F_{\min}$ increases again. Optimum drain currents for low-noise operation are typically at 0.15–0.25I_{DSS}. In Equation (2.30), $\mathcal{E}_{\text{crit}}$ is the critical electric field for velocity saturation.

Low-frequency noise

Low-frequency noise is only discussed briefly here; however, it will be shown that it has significant impact on circuit performance, especially in oscillators.

While there are quantum mechanical reasons for low-frequency noise occurring in any conducting or semi-conducting material, practical devices exhibit low-frequency noise levels significantly above the quantum limit. This excess noise is due to interaction with impurities or dislocations which create energy levels inside of the forbidden gap of semiconductor materials. These traps may

- locally lead to enhanced scattering of charge carriers – *mobility fluctuation noise*; or
- modify the number of charge carriers through trapping and release, with a characteristic time constant τ – *number fluctuation noise*.

Even though it was derived at first only for mobility fluctuation noise in bulk semiconductors, the empirical Hooge equation [24] is often applied to low-frequency noise parameters. Applied to the drain current I_D, the Hooge relationship finds for the spectral power density of the drain current fluctuations:

$$S_{ID} = I_D^2 \frac{\alpha_H}{N \cdot f}, \tag{2.31}$$

where N is the number of carriers in a given volume and f is the frequency. Due to the observed frequency relationship, low-frequency noise is often coined *1/f noise*.

This ideal $1/f$ noise spectrum is frequently superimposed by generation-recombination spectra through a trap with distinct capture and re-emission time constant τ. Such traps lead to noise with a low-pass limited spectral noise power density:

$$S_N(f) \sim \frac{1}{1 + (2\pi f)^2 \tau^2}. \tag{2.32}$$

Figure 2.13 shows qualitatively a low-frequency noise spectrum of the drain current in the presence of a distinct trap with generation-recombination noise, a $1/f$ noise component and white noise at higher frequencies.

The absolute spectral density depends very strongly on the technology. A high spectral density at a given frequency is indicative for a high number of defects or deep impurities. So it is not surprising that low-frequency noise is much more pronounced for ion-implanted MESFETs (with significant radiation-induced defects) than for epitaxially grown structures.

2.2.7 MESFETs in the third millennium

Commercially, MESFETs were the transistors of choice for microwave circuits, including monolithic microwave ICs (MMICs), from the 1970s well into the 1990s. Initially,

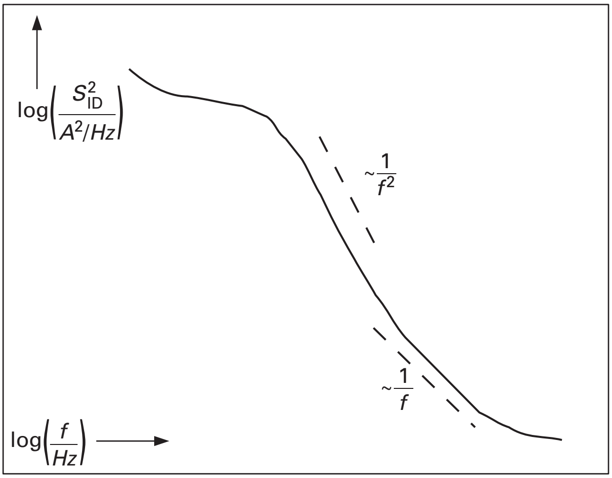

$\log\left(\frac{S_{ID}^2}{A^2/Hz}\right)$
$\sim\frac{1}{f^2}$
$\sim\frac{1}{f}$
$\log\left(\frac{f}{Hz}\right)$

Fig. 2.13 Qualitative low-frequency noise spectrum of the drain current in the presence of $1/f$ noise, generation-recombination noise and white noise.

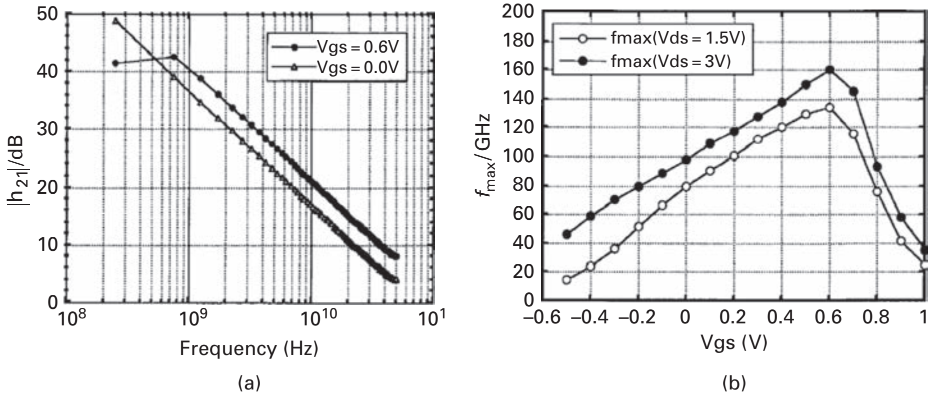


Fig. 2.14 (a) $|h_{21}|$ of a $2 \cdot 0.12\,\mu\text{m} \cdot 75\,\mu\text{m}$ ion-implanted MESFET ($V_{DS} = 1.2\,\text{V}$). (b) $f_{\max}$ of a $2 \cdot 0.12\,\mu\text{m} \cdot 25\,\mu\text{m}$ device in the same technology (H. Hsia, Z. Tang, D. Caruth, D. Becher and M. Feng, *IEEE Electron Device Letters*, Vol. EDL-20, No. 5, pp. 245–247, May 1999. ©1999 IEEE).

they only gained importance on GaAs substrates – MESFET structures on other materials, including Si, were but an academic curiosity. GaAs MESFETs have been almost exclusively replaced in contemporary designs – either by HBTs or HFETs. MOSFETs are also making inroads into the former realm of GaAs MESFETs.

It should be noted that MESFETs did reach cutoff frequencies in excess of 100 GHz, even for devices fabricated by ion implantation. Hsia and co-workers [25] reported a device with $L_G = 0.12\,\mu\text{m}$, which exhibited $f_T = 121\,\text{GHz}$ and $f_{\max} = 160\,\text{GHz}$, albeit not at the same bias point or the same device size. Figure 2.14 shows h_{21} versus frequency and $f_{\max}$ versus V_{GS} for this technology.

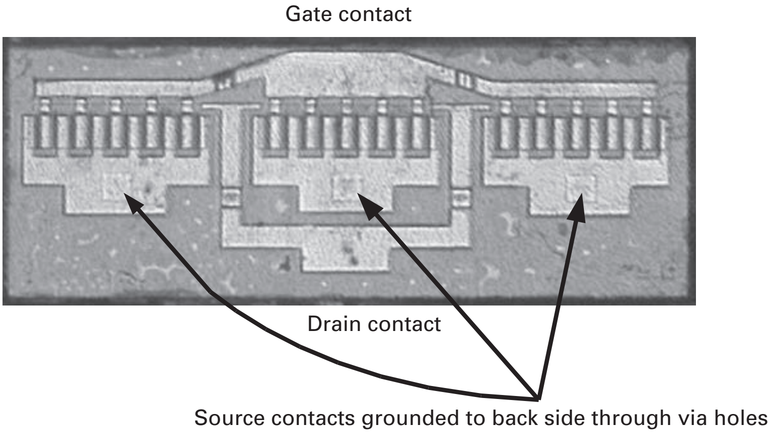


Fig. 2.15 Example of a SiC power MESFET (Chip photo adapted from M. Südow, K. Andersson, N. Billström, J. Gran, H. Hjelmgren, J. Nilsson, P.-A. Nilsson, J. Stahl, H. Zirath and N. Rorsman, *IEEE Transactions on Microwave Theory and Techniques*, Vol. MTT-54, No. 12, pp. 4072–4078, December 2006. ©2006 IEEE).

Note that the maximum f_T is measured at $V_{GS} = 0.6\,V$, i.e. the gate electrode starts to be forward-biased. This is also visible from the lower $|h_{21}|$ below 1 GHz. For a more practical $V_{GS} = 0\,V$, $f_T = 70\,GHz$. f_{max} peaks also at $V_{GS} = 0.6\,V$, but is improved by a larger V_{DS}, because the latter will further reduce C_{DG}.

The record f_T and f_{max} values are measured for different device geometries. This is a common trick – f_T is measured for a larger gate finger width (here, 75 μm), because R_G does not matter, and the wider finger leads to a better ratio of intrinsic and parasitic C_{GS}. For f_{max}, R_G does matter, and hence a smaller gate finger width is chosen (here 25 μm).

The MESFET structure makes a strong comeback on SiC, with important applications in power amplifiers, e.g. for mobile radio base stations in the lower GHz range. Figure 2.15 shows an example of such a structure [60].

Note the multi-finger layout which is very common in power FETs. Due to the limited current-carrying ability per unit width (in this case $350\,mA\,mm^{-1}$), the total device periphery needs to be extended. As the series resistance per unit length of the gate stripe is quite high in case of submicron gate length (here, $L_G = 0.4\,\mu m$), R_B can be kept small by choosing a short individual gate finger length and connecting transistor cells in parallel, in this example for a total gate width $W_G = 0.4\,mm$.

The major advantage of a semiconductor material with a large band gap is the very high electric field at breakdown. In this case, the gate-drain breakdown voltage is 180 V. The transistor shown produces a saturated output power of 3.1 W at 3 GHz, or 7.8 W/mm gate width, when biased at a drain-source voltage $V_{DS} = 65\,V$. The power added efficiency in this mode of operation is 70%.

The device has a small-signal $f_T = 8\,GHz$ and a maximum frequency of oscillation $f_{max} = 20\,GHz$. Record transit frequencies were reported at 28 GHz, and record maximum frequencies of oscillation at 50 GHz.

2.3 High electron mobility transistor

While the MESFET is conceptually a very simple device, yielding sufficient performance well into the millimetre wave range, it does not unleash the full potential of group III–V semiconductor materials. The fact that free charge carriers and ionised dopants share the same space in MESFETs leads to a reduction of low-field mobility through electrostatic fields, a major effect which we will consider first.

2.3.1 The importance of Coulomb scattering

Figure 2.16 shows the electron mobility for nominally undoped GaAs as a function of the absolute temperature, along with the two dominant scattering mechanisms. Other scattering mechanisms have been omitted for clarity.

We notice that at room temperature (300 K) the scattering of electrons is mostly due to lattice vibrations – longitudinal optical phonons. As we lower the temperature and lattice vibrations are increasingly suppressed, another mechanism becomes dominant – *Coulomb* scattering. Coulomb scattering is due to the electrostatic force between the mobile charge carriers and the fixed ionised atoms. In doped semiconductors, the main source of fixed charge are the ionised doping atoms. Therefore, the main electrostatic effect we need to consider in an n-channel MESFET is between the negatively charged electrons and the positively charged ionised donors. This is clearly shown in Figure 2.16 through the strong doping dependence of the mobility.

From electrostatic theory we know that the force created between two objects with a charge of magnitude q – the elementary charge – and the opposite sign of charge is

$$F = \frac{q^2}{4\pi \epsilon_s d^2} \propto \frac{1}{d^2}. \qquad (2.33)$$

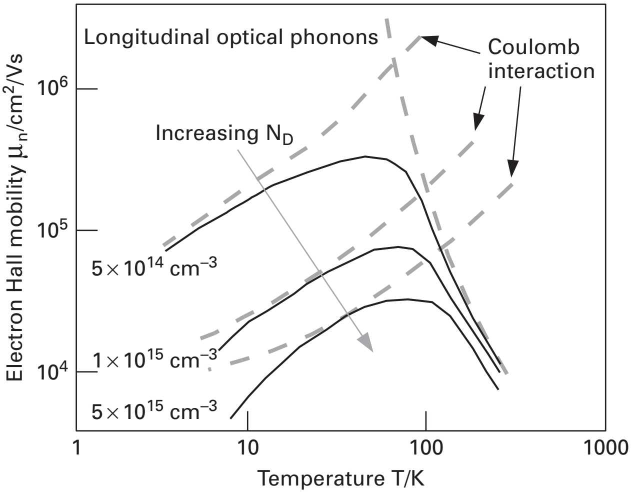


Fig. 2.16 Electron mobility versus absolute temperature for nominally undoped GaAs, and the underlying dominant scattering mechanisms. Data adapted from [59].

With increasing doping concentration, the mobility limiting effect of Coulomb scattering will become more pronounced. This is also shown in Figure 2.16.

Coulomb scattering becomes an increasing problem as we reduce the gate length in MESFETs:

- As we reduce the gate length L_G, we also have to reduce the channel thickness a to keep the *aspect ratio* L_G/a constant.[3]
- To compensate for the reduction in a, we need to increase the *channel doping concentration* N_D.
- Then, however, the significance of Coulomb scattering will increase and reduce the mobility!

If the physical co-location of free and fixed charge is the reason for the increased dominance of Coulomb scattering, then the following idea is immediately apparent: why not physically separate free and fixed charge, i.e. the electrons and the ionised donors in an n-channel device?

To find out how this may be done, let us investigate a *heterojunction* in the n^+ AlGaAs/p^- GaAs material system. The AlGaAs/GaAs material system has the advantage that the lattice constant is almost independent of the material composition.

The band gap in an $Al_xGa_{1-x}As$/GaAs heterojunction adjusts as follows:

E_g(GaAs)	1.42 eV	
ΔE_C(AlGaAs – GaAs)	$0.62\,\Delta E_g$	for $x_{Al} < 0.37$
ΔE_g(AlGaAs – GaAs)	$1.255\,\text{eV}\ x_{Al}$	as above

In this example, the Al concentration, doping types and concentrations are:

$Al_{0.25}Ga_{0.75}As$	n-type	$N_D = 10^{18}\,\text{cm}^{-3}$
$GaAs$	p-type	$N_A = 10^{15}\,\text{cm}^{-3}$

Further, a thin (~5–10 nm) layer of undoped AlGaAs is inserted at the heterojunction – this is the *spacer layer*. Figure 2.17 shows this material combination schematically. When discussing heterostructures, the doping type of large-gap materials will be denoted with capital letters, while the doping type of narrow-gap materials is shown in lower case letters.

The conduction band diagram of this heterostructure is shown in Figure 2.18. The discontinuity at the AlGaAs/GaAs interface, $\Delta E_C = 0.2\,\text{eV}$, and the potential barrier towards the p^--GaAs form a triangular quantum well structure, which is the most important feature – note how it dips below the Fermi level. Close to the hetero-interface, the potential in the GaAs layer can be approximated by a linear function.

[3] Otherwise, the assumption that the electric field is directed predominantly in parallel to the surface will break down. Among other things, this would significantly increase the output conductance in the saturated regime.

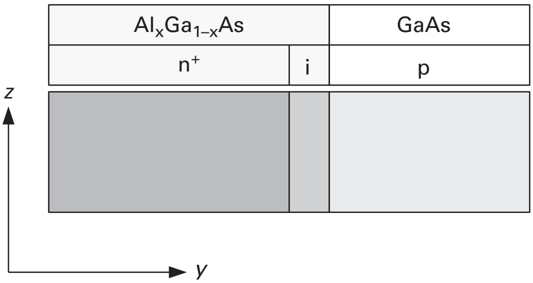
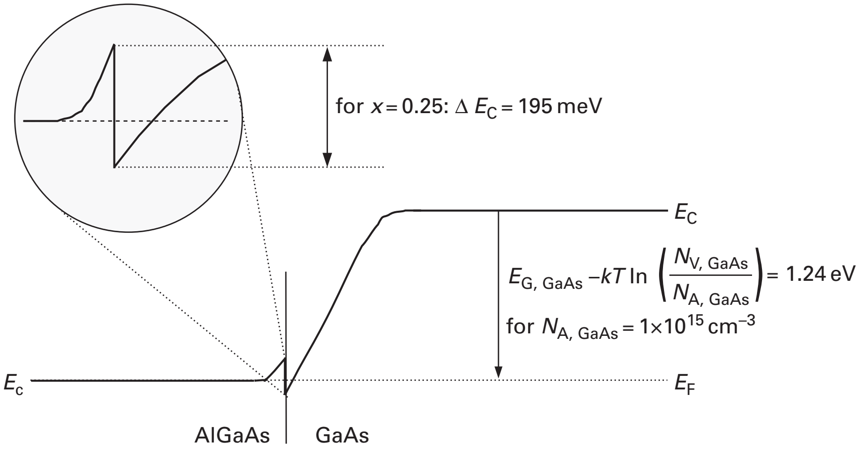


Fig. 2.17 AlGaAs/GaAs n^+–i–p heterostructure.



for $x = 0.25$: $\Delta E_C = 195$ meV
E_C
$E_{G,\,GaAs} - kT \ln\left(\frac{N_{V,\,GaAs}}{N_{A,\,GaAs}}\right) = 1.24$ eV
for $N_{A,\,GaAs} = 1\times10^{15}$ cm^{-3}
E_c
E_F
AlGaAs
GaAs

Fig. 2.18 Conduction band diagram of the AlGaAs/GaAs heterostructure.

2.3.2 Charge control

If we force the free electrons out of the n-AlGaAs layer, they will congregate in the potential well where they are separated from the ionised donor atoms by the undoped AlGaAs spacer layer – the sought-after *reduction in Coulomb scattering* can be achieved this way.

We can ‘force’ the free electrons to leave the AlGaAs when we *deplete* the doped AlGaAs layer (the *supply layer*) by means of a Schottky contact. The electrons can then either tunnel through the spacer layer or overcome the conduction band spike at the heterostructure by thermionic emission.

Note that from now on, it will suffice to draw just the conduction band diagram, because we consider electrons only.

Figure 2.19 represents the band diagram, without applied external voltage, of a high electron mobility transistor, or HEMT. By applying a positive gate voltage, the density of free electrons in the potential well increases; a negative gate voltage will decrease it.

An important difference between the MESFET and the HEMT is the current control mechanism: in the MESFET, we controlled the thickness of the channel, while the

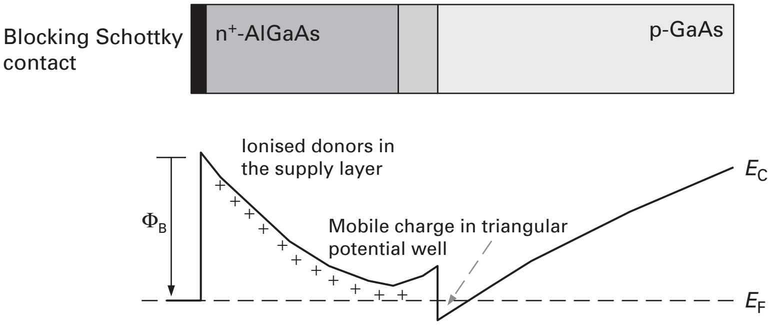

Blocking Schottky contact
n⁺-AlGaAs
p-GaAs
Ionised donors in the supply layer
Φ_B
Mobile charge in triangular potential well
E_C
E_F

Fig. 2.19 Conduction band diagram of HEMT in thermodynamic equilibrium.

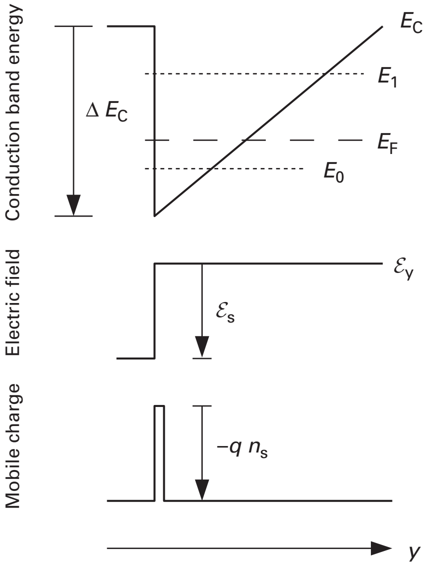


Fig. 2.20 Approximation of the HEMT channel region as a triangular quantum well.

density of charge carriers in the channel remained constant. Here we change the density of carriers in the channel, while the thickness of the channel, given by the triangular well, remains approximately constant.

The free carrier ensemble in the channel is called a *two-dimensional electron gas* (2DEG).

Let us investigate the triangular potential well in more detail. First, we have to be aware that the triangular potential well is narrow enough to introduce *quantisation of energy levels*. Consider Figure 2.20. Note that for convenience, $y = 0$ at the hetero-interface.

We initially assume that the potential walls are infinitely high. In the potential well, electrons may only occupy the discrete energy levels E_1, with $l \in [0, 1, 2, \ldots]$.

Solution of Schrödinger's equation yields these energies:

$$E_l = \left(\frac{h^2}{8\pi^2 m_n^*}\right)^{\frac{1}{3}} \left[\frac{3}{2} q\, \mathcal{E}_y\, \pi \left(l + \frac{3}{4}\right)\right]^{\frac{2}{3}}, \tag{2.34}$$

where $\mathcal{E}_y$ is the y component of the electric field in the well.

The potential increases linearly beyond the heterostructure. The electric field as the gradient of the potential is therefore constant:

$$\mathcal{E}_y = \mathcal{E}_S = -dV(y)/dy.$$

The discontinuity of the electric field at $y = 0$ necessitates a sheet charge in this plane, whose charge density is

$$q\, n_S = \epsilon_1 \mathcal{E}_S = -\epsilon_1 \frac{dV(y)}{dy}, \tag{2.35}$$

where ϵ_1 is the dielectric constant of the semiconductor in the channel region – here GaAs. This sheet charge is the 2DEG.

n_S is the sum over the sheet charge densities in the discrete energy levels: $n_S = \sum_{l=0}^{\infty} n_l$, where only the first two ($l = 0, 1$) typically need to be evaluated, because in practice the walls of the quantum well have a finite height, set by the conduction band discontinuity ΔE_C.

Using $\mathcal{E}_S = q\, n_S/\epsilon_1$ we find

$$E_l = \left(\frac{h^2}{8\pi^2 m_n^*}\right)^{\frac{1}{3}} \left[\frac{3}{2} \frac{q^2}{\epsilon_1} \pi \left(l + \frac{3}{4}\right)\right]^{\frac{2}{3}} n_S^{\frac{2}{3}}. \tag{2.36}$$

The first two terms are material-dependent and shall be combined into a constant γ_l:

$$\left(\frac{h^2}{8\pi^2 m_n^*}\right)^{\frac{1}{3}} \left[\frac{3}{2} \frac{q^2}{\epsilon_1} \pi \left(l + \frac{3}{4}\right)\right]^{\frac{2}{3}} \equiv \gamma_l,$$

and therefore,

$$E_l = \gamma_l\, n_S^{\frac{2}{3}}. \tag{2.37}$$

For GaAs,

$$\gamma_0 = 2.5 \times 10^{-12}\, e\mathrm{V\,m}^{\frac{4}{3}}$$
$$\gamma_1 = 3.2 \times 10^{-12}\, e\mathrm{V\,m}^{\frac{4}{3}}.$$

Next, we need to calculate the sheet charge density of the 2DEG as a function of the Fermi energy (note that the Fermi energy is referenced to the conduction band minimum here).

Consider the density of states for the two-dimensional electron gas in Figure 2.21. The constant $D = q\, m_n^*/(2\pi^2 h^2)$ is $D = 3.24 \times 10^{17}\ \mathrm{m^{-2}\,V^{-1}}$ for GaAs.

The density of the *occupied* states can be calculated from

$$n_S = \textit{density of states} \cdot \textit{occupation probability.}$$

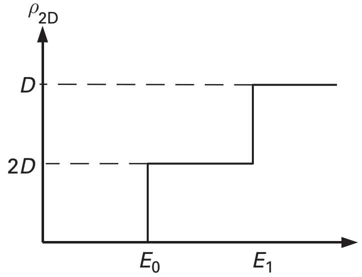


Fig. 2.21 Density of states in the triangular quantum well.

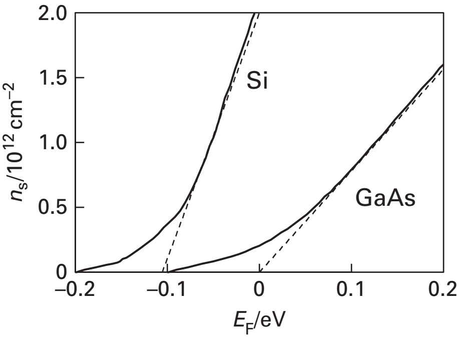


Fig. 2.22 Density of the 2DEG in a HEMT structure as a function of the Fermi energy.

The probability that an allowed state is occupied must be calculated using *Fermi–Dirac statistics* here, because the Fermi energy is inside the conduction band. Only two discrete energy levels are considered:

$$n_{\mathrm{S}} = D \int_{E_0}^{E_1} \frac{dE}{1 + \exp\left(\frac{E - E_{\mathrm{F}}}{kT}\right)} + 2D \int_{E_1}^{\infty} \frac{dE}{1 + \exp\left(\frac{E - E_{\mathrm{F}}}{kT}\right)}. \tag{2.38}$$

For the integral, we find

$$\int \frac{dx}{1 + \exp(a\,x)} = -\frac{1}{a}\ln(1 + e^{-ax}),$$

and therefore, the sheet charge density is

$$n_{\mathrm{S}} = D\,kT \sum_{l=0}^{1} \left[(l + 1) \cdot \ln\left(1 + e^{\frac{E_{\mathrm{F}} - E_{\mathrm{l}}}{kT}}\right)\right]. \tag{2.39}$$

Because on the other hand $E_{\mathrm{l}} = \gamma_l\, n_{\mathrm{S}}^{2/3}$, this transcendental equation has to be solved iteratively.

Figure 2.22 [64] shows its solution for the case of Si and for the case of GaAs. For larger charge carrier densities, $n_{\mathrm{S}}(E_{\mathrm{F}})$ can be approximated by a linear relationship:

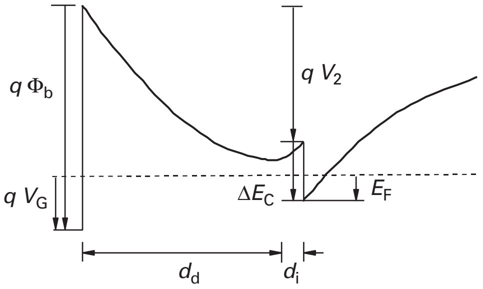


Fig. 2.23 Conduction band diagram of a HEMT structure under gate bias control ($V_G \neq 0$).

$$n_S \approx \frac{E_F - \Delta E_{F0}}{q\,a}. \tag{2.40}$$

For GaAs,

$$\begin{aligned} \Delta E_{F0}(300\,\mathrm{K}) &= 0\,\mathrm{eV} \\ \Delta E_{F0}(77\,\mathrm{K}) &= 25\,\mathrm{meV} \\ a &= 0.125 \times 10^{-12}\,\mathrm{V\,cm^2}. \end{aligned}$$

It is this linear relationship which we will use in our future calculations.

Finally, we need the relationship between n_S and the gate-channel voltage V_G. This means the potential across supply layer and spacer has to be included in the calculation.

We consider a structure where the supply layer is homogeneously doped and the spacer undoped. Integrating Poisson's equation twice, we find a parabolic potential in the supply layer, and a linear potential in the spacer (see Figure 2.23).

The built-in voltage drop over the AlGaAs layer V_2 is

$$V_2 = \frac{q\,N_D}{2\epsilon_2} d_d^2 - \mathcal{E}_S(d_d + d_i),$$

with $\mathcal{E}_S = q\,n_S/\epsilon_1$. ϵ_1 is the dielectric constant in the small-bandgap material (here, GaAs) and ϵ_2 is the corresponding value in the large-bandgap region (here, AlGaAs).

For the relationship between E_F and n_S, we use the linear approximation, Equation (2.40). Solving for n_S,

$$n_S = \frac{\epsilon_1}{q\left(d_d + d_i + \frac{\epsilon_2\,a}{q}\right)} \left(\frac{q\,N_D}{2\,\epsilon_2} d_d^2 + V_G - \Phi_b - \frac{\Delta E_{F0} - \Delta E_C}{q} \right). \tag{2.41}$$

We introduce a *threshold voltage* V_{off} as the gate-channel voltage where the interface carrier density disappears:

$$V_{\mathrm{off}} = \Phi_b + \frac{\Delta E_{F0} - \Delta E_C}{q} - \frac{q\,N_D}{2\,\epsilon_2} d_d^2. \tag{2.42}$$

For simplification, we define a *virtual increase of the supply layer thickness*:

$$\Delta d = \frac{\epsilon_2\,a}{q}. \tag{2.43}$$

We can now write Equation (2.41) in a more compact form:

$$n_S = \frac{\epsilon_1}{q} \frac{V_G - V_{off}}{d_d + d_i + \Delta d}. \tag{2.44}$$

The threshold voltage can be controlled via the thickness of the doped layer. Calculate the supply layer thickness where $V_{off} = 0$:

$$d_{d0} = d_d(V_{off} = 0) = \sqrt{\frac{2\,\epsilon_2}{N_D\, q}\left(\Phi_b + \frac{\Delta E_{F0} - \Delta E_C}{q}\right)}. \tag{2.45}$$

If now:

$d_d > d_{d0}$: The HEMT is normally on or operating in 'depletion-mode' – it will pass drain current for $V_{GS} = 0$.

$d_d < d_{d0}$: The HEMT is normally off or operating in 'enhancement-mode' – it will not pass drain current for $V_{GS} = 0$.

The threshold voltage in practical HEMTs is often tailored for the maximum transconductance to occur for $V_{GS} = 0$, which implies $V_{off} < 0$, as we will see further down.

Gate-channel capacitance. The gate-channel capacitance can be easily calculated by differentiating the charge in the 2DEG with respect to V_G:

$$C_0 = q\, W_G\, L_G \frac{dn_S}{dV_G} = \epsilon_2 \frac{W_G L_G}{(d_d + d_i + \Delta d)}, \tag{2.46}$$

for $V_G > V_{off}$.

For $V_G \leq V_{off}$, the 2DEG will be depleted, and in first-order approximation, the gate-channel capacitance disappears.

A practical HEMT example

Before we continue to consider the channel current as a function of gate-source and drain-source voltages, let us briefly look at a practical transistor structure.

The structure shown in Figure 2.24 is the classic cross-section of a HEMT. The source and drain contacts are non-rectifying ('Ohmic') contacts. To facilitate a low contact resistance, they sit on highly n-doped GaAs. AlGaAs habitually forms aluminium

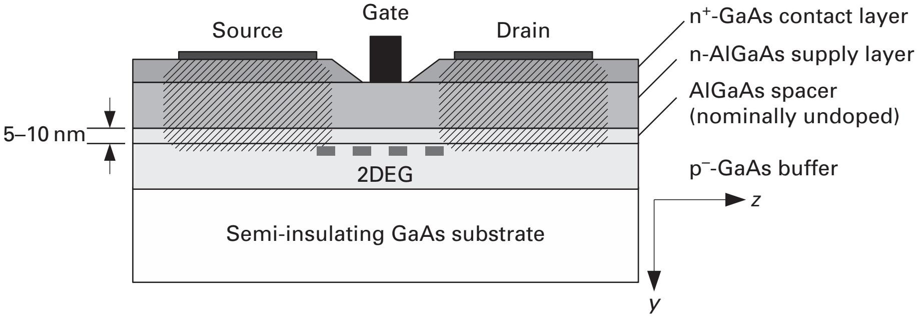


Fig. 2.24 Classic AlGaAs/GaAs HEMT structure.

oxide at the surface which would seriously increase the contact resistance. The alloying process will drive the contacts down through the heterostructure to make contact with the 2DEG. This is indicated by the shaded regions in Figure 2.24.

The gate contact must be a Schottky contact as shown. It sits in a recess through the GaAs contact layer. The recess depth controls the gate-channel separation ($d_\mathrm{d} + d_\mathrm{i}$) and is an important technological parameter.

The n-doped AlGaAs layer is called *supply layer* because its doping atoms supply the free carriers in the channel.

The thickness of the spacer layer (typically 5–10 nm) controls not only the reduction of Coulomb scattering, but also the transfer of electrons from the supply layer into the channel. It must be carefully optimised. The AlGaAs in the spacer is not actually *intrinsic* – it is just not intentionally doped.

The GaAs layer should be low doped, but it must be p-type. The free carriers in the channel must come from the supply layers and not from the GaAs buffer, otherwise the HEMT cannot be shut off under gate control – it exhibits *parallel conduction*. According to the mass action law,

$$n_\mathrm{p} = \frac{n_\mathrm{i}^2}{N_\mathrm{A}}.$$

As in GaAs the intrinsic carrier concentration is[4] $n_\mathrm{i} = 2.1 \cdot 10^6\,\mathrm{cm}^{-3}$, even a very low acceptor doping concentration, e.g. $N_\mathrm{A} = 10^{15}\,\mathrm{cm}^{-3}$, will virtually eliminate free electrons in the p-buffer.

The AlGaAs/GaAs heterostructure was first grown and analysed by R. Dingle at Bell Laboratories in 1974. For a review of early work on AlGaAs/GaAs heterostructures, refer to [13]. Mimura [39] was the first to practically realise a HEMT.

Channel current – constant mobility

So far, we only considered the case of $V_\mathrm{DS} = 0$. Now, we will allow $V_\mathrm{DS} > 0$, i.e. a current will flow between source and drain. This current is

$$I_\mathrm{D} = q\, n_\mathrm{S}(z)\, v_\mathrm{n}(\mathcal{E}_\mathrm{z})\, W_\mathrm{G} = \mathrm{const},$$

due to current continuity in the channel. Because of the voltage drop along the channel between a point z and a source $V(z)$, the density of the 2DEG now becomes z-dependent:

$$n_\mathrm{S}(z) = \frac{\epsilon_1}{(d_\mathrm{d} + d_\mathrm{i} + \Delta d)\, q}[V_\mathrm{GS} - V_\mathrm{off} - V(z)].$$

Figure 2.25 shows the immediate channel region and the appropriate voltages affecting the channel.

As in the MESFET, we assume that the channel is

- *one-dimensional*, i.e. the electric field $\mathcal{E}$ has only a component in z direction ($\mathcal{E}_\mathrm{z}$);
- *gradual*, i.e. the carrier densities change so slowly that diffusion currents can be neglected.

[4] www.ioffe.rssi.ru/SVA/NSM/Semicond/GaAs/bandstr.html

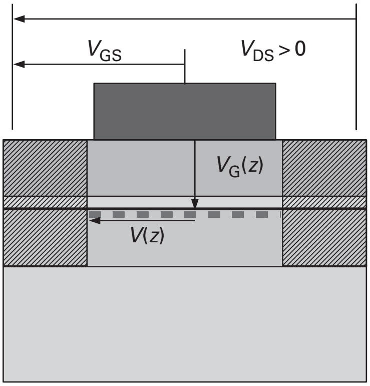


Fig. 2.25 HEMT channel region.

As in the MESFET, we will first consider the low-field case where $\mu_n = \text{const}$.

$$\begin{aligned} I_D &= q\, n_S(z)\, W_G\, \mu_n\, \mathcal{E}_z(z) \\ &= \frac{\epsilon_1\, W_G}{d_d + d_i + \Delta d}\, [V_{GS} - V_{off} - V(z)]\, \mu_n\, \frac{dV(z)}{dz}. \end{aligned} \tag{2.47}$$

Obeying current continuity, we find

$$I_D = \frac{\epsilon_1\, \mu_n\, W_G}{(d_d + d_i + \Delta d)\, L_G} \int_0^{L_G} [V_{GS} - V_{off} - V(z)]\, \frac{dV(z)}{dz} dz.$$

Let β be the transconductance parameter:

$$\beta = \frac{\epsilon_1\, \mu_n\, W_G}{(d_d + d_i + \Delta d)\, L_G}. \tag{2.48}$$

Then, using parameter substitution to integrate over V instead of z:

$$I_D = \beta \int_{V(0)}^{V(L_G)} [V_{GS} - V_{off} - V(z)]\, dV.$$

Note that $V(L_G) = V_{DS},\ V(0) = 0$.

Hence, we obtain the current–voltage characteristics in the linear regime (small V_{DS}):

$$I_D = \beta \left[(V_{GS} - V_{off})\, V_{DS} - \frac{V_{DS}^2}{2} \right]. \tag{2.49}$$

For very small $V_{DS} \ll 2(V_{GS} - V_{off})$, we note $I_D \approx \beta\, V_{DS}\, (V_{GS} - V_{off})$. In this regime, the HEMT acts as a 'voltage-controlled resistor'. The parameters β and V_{off} can be easily extracted if $V_{DS} = \text{const}$. This is shown in Figure 2.26. The drain current is measured for two V_{GS} while keeping $V_{DS} = \text{const} \ll 2(V_{GS} - V_{off})$. Linear extrapolation towards small V_{GS} provides V_{off} at the intersection with the V_{GS} axis. Once V_{off} is known, the transconductance parameter can be calculated as

$$\beta = -\frac{I_D(V_{GS} = 0)}{V_{DS}\, V_{off}}.$$

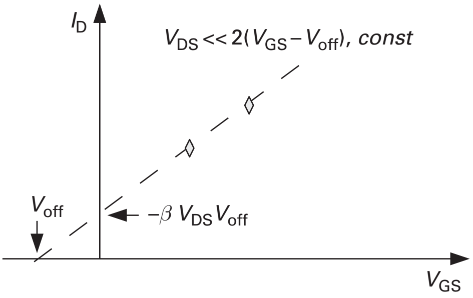


Fig. 2.26 Extraction of V_{off} and β at small V_{DS}.

Channel current – constant velocity

The *ansatz* in Equation (2.47) assumes that $n_{\text{S}}(z) > 0$ for all $0 < z < L_{\text{G}}$. If V_{DS} is sufficiently large, the gate-channel voltage may drop below V_{off} in the channel and the channel will become fully depleted. As $V(z) \leq V_{\text{DS}}$, this happens first at the drain end: $V_{\text{GS}} - V_{\text{DS}} = V_{\text{off}}$ or

$$V_{\text{DS}} \equiv V_{\text{k}} = V_{\text{GS}} - V_{\text{off}}. \tag{2.50}$$

This corresponds to what we saw in the MESFET, where the undepleted channel height disappeared at the drain end.

As in the MESFET, we can argue that velocity saturation prevents channel closure – $n_{\text{S}}(z) \rightarrow 0$ implies $\mathcal{E}_{\text{z}}(z) \rightarrow \infty$ due to the current continuity requirement, so that the constant-mobility assumption breaks down much earlier and $v_{\text{n}}(z) \rightarrow v_{\text{sat}}$.

Using a two-region model for the electron velocity, where mobility is constant for $|\mathcal{E}_{\text{z}}| < \mathcal{E}_{\text{crit}}$ and velocity is constant for $|\mathcal{E}_{\text{z}}| > \mathcal{E}_{\text{crit}}$, we can calculate the drain-source voltage V_{DSS} for which velocity saturation happens at the drain end of the channel ($z = L_{\text{G}}$). At this point the local electric field is $\mathcal{E}_{\text{z}}(z = L_{\text{G}}) = \mathcal{E}_{\text{crit}}$. Using the constant-mobility current Equation (2.49), we find

$$I_{\text{D}} = \frac{\epsilon_1 \mu_{\text{n}} W_{\text{G}}}{L_{\text{G}}(d_{\text{d}} + d_{\text{i}} + \Delta d)} \left[(V_{\text{GS}} - V_{\text{off}}) V_{\text{DSS}} - \frac{V_{\text{DSS}}^2}{2} \right].$$

On the other hand, I_{D} can be calculated using a constant-velocity *ansatz*:

$$I_{\text{D}} = q\, n_{\text{S}}\, W_{\text{G}}\, v_{\text{sat}}.$$

In the two-region model, $\mu_{\text{n}} \mathcal{E}_{\text{crit}} \equiv v_{\text{sat}}$. Further, n_{S} can be calculated from Equation (2.44) using $V_{\text{G}} = V_{\text{GS}} - V_{\text{DSS}}$, so that

$$I_{\text{D}} = \frac{\epsilon_1\, W_{\text{G}}\, \mu_{\text{n}}}{(d_{\text{d}} + d_{\text{i}} + \Delta d)} (V_{\text{GS}} - V_{\text{off}} - V_{\text{DSS}})\, \mathcal{E}_{\text{crit}}\, L_{\text{G}}.$$

The drain current expressions for constant mobility and constant velocity must be equal for $V_{\text{DS}} = V_{\text{DSS}}$, because we are transitioning from the constant mobility to

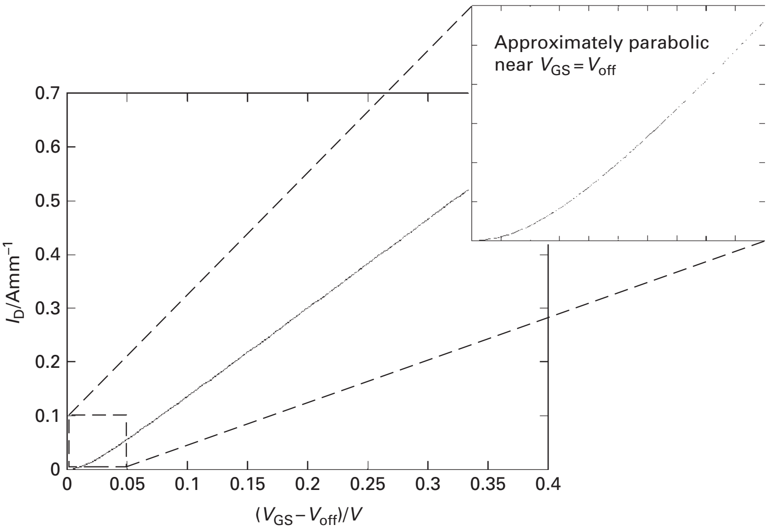


Fig. 2.27 Sample calculation of I_D following Equation (2.52). Parameters are $\beta = 83\,\mathrm{S\,(Vm)}^{-1}$, $\mathcal{E}_{crit} = 1\,kV\,\mathrm{cm}^{-1}$ and $L_G = 0.2\,\mu\mathrm{m}$.

the constant-velocity regime. This leads to a quadratic equation in V_{DSS}, which we can solve to find the necessary drain-source current for velocity saturation to set in:

$$V_{DSS} = V_0\left[1 + \frac{V_{GS}-V_{off}}{V_0} - \sqrt{1+\left(\frac{V_{GS}-V_{off}}{V_0}\right)^2}\right], \tag{2.51}$$

where $V_0 = \mathcal{E}_{crit}\,L_G$.

For the corresponding drain current in the velocity-saturated case, we finally find

$$I_{DSS} = \beta\,V_0^2\left[\sqrt{1+\left(\frac{V_{GS}-V_{off}}{V_0}\right)^2} - 1\right]. \tag{2.52}$$

In Figure 2.27, a sample calculation has been performed using Equation (2.52). Note that for the most part, I_D is a strictly linear function of V_{GS}. Very close to V_{off}, a more parabolic behaviour dominates.

2.3.3 Small-signal parameters

Transconductance in the saturated regime is calculated by differentiating Equation (2.52) with respect to V_{GS}:

$$g_m = \frac{d\,I_{DSS}}{d\,V_{GS}} = \frac{\beta\,(V_{GS}-V_{off})}{\sqrt{1+\left(\frac{V_{GS}-V_{off}}{V_0}\right)^2}}. \tag{2.53}$$

For large V_{GS}, transconductance is predicted to be constant, while it is approximately linearly dependent on V_{GS} for small V_{GS}. We will see later that this ideal behaviour is superseded by parasitic effects, however.

The gate-source capacitance C_{GS} can be found by differentiating the total channel charge Q_{T} with respect to V_{GS}:

$$C_{\mathrm{GS}} = \frac{dQ_{\mathrm{T}}}{dV_{\mathrm{GS}}} = \frac{d}{dV_{\mathrm{GS}}} W_{\mathrm{G}}\, q \int_0^{L_{\mathrm{G}}} n_{\mathrm{S}}(z) dz.$$

Particularly simple – and practically important – is the case of velocity saturation in the whole channel. As

$$I_{\mathrm{D}} = q\, n_{\mathrm{S}}(z)\, v_{\mathrm{sat}}\, W_{\mathrm{G}} = \mathrm{const} \Rightarrow n_{\mathrm{S}}(z) = \mathrm{const} = n_{\mathrm{SS}},$$

where

$$n_{\mathrm{SS}} = \frac{\beta\, V_0^2}{q\, v_{\mathrm{sat}}\, W_{\mathrm{G}}} \left[\sqrt{1 + \left(\frac{V_{\mathrm{GS}} - V_{\mathrm{off}}}{V_0} \right)^2} - 1 \right]$$

and therefore

$$Q_{\mathrm{T}} = q\, n_{\mathrm{SS}}\, W_{\mathrm{G}}\, L_{\mathrm{G}} = \frac{\beta\, V_0^2\, L_{\mathrm{G}}}{v_{\mathrm{sat}}} \left[\sqrt{1 + \left(\frac{V_{\mathrm{GS}} - V_{\mathrm{off}}}{V_0} \right)^2} - 1 \right].$$

The gate-source capacitance becomes in this case

$$\begin{aligned} C_{\mathrm{GS}} &= \frac{dQ_{\mathrm{T}}}{dV_{\mathrm{GS}}} \\ &= \frac{L_{\mathrm{G}}}{v_{\mathrm{sat}}} \frac{\beta\, V_0\, (V_{\mathrm{GS}} - V_{\mathrm{off}})}{\sqrt{(V_{\mathrm{GS}} - V_{\mathrm{off}})^2 + V_0^2}} \\ &= \frac{L_{\mathrm{G}}}{v_{\mathrm{sat}}} g_{\mathrm{m}}. \end{aligned} \tag{2.54}$$

As in case of the MESFET, we find for the transit time of carriers under the gate:

$$\tau_{\mathrm{T}} = \frac{L_{\mathrm{G}}}{v_{\mathrm{sat}}} = \frac{C_{\mathrm{GS}}}{g_{\mathrm{m}}}.$$

The gate-drain capacitance can be calculated similarly:

$$C_{\mathrm{GD}} = \frac{dQ_{\mathrm{T}}}{dV_{\mathrm{GD}}} = 0$$

in this simple model because $Q_{\mathrm{T}} \neq f(V_{\mathrm{GD}})$. In reality, C_{GD} is non-zero because of the geometric capacitance between the metal contacts and other parasitic effects. As in the MESFET, $C_{\mathrm{GD}} \ll C_{\mathrm{GS}}$ in saturation.

For the small-signal equivalent circuit, we can use the same topology as for the MESFET. Accordingly, the transit frequency can be approximated by

$$f_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{2\pi C_{\mathrm{GS}}}. \tag{2.55}$$

2.3.4 'High electron mobility'?

The common name 'high electron mobility transistor' deserves some critical reflection. Remember that while Coulomb scattering is the dominant mobility-limiting mechanism at cryogenic temperatures, phonon scattering is dominant at room temperature (Figure 2.16).

Realistic enhancement factors for the electron mobility in HEMTs are

- a factor of two at room temperature;
- up to a factor of 100 at cryogenic temperatures (e.g. 77 K – liquid nitrogen).

Furthermore, we found that in short-channel FETs velocity saturation dominates in the channel – hence the enhancement in mobility has two major advantages:

(i) reduction of series resistances, most importantly R_S;
(ii) lowering of the critical field for velocity saturation, i.e. saturated velocity will be reached sooner.

However, there are other advantages of the HEMT structure which are also significant:

- The carrier distribution is similar to that of a pulse-doped MESFET – we expect a constant transconductance g_m and therefore a high linearity. There are parasitic effects which prevent this from happening – more about this later.
- In active operation, the supply layer is fully depleted and hence the gate-source capacitance C_{GS} should be constant. Again, this is not quite true in reality (see p. 81).
- The potential barrier towards the substrate reduces carrier injection into the substrate and increases output resistance.

Note that, compared to a MESFET with an epitaxially grown channel, the HEMT structure is technologically not much more complicated.

As an aside, the HEMT has many other names and has jokingly been called a *multi-acronym device* (MAD). To name but a few:

HFET	heterostructure field effect transistor
MODFET	modulation-doped field effect transistor
TEGFET	two-dimensional electron gas FET
SDFET	selectively doped field effect transistor

2.3.5 Non-ideal behaviour

In the following pages, we will discuss how in practical HEMTs the experimentally observed behaviour deviates from the theory developed so far. The explanation of these non-ideal features will have important implications for the design of optimised HEMT devices.

Non-ideal HEMT behaviour for large V_{GS}

From simple HEMT theory as outlined above, we expect that for sufficiently large $V_{GS} - V_{off}$, the drain current increases linearly with V_{GS} and hence the transconductance is constant. Also, we would expect that the gate-source capacitance is constant in the same region.

Experimentally, however, transconductance and gate-source capacitance show the behaviour in Figure 2.28 [1]: after a sharp increase above the threshold voltage, the transconductance goes through a maximum, then decreases again for higher V_{GS}. The gate-source capacitance initially tracks the transconductance, as predicted by Equation (2.54) (save for a constant parasitic contribution), but then increases further for higher V_{GS}.

This compression of the transconductance is due to a 'parasitic MESFET' effect [33]. To understand its origin, please consider Figure 2.29.

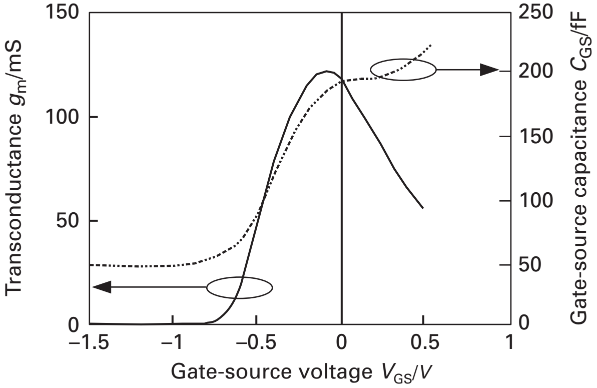


Fig. 2.28 Experimental transconductance and gate-source capacitance versus gate-source voltage.

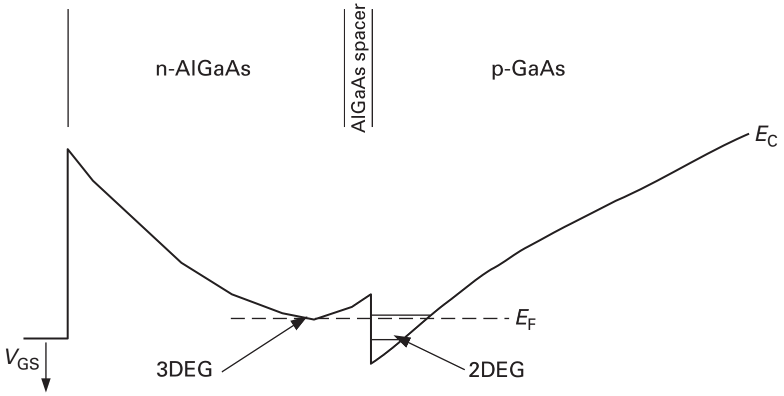


Fig. 2.29 Conduction band diagram of a HEMT under high V_{GS}. The arrows indicate the locations of the 2DEG and three-dimensional electron gas (3DEG).

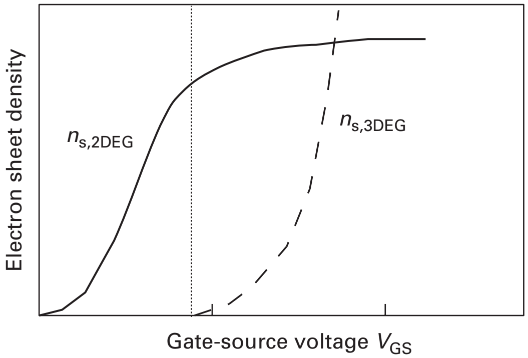


Fig. 2.30 Dependence of the electron sheet densities in the 2DEG and 3DEG as a function of the gate-source voltage.

If V_{GS} is sufficiently high, the conduction band minimum in the AlGaAs supply layer dips below the Fermi level. At this point, the free electron density in the supply layer will rapidly increase, the supply layer is no longer depleted. Because the free electron population in the AlGaAs conduction band minimum has very little confinement, it is referred to as the *three-dimensional electron gas* or *3DEG*. In its low confinement, this channel is very similar to a MESFET's, hence the term *parasitic MESFET*.

Once the 3DEG builds up, it electrostatically shields the 2DEG from the gate electrode – the 2DEG density $n_{S,2D}$ saturates; any further increase in charge density due to a further increase in V_{GS} will benefit only $n_{S,3D}$. This is schematically shown in Figure 2.30.

The rise of the 3DEG has two substantial effects:

- Because the mobility is much lower in the ternary AlGaAs supply layer than in the GaAs channel region, the resulting transconductance due to the 3DEG channel is lower, causing the overall transconductance to decrease.
- The additional charge under the gate leads to a strong increase in the gate-source capacitance.

Recall Equation (2.55) – the simultaneous decrease in transconductance and increase in gate-source capacitance will have a very negative impact on the transit frequency f_T. Using the data from Figure 2.28, this is exemplified in Figure 2.31.

The transit frequency, which in our simple theory was predicted to be independent of frequency, now shows a pronounced maximum, which occurs for gate-source voltages slightly lower than the transconductance maximum. For the design of high-speed circuits, this is an important observation.

The ungated FET structure can be used as a model for the situation at the onset of the parasitic MESFET effect, where $n_{S,3D} = N_D$. Figure 2.32 shows the conduction band diagram.

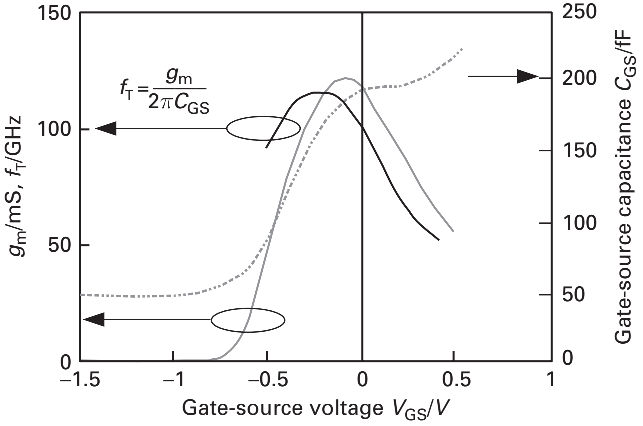


Fig. 2.31 Calculated transit frequency of the HEMT in Figure 2.28 versus gate-source voltage.

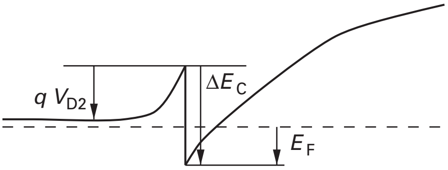


Fig. 2.32 Conduction band diagram of an ungated HEMT structure.

We find that

$$E_F = \Delta E_C - qV_{D2} - kT\,\ln\frac{N_{C,\text{AlGaAs}}}{N_D} \approx \Delta E_C - qV_{D2},$$

for large N_D. V_{D2} is the built-in potential in the large-band-gap part of the heterostructure. Using Equation (2.40),

$$n_{S,\max} \approx \frac{\Delta E_C - \Delta E_{F0} - qV_{D2}}{q\,a}.$$

To maximise $n_{S,\max}$, we must therefore choose a material combination with large ΔE_C. In a conventional HEMT structure, $n_{S,\max} \approx 1 \cdot 10^{12}\,\text{cm}^{-2}$.

Trapping effects

In an $\text{Al}_x\text{Ga}_{1-x}\text{As}$/GaAs HEMT, we may obtain a larger ΔE_C by increasing the Al mole fraction x. However, consider the following detrimental effects.

For $x_{\text{Al}} > 0.3$, the *effective energy depth of the donor level* increases – the number of free carriers provided by a given doping density N_D will decrease.

Earlier, for $x_{\text{Al}} > 0.25$, the *density of deep traps* (*DX centres*) will increase. These traps are energy states in the forbidden gap which can interact with the valence or conduction band. In this case, they are closer to the conduction band, at an energetic depth E_T – they are 'donor-like'. The 'X' denotes that their physical origin was long unknown. A trap will capture a free electron from the conduction band and eventually re-emit it.

The characteristic time constant for re-emission is strongly temperature-dependent:

$$\tau_{\mathrm{RE}}(T) = \tau_0 \exp\left(\frac{E_{\mathrm{T}} + E_{\mathrm{B}}}{kT}\right), \tag{2.56}$$

where E_{B} is an additional energy barrier for re-emission. In AlGaAs, $E_{\mathrm{T}} \approx 50\,\mathrm{meV}$ and $E_{\mathrm{B}} \approx 300\,\mathrm{meV}$.

When reducing the temperature, formerly free carriers will be ‘frozen’ and as a consequence, will no longer be available for the channel. This can be described by a shift in threshold voltage:

$$\Delta V_{\mathrm{off}} = -\frac{q}{2\epsilon} N_{\mathrm{DT,ion}}\, d_{\mathrm{d}}^2.$$

The density of ionised traps is, using Fermi–Dirac statistics,

$$N_{\mathrm{DT,ion}} = \frac{N_{\mathrm{DT}}}{1 + \exp\left(\frac{E_{\mathrm{F}} - E_{\mathrm{T}}}{kT}\right)}.$$

Note that the trap density N_{DT} has been experimentally observed to be proportional to the donor density N_{D}. It is now accepted that the donor atoms themselves introduce two different energy levels in the forbidden gap in AlGaAs: a shallow one associated with the Γ-minimum (the direct minimum) – this is the proper donor level – and a deep energy state associated with the L-minimum (an indirect minimum) – this is the DX centre [5].

The effect of DX centres in the supply layer made early HEMT structures very problematic in cryogenic operation.

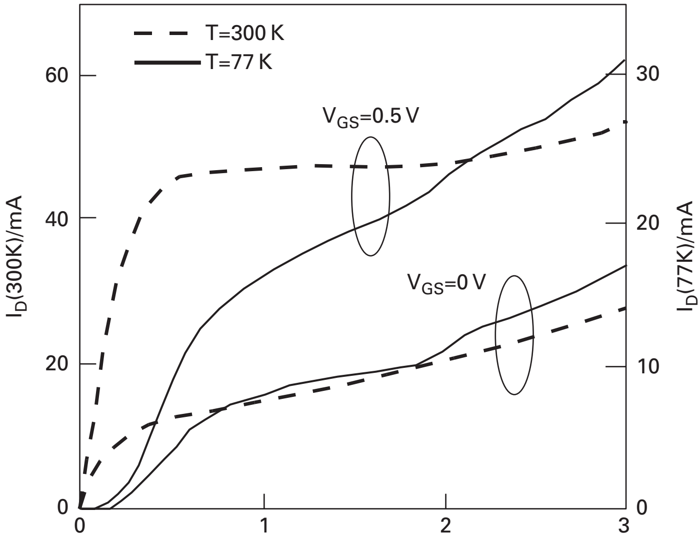


Fig. 2.33 Output I–V characteristics of a HEMT device with a low-temperature current collapse phenomenon (A. Belache, A. Vanoverschelde, G. Salmer and M. Wolny, *IEEE Transactions on Electron Devices*, Vol. ED-38, No.1, pp. 3–13, January 1991. ©1991 IEEE).

Figure 2.33 shows an example of a device showing such a current collapse phenomenon [3]. Apart from the change in output conductance in saturation and the various ‘kink’ effects, which shall not be discussed here, we note

- At low V_{DS}, the output conductance in the linear regime decreases considerably – this is due to an increase in the source resistance R_S. The decrease in R_S with decreasing T is unexpected, as the mobility itself will increase. The decrease in the free carrier concentration, however, dominates.
- In the saturated regime, $V_{DS} > 0.5\,\text{V}$, the transconductance also decreases significantly with decreasing temperature.

The occurrence of DX centres is closely linked to the use of AlGaAs as the barrier material – other supply layer materials such as GaInP do not show this effect and will correspondingly fare better in their low-temperature performance [7].

2.3.6 Structural HEMT variations

Increasingly, structural variations of the original HEMT concept are being used to circumvent the non-ideal behavioural effects explained above and to improve performance.

Pulse-doped HEMT structure

Due to the severeness of the DX centre limitation, a method to eliminate this limitation has the highest priority.

Because $N_{DT} \sim N_D$, the trap-induced threshold voltage shift is

$$\Delta V_{\text{off}} \sim N_D\, d_d^2.$$

On the other hand, the supply layer must be able to supply the necessary carrier density in the 2DEG:

$$N_D\, d_d > n_S.$$

If, therefore, we concentrate the doping in a narrow sheet – increase N_D and decrease d_d – the trap-induced threshold voltage shift can be drastically reduced.

This concept leads to the *delta-doped* (or pulse-doped) HEMT structure (see Figure 2.34).

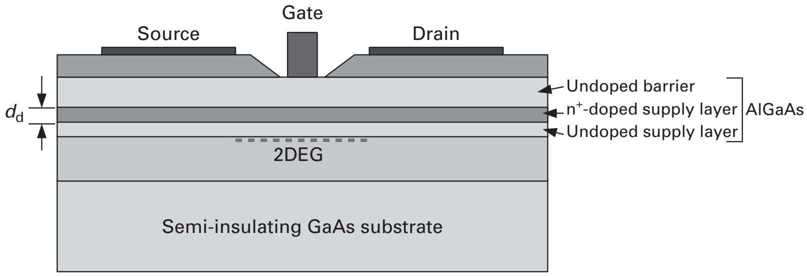


Fig. 2.34 Layer structure of a delta-doped HEMT.

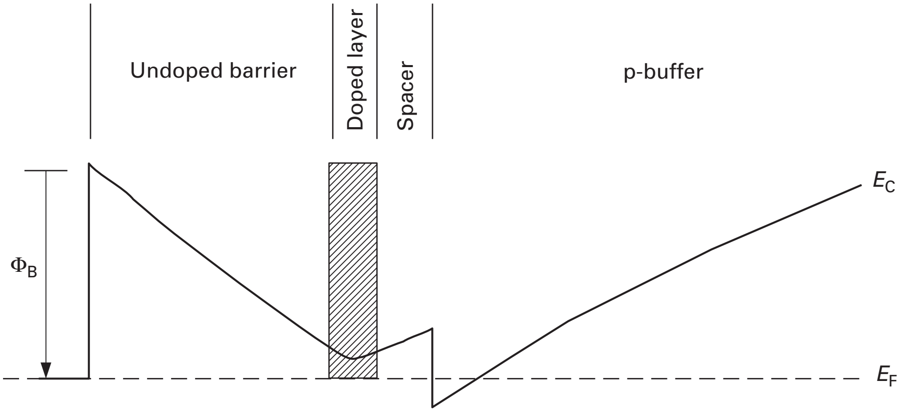


Fig. 2.35 Conduction band diagram of a pulse-doped HEMT structure.

The restriction of doping to only a narrow sheet of the wide-gap layer, of course, also modifies the band structure. The Poisson equation

$$\frac{d^2V}{dy^2} = -\frac{\rho}{\epsilon}$$

tells us that the potential V will have a linear y dependence if $\rho \simeq 0$. Where $\rho \neq 0 = \text{const}$, V will have a parabolic dependence on y. These principles are visible in the conduction band diagram of a pulse-doped HEMT structure (see Figure 2.35).

As an additional advantage, the pulse-doped HEMT can be expected to have lower gate leakage because of the lower doping of the region immediately under the gate.

Pseudomorphic HEMT structure

So far, the AlGaAs/GaAs HEMT structures considered were lattice-matched – a significant advantage of the (Al,Ga)As material system is that the lattice constant is almost independent of the Al content.

We will now deliberately leave the lattice match principle behind and allow for material combinations which are lattice-mismatched, but where the lattice difference is accommodated by elastic deformation of the crystal – *pseudomorphic* structures. This gives us greater flexibility in the choice of materials.

Let us replace the GaAs channel in a conventional HEMT with an InGaAs channel. This leads to a double-heterostructure because the GaAs buffer and substrate shall be maintained. In Figure 2.36, the conduction band diagram of an example structure combining the pseudomorphic channel layer with a pulse-doped barrier is shown.

Compared to the conventional HEMT, this structure has several advantages:

- The significantly higher conduction band discontinuity increases the maximum density of the 2DEG from about $1 \cdot 10^{12}\,\text{cm}^{-2}$ for the conventional HEMT to about $2 \cdot 10^{12}\,\text{cm}^{-2}$ for the pseudomorphic HEMT as shown.
- The low Al content in the supply layer reduces the density of DX centres.

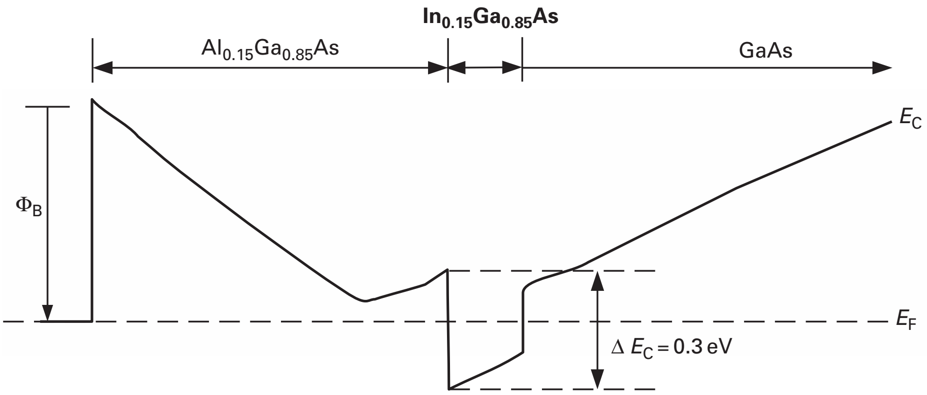


Fig. 2.36 Conduction band diagram of a pseudomorphic HEMT structure with a pulse-doped barrier.

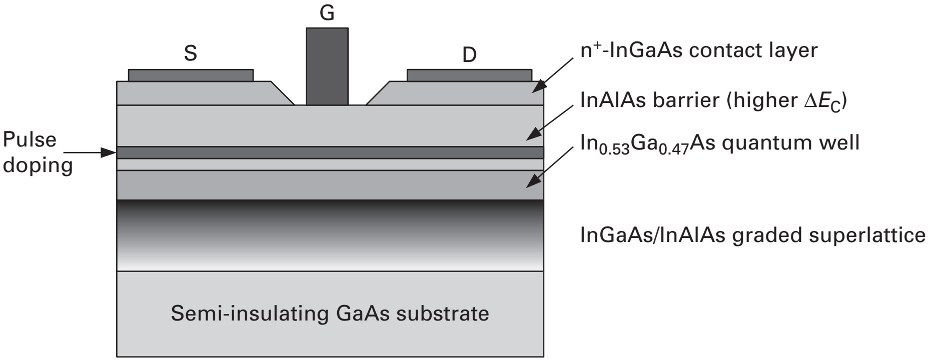

G
S
D
n$^+$-InGaAs contact layer
InAlAs barrier (higher ΔE_C)
Pulse doping
In$_{0.53}$Ga$_{0.47}$As quantum well
InGaAs/InAlAs graded superlattice
Semi-insulating GaAs substrate

Fig. 2.37 Layer structure of a metamorphic HEMT.

- The addition of In in the channel enhances the *low-field mobility* and, to a lesser extent, the peak velocity in the channel.
- The added heterostructure towards the GaAs buffer reduces injection of carriers into the buffer and substrate.

Higher ΔE_C are possible with higher In concentrations in the channel. However, note that with increasing In content, the InGaAs layer thickness must be reduced. In practical pseudomorphic HEMTs on GaAs substrates, $x_{\mathrm{In}} = 0.15\ldots 0.25$. This increases the maximum density of the 2DEG to $n_{\mathrm{S,max}} \simeq 2 \cdot 10^{12}\,\mathrm{cm}^{-2}$.

Metamorphic HEMT

Mobility in the channel would benefit from even higher In mole fractions, e.g. $x_{\mathrm{In}} = 0.53$, as in InGaAs lattice matched to InP. However, InP substrates are still considerably more expensive than GaAs wafers.

The metamorphic HEMT concept enables high In mole fractions in the channel layer on GaAs substrates, through a modification of the lattice constant in a *graded superlattice*. Such a layer structure is shown in Figure 2.37.

An InAlAs/InGaAs superlattice with varied thickness and composition is grown on top of the GaAs substrate such that the effective lattice constant (modified by composition and built-in mechanical strain) is adapted from that of GaAs to (in this case) the one of InGaAs with an In mole fraction of 0.53. The use of a low-temperature grown superlattice allows the change of the lattice constant while keeping the density of deformation-related crystal defects low.

As in all modern HEMTs, the barrier layer is assumed to be pulse-doped. Another modification is the use of InAlAs instead of AlGaAs as barrier material. InAlAs has a higher conduction band discontinuity towards InGaAs than AlGaAs for comparable Al mole fractions; furthermore, the InAlAs/InGaAs heterostructure stack is easier to grow. As contact layer material, InGaAs is used here because it has a much lower Schottky barrier height than GaAs.

2.3.7 CAD modelling of HEMTs

Due to the similarity of the HEMT to MESFETs and (as we will see in the next section) MOSFETs, CAD models of these two devices are often re-used to simulate HEMTs.

In the discussion of CAD modelling, we will go beyond the rather simple models provided for the MESFET and describe a high-accuracy semi-empirical approach. It is equally suitable for an enhanced precision model of the MESFET.

Static current equations

A HEMT-specific problem is the simulation of transconductance suppression at large V_{GS} (see p. 81). A suitable drain current expression which accommodates this (the discussion follows I. Kallfass [27]) is

$$I_{\mathrm{DS}}(V_{\mathrm{GS}}) = \beta\ (V_{\mathrm{GS}} - V_{\mathrm{off}})^{\lambda/(1+\xi\ (V_{\mathrm{GS}}-V_{\mathrm{off}}))} \tag{2.57}$$

in saturation – neglecting the V_{DS} dependence of I_{DS}.

The non-saturated region at small V_{DS} can be included using the *tanh* term already discussed in the context of the MESFET Curtice model:

$$I_{\mathrm{DS}} = \beta\ (V_{\mathrm{GS}} - V_{\mathrm{off}})^{\lambda/(1+\xi\ (V_{\mathrm{GS}}-V_{\mathrm{off}}))} \tanh\ (\alpha\ V_{\mathrm{DS}}). \tag{2.58}$$

In real devices, the drain-source voltage has a non-linear influence (so far neglected) on the current in saturation, e.g. through impact ionisation effects. In the non-saturated regime, on the other hand, the tanh $(\alpha\ V_{\mathrm{DS}})$ expression is not always sufficient, because the V_{GS} dependence is not adequately modelled. An effective voltage V_{eff} is introduced, replacing the simple $V_{\mathrm{GS}} - V_{\mathrm{off}}$ term in Equation (2.58):

$$\begin{aligned} I_{\mathrm{DS}} &= \beta\ V_{\mathrm{eff}}^{\frac{\lambda}{1+\mu V_{\mathrm{DS}}^2+\xi V_{\mathrm{eff}}}} \tanh\left[\alpha\ V_{\mathrm{DS}}\ (1+\zeta\, V_{\mathrm{eff}})\right] \\ V_{\mathrm{eff}} &= \frac{1}{2}\left(V_{\mathrm{GSt}} + \sqrt{V_{\mathrm{GSt}}^2 + \delta^2}\right) \\ V_{\mathrm{GSt}} &= V_{\mathrm{GS}} - (1+\beta_{\mathrm{r}}^2)\ V_{\mathrm{T0}} + \gamma\, V_{\mathrm{DS}}. \end{aligned} \tag{2.59}$$

This expression, introduced by Cojocaru and Brazil in 1997 [8], is called the *COBRA* current equation. Its advantage is that it is continuous in the entire bias plane, and also its derivatives are continuous, which is very important for simulations of the non-linear behaviour of circuits.

$\beta, \lambda, \mu, \xi, \alpha, \zeta, \delta, \gamma$ and V_{T0} are model parameters to be extracted by measurements. β_r is equal to β, but dimensionless. They affect I_{DS} as follows:

α, ζ affect the linear regime of the device – α is the main parameter modelling the V_{DS} dependence; ζ modifies the V_{GS}-dependent behaviour.
β is the main transconductance parameter.
ξ is the parameter which adjusts the transconductance compression.
γ introduces a V_{DS} dependence to the drain current in the saturated regime and is hence responsible for the output conductance.
μ equally introduces a V_{DS} dependence in the linear regime. It is used to model impact ionisation effects in the saturated regime.
λ adjusts the curvature of $I_{DS}(V_{GS})$ for small V_{DS} and close to threshold.
V_{T0} is the threshold voltage for small V_{DS}.

The drain current source $I_{DS} = f(V_{GS}, V_{DS})$ is embedded into an equivalent circuit to account for the series resistances and the non-linear gate-source and gate-drain contacts. This is shown in Figure 2.38. Note that the controlling voltages drop between the internal nodes! The diodes, D_{GS} and D_{GD}, are used to model the non-linear gate current. Breakdown behaviour can equally be included here:

$$\begin{aligned} I_{GS}(V_{GS}) = {} & I_{sgs}\left(\exp\frac{V_{GS}}{n_{id}\,V_T} - 1\right) \\ & + I_{bv}\exp\left(-\frac{V_{GS} - V_{bv}}{n_{bv}V_T}\right)\frac{V_{GS}}{V_{bv}} \end{aligned} \tag{2.60}$$

$$\begin{aligned} I_{GD}(V_{GD}) = {} & I_{sgd}\left(\exp\frac{V_{GD}}{n_{id}\,V_T} - 1\right) \\ & + I_{bv}\,\exp\left(-\frac{V_{GD} - V_{bv}}{n_{bv}V_T}\right)\frac{V_{GD}}{V_{bv}}, \end{aligned} \tag{2.61}$$

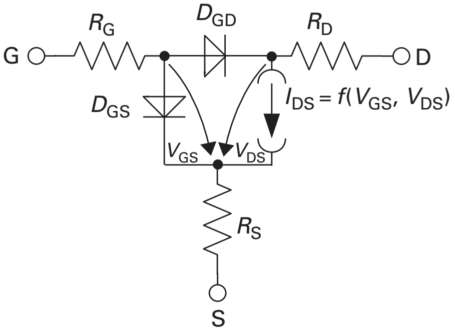


Fig. 2.38 Quasi-static equivalent circuit used in the COBRA model.

where I_{sgs} and I_{sgd} are saturation currents for the gate-source and gate-drain diodes, respectively, and n_{id} is the emission factor for these diodes. The second term in each equation models breakdown with an exponential diode term. V_{bv} is the breakdown voltage and I_{bv} and n_{bv} are used to model the current increase beyond breakdown. The very last product term simply makes sure that the breakdown current is zero, if either V_{GS} or V_{GD} are zero in Equations (2.60) or (2.61), respectively, but has no other major effect.

Non-linear capacitance equations

To properly model the non-linear behaviour in any FET, we need to account for several contributions:

- parasitic (non-bias-dependent) capacitance,
- the junction capacitance,
- the change in channel charge with varying voltage.

The first two are straightforward to model: the parasitic capacitance is C_{pgs} for the gate-source diode and C_{pgd} for the gate-drain diode. For the junction capacitance, the common form also implemented in SPICE is used:

$$C(V) = \frac{C_0}{\left(1 - \frac{V}{V_{\mathrm{bi}}}\right)^{\mathrm{m}}},$$

where C_0 is the capacitance without any external voltage, V_{bi} is the built-in voltage of the junction and m is an exponent.

Inclusion of the channel charge is much more complicated. For once, the channel charge depends on V_{GS} and V_{GD} simultaneously. Then, charge conservation needs to be satisfied. This means [28]

$$\frac{\delta C_{\mathrm{GS}}}{\delta V_{\mathrm{GS}}} = \frac{\delta^2 Q_G}{\delta V_{\mathrm{GS}} \delta V_{\mathrm{GD}}} = \frac{\delta^2 Q_G}{\delta V_{\mathrm{GD}} \delta V_{\mathrm{GS}}} = \frac{\delta C_{\mathrm{GD}}}{\delta V_{\mathrm{GS}}}. \tag{2.62}$$

Any empirical expressions for $C_{\mathrm{GS}}(V_{\mathrm{GS}}, V_{\mathrm{GD}})$ or $C_{\mathrm{GD}}(V_{\mathrm{GS}}, V_{\mathrm{GD}})$ must fulfil Equation (2.62).

Figure 2.39 shows gate-source and gate-drain capacitances experimentally determined from S-parameter measurements, as a function of V_{GS}, for V_{DS} values in the linear and the saturated regime of FET operation. Note the rather strong variation near pinch-off, and generally in the linear regime.

In the following empirical equations [29], the tanh(x) function is again exploited, similar to the Curtice models.

$$\begin{aligned} C_{\mathrm{GS}}(V_{\mathrm{GS}}, V_{\mathrm{GD}}) = C_{\mathrm{pgs}} + \frac{C_{\mathrm{gs1}}}{\left(1 - \frac{V_{\mathrm{GS}}}{V_{\mathrm{bi}}}\right)^{\mathrm{m}}} \\ + C_{\mathrm{gs2}} \{1 + \tanh[\kappa (V_{\mathrm{GS}} - V_{\mathrm{t2}})]\} \\ + C_{\mathrm{S}}(V_{\mathrm{GS}}) \{1 + \tanh[\iota (V_{\mathrm{GS}} - V_{\mathrm{GD}} - V_{\mathrm{t4}})]\} \\ - \frac{\delta C_{\mathrm{S}}(V_{\mathrm{GS}})}{\delta V_{\mathrm{GS}}} \left(V_{\mathrm{GD}} - \frac{1}{\iota} \ln \{\cosh[\iota (V_{\mathrm{GS}} - V_{\mathrm{GD}} - V_{\mathrm{t4}})]\} \right) \end{aligned} \tag{2.63}$$

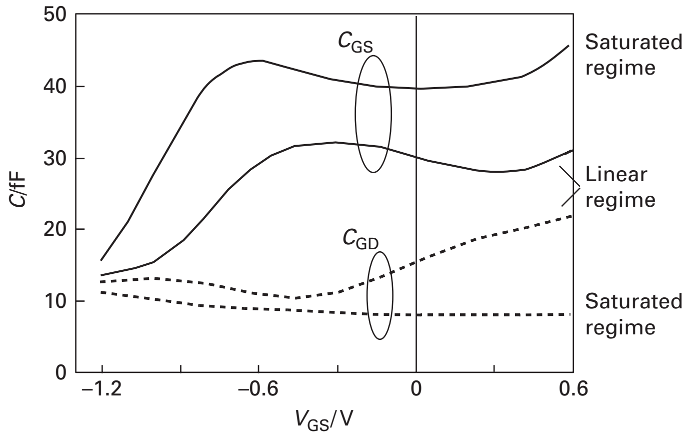


Fig. 2.39 Experimentally determined C_{GS} and C_{GD} of pseudomorphic GaAs HEMT ($L_G = 0.15\,\mu$m, $W_G = 2 \times 20\,\mu$m).

$$\begin{aligned} C_{GS}(V_{GS}, V_{GD}) = C_{pgd} + &\frac{C_{gd1}}{\left(1 - \frac{V_{GD}}{V_{bi}}\right)^m} \\ &+ C_{gd2}\{1 + \tanh[\kappa(V_{GD} - V_{t5})]\} \\ &- C_S(V_{GS})\{1 + \tanh[\iota(V_{GS} - V_{GD} - V_{t4})]\}\,. \end{aligned} \tag{2.64}$$

The capacitance

$$C_S(V_{GS}) = C_3 V_{eff}^{\psi}$$

with

$$V_{eff} = \frac{1}{2}\left(v_{GS} - V_{t3} + \sqrt{(V_{GS} - V_{t3})^2 + \theta^2}\right)$$

is closely related to the drain saturation current (compare Equation (2.59)). C_{gs1}, C_{gs2}, C_{gd1}, C_{gd2}, m, V_{bi}, V_{t2}, V_{t3}, V_{t4}, V_{t5}, ι, κ, θ and ψ are fitting parameters.

The non-linear capacitances, along with an additional parasitic drain-source capacitance C_{DS} and a parasitic channel resistance R_i, have been combined in Figure 2.40 to form a basic non-linear dynamic model of the HEMT. More bias-independent parasitic parameters may be added, as needed.

2.3.8 MESFET versus HEMT: a small-signal comparison

When the non-linear equivalent circuit in Figure 2.40 is linearised in a given bias point, the resulting small-signal equivalent circuit is identical to that derived for the MESFET in the previous section, with the exception of the domain capacitance C_{DC}, which is often neglected anyhow. Many results obtained for the MESFET can therefore be directly applied. Rather than repeating the results here, let us discuss how the achievable small-signal performance differs between MESFET and HEMT.

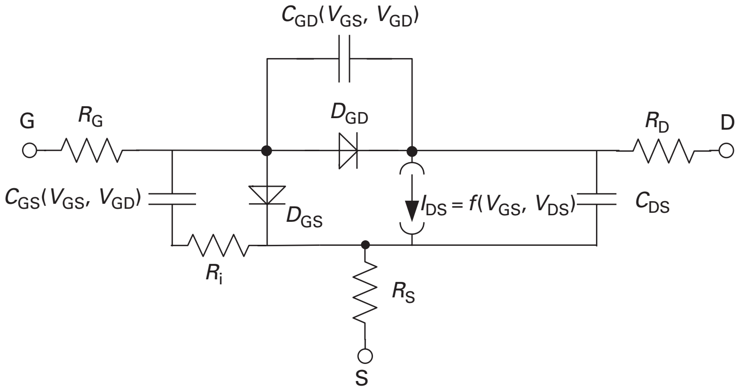


Fig. 2.40 A dynamic non-linear model of the HEMT.

As discussed, the higher low-field mobility affects the series resistances which represent semiconductor regions outside of the velocity-saturated channel. These are R_S, R_i and, of lesser importance, R_D. Equally, it increases the transconductance g_m (see Equation (2.53)).

The larger potential barrier between the channel and the substrate reduces the output conductance g_{ds} in the HEMT.

These findings directly translate into a significant advantage in terms of the maximum frequency of oscillation, f_{max}:

$$f_{max} = \frac{f_T}{2\sqrt{(R_G + R_S + R_i)g_{ds} + 2\pi f_T R_G C_{DG}}}.$$

The gate resistance R_G is, of course, independent of the device structure.

The HEMT structure also has a positive impact on the noise performance. This can be shown using the Fukui equation already introduced for the MESFET:

$$\begin{aligned} F_{min} &= 1 + k_F \frac{f}{f_T}\sqrt{g_m(R_G + R_S)} \\ &= 1 + k_F 2\pi f C_{GS}\sqrt{\frac{R_G + R_S}{g_m}}, \end{aligned}$$

using

$$f_T \sim \frac{g_m}{2\pi C_{GS}}.$$

The noise performance is improved not only by the reduction in R_S and the increase in g_m. The fitting factor k_F, which is typically 2.5 in MESFETs, decreases to $k_F = 1\ldots2$ in HEMTs. This is commonly explained by the higher correlation between channel noise and induced gate noise, and the reduction in channel noise due to the smaller degree of freedom of carrier movement in the 2DEG.

2.3.9 A practical HEMT example

The example chosen here is a metamorphic HEMT [4] because it illustrates many of the concepts discussed.

The device structure is shown in Figure 2.41. The rather thick (1 μm) linearly graded buffer adapts the lattice constant of the GaAs substrate to the much larger lattice constant of $In_{0.53}Ga_{0.47}As$ and $Al_{0.48}In_{0.52}As$ (the ternary compounds are lattice-matched to each other). Note the 'double doping' structure – there are δ-doped AlInAs layers below and above the InGaAs quantum well. In this case, it allows a density of the 2DEG of $n_{S,\max} = 4 \cdot 10^{12}\,\text{cm}^2$, together with the excellent carrier confinement in the quantum well. The excellent confinement is due to the large conduction band discontinuity between $In_{0.53}Ga_{0.47}As$ and $Al_{0.48}In_{0.52}As$.

The ohmic contacts are placed on an $In_{0.53}Ga_{0.47}As$ cap layer, which reduces the contact resistance and shields the metal–semiconductor interface from the Al-containing alloy, which is prone to formation of Al oxides at the exposed surface. The gate contact has a T shape which reduces the series resistance of the gate stripe and hence R_G. It is again shown in Figure 2.42.

The cross-section of the gate metallisation is significantly larger than what would be possible for a simple stripe with a 250 nm footprint, due to the T-gate structure. A refractory metal is used here so that the gate can be fabricated before the ohmic contacts – this allows an easy self-alignment of the ohmic contacts with respect to the T-gate structure, minimising the distance between the source and drain contacts and the channel, reducing R_S and R_D. The surface between the gate and the ohmic contacts is passivated by a SiN layer.

The plot in Figure 2.43 shows the drain current and transconductance of the device, normalised to 1 mm gate width, at $V_{DS} = 1$ V, which is well into the saturated regime for this device. The actual gate width of the characterised device is 20 μm. The g_m maximum is placed at $V_{GS} = 0$ – this is frequently done as it facilitates gate biasing. The threshold voltage is slightly below 0.6 V. The g_m depression at higher V_{GS} is also clearly visible.

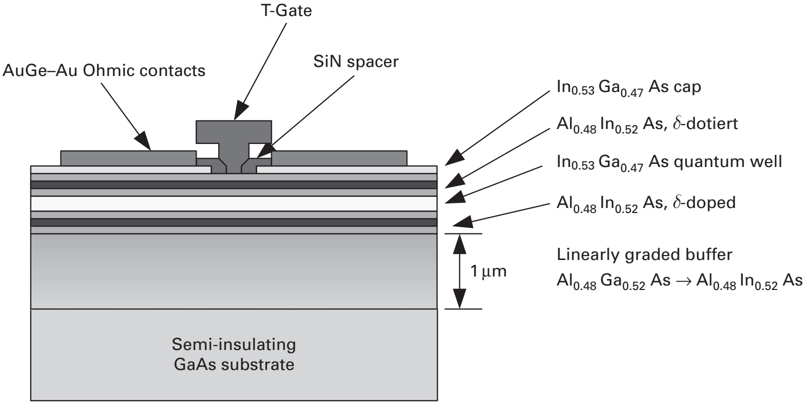

T-Gate
AuGe–Au Ohmic contacts
SiN spacer
$In_{0.53}Ga_{0.47}As$ cap
$Al_{0.48}In_{0.52}As$, δ-dotiert
$In_{0.53}Ga_{0.47}As$ quantum well
$Al_{0.48}In_{0.52}As$, δ-doped
1 μm
Linearly graded buffer
$Al_{0.48}Ga_{0.52}As \rightarrow Al_{0.48}In_{0.52}As$
Semi-insulating
GaAs substrate

Fig. 2.41 Layer structure of the metamorphic HEMT structure discussed here.

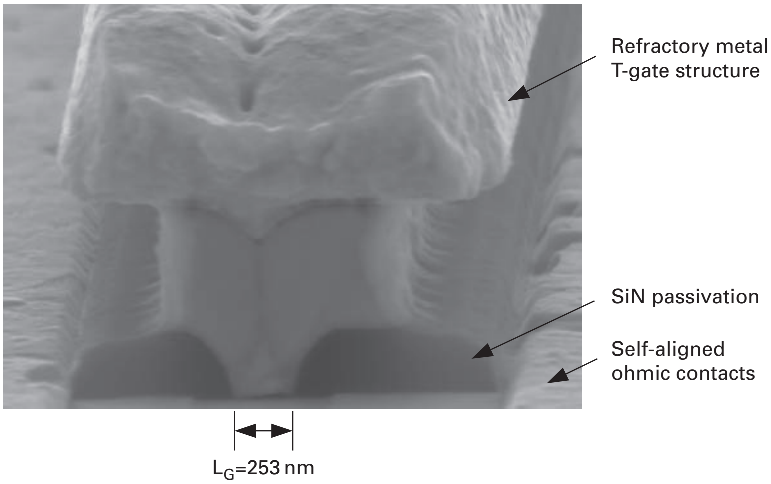

Refractory metal T-gate structure
SiN passivation
Self-aligned ohmic contacts
L_G=253 nm

Fig. 2.42 SEM micrograph of the gate structure (F. Benkhelifa, M. Chertouk, M. Dammann, M. Massler, H. Walther and G. Weimann, *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest*, May 2001).

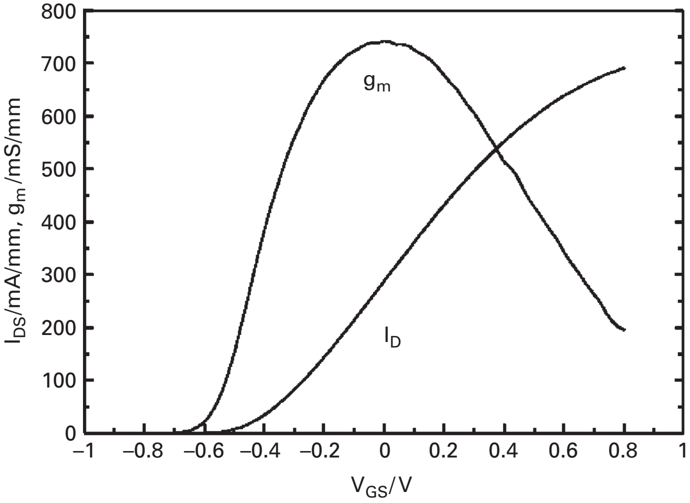


Fig. 2.43 Drain current (I_D) and transconductance (g_m) of the metamorphic HEMT, normalised to 1 mm gate width (F. Benkhelifa, M. Chertouk, M. Dammann, M. Massler, H. Walther and G. Weimann, *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest*, May 2001).

Figure 2.44, finally, shows the short-circuit current gain h_{21} as well as the maximum available gain (MAG) and the maximum stable gain (MSG) as a function of frequency, on a logarithmic scale. The current gain rolls off with an expected −20 dB/decade, and the transit frequency f_T, measured at $|h_{21}| = 0$ dB, is 110 GHz. The extraction of the claimed f_{max} of 300 GHz is less certain. As is explained in Chapter 5, f_{max} can be

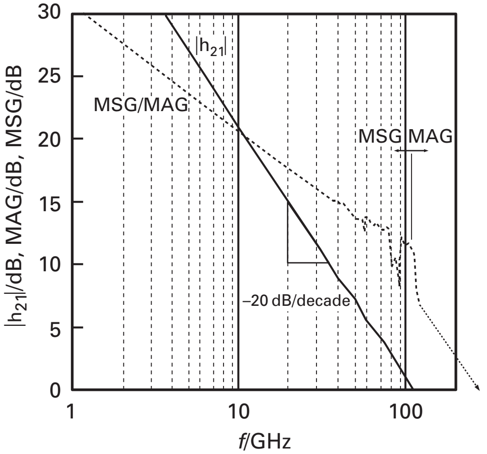


Fig. 2.44 Short-circuit current gain $|h_{21}|$, and power gains MAG and MSG, as a function of frequency, for the metamorphic HEMT structure (F. Benkhelifa, M. Chertouk, M. Dammann, M. Massler, H. Walther and G. Weimann, *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest*, May 2001).

extracted at the frequency where MAG $= 0$ dB. The problem is that MAG only exists where Rollet's stability factor $k > 1$, otherwise it is replaced by MSG. The change in slope of the MSG/MAG curve suggests that the transition between MSG and MAG happens only above 100 GHz, close to the upper end of the measurement range. From there, $f_{\max}$ seems to be extrapolated also at -20 dB/decade, even though the true roll-off is much steeper. The problem in determining $f_{\max}$ from MAG can be circumvented if an extraction from Mason's unilateral gain u is used. This is also explained elsewhere, and leads to different values for $f_{\max}$.

The important finding, however, is that using a metamorphic HEMT with an *optically defined* gate of $L_G = 0.25\,\mu$m provides sufficient gain for applications at 100 GHz.

2.4 Radio Frequency MOSFETs

2.4.1 Introduction

The silicon MOSFET is by a huge margin the most popular transistor structure. Long confined to either digital circuits or lower-frequency analogue applications, it now makes significant inroads into the realm of micro- and millimetre-wave circuits. With gate lengths below 100 nm, its cutoff frequencies f_T and $f_{\max}$ now rival those of the already introduced HEMTs or advanced HBTs, which will be introduced in the next section of this chapter.

We will briefly review the fundamental aspects of MOSFET operation and then proceed to the analogue aspects of RF CMOS operation which are commonly not covered in texts dealing primarily with MOSFETs as components in digital VLSI and ULSI.

The unparalleled success of silicon as the material of choice in fabricating electronic components is due to several factors:

- Silicon is cheap and has an almost limitless supply.
- Silicon is mechanically robust.
- Silicon has a high thermal conductivity, at least compared to GaAs and InP.
- But most importantly, silicon has a highly stable native oxide, SiO_2, which forms a high-quality interface with silicon.

This latter property led to the unequalled victory of metal-oxide–semiconductor (MOS) technology.

Basic MOSFET structure

Consider a somewhat schematic cross-section of a MOSFET (Figure 2.45).

It is not intended to do justice to the complexity of modern MOSFET devices, but shows their fundamental components. This is an *n-channel* device – the current in the channel will be carried by electrons. We will focus on n-channel devices here as this facilitates comparison with the previously discussed MESFETs and HEMTs (which are almost exclusively n-channel devices), but all findings relate analogously to p-channel devices as well, with appropriate modifications reflecting the differences in doping and free carrier type.

First, we note that the electron channel will actually form in a p-type semiconductor region, called the *bulk*. This can be either the substrate or a p-doped layer formed by epitaxy or diffusion. This will be explained in the next paragraph. Secondly, the stable SiO_2 is used in two different ways:

(i) as a thin *gate oxide* which covers the surface between the source (S) and drain (D) contacts and carries the gate (G) electrode on top; and
(ii) as a thick *field oxide* which covers the remainder of the structure.

The source and drain contacts are non-blocking (ohmic). The gate is physically separated from the semiconductor by the gate oxide – this arrangement is called an *MOS* (*m*etal *o*xide-*s*emiconductor) diode, even though the gate electrode in modern MOSFETs is actually not metallic, but fabricated from highly doped *polycrystalline Silicon* (poly-Si).

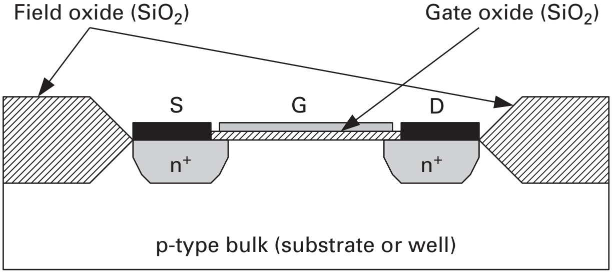


Fig. 2.45 Basic structure of an n-channel MOSFET.

MOS diode operation

Let us now investigate how an electron channel can form in the p-type semiconductor. To this end, we look more closely at the band diagram in a region below the gate electrode. Initially, no external voltages shall be applied to the device.

We construct the band diagram of the MOS diode (Figure 2.46), using *Anderson's rule*.

In the n$^+$-doped poly-Si gate, we assume that the conduction band energy E_G coincides with the Fermi energy E_F. The bulk Si is p-doped, and we calculate the distance between the Fermi level and the valence band energy E_V, assuming that the Boltzmann approximation to the Fermi–Dirac statistics is valid:

$$E_F - E_V = kT \ln\left(\frac{N_V}{N_A}\right), \tag{2.65}$$

where N_V is the density of states in the valence band and N_A is the acceptor concentration in the p-Si.

The SiO$_2$ is handled as a semiconductor with a very large band gap.

The distance between the conduction bands and the vacuum level, E_{vac} is given by the electron affinities in the Si and SiO$_2$ – χ_{Si} and χ_{SiO_2}, respectively. Note $\chi_{Si} > \chi_{SiO_2}$.

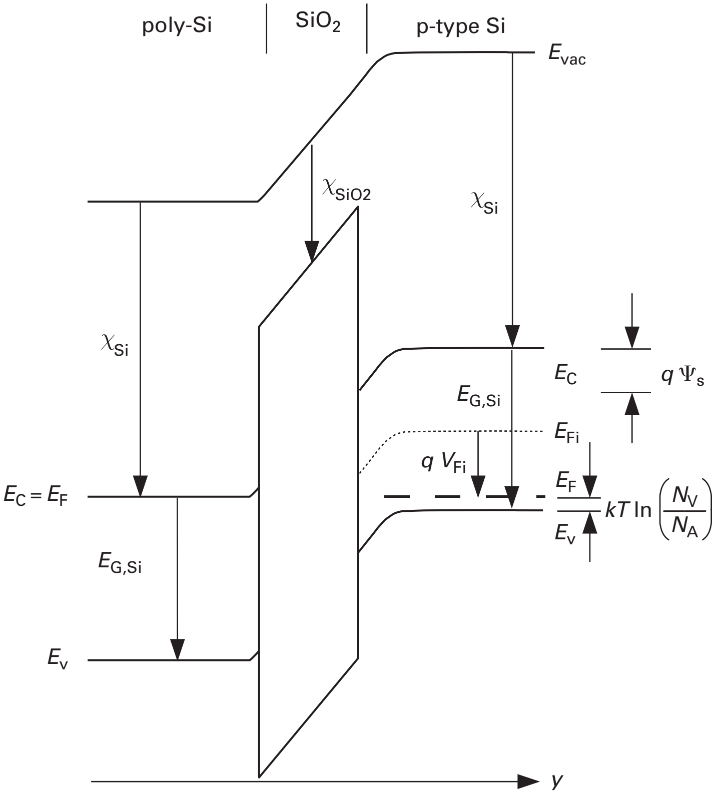


Fig. 2.46 Band diagram of an MOS diode structure.

The continuity of the vacuum level (postulated by Anderson's rule), along with the fact that the poly-Si is n-type, makes the bands in the p-type Si 'dip' towards the SiO_2/Si interface.

We introduced two new potentials and their corresponding energies:

(i) $q\,V_{\mathrm{Fi}}$ is the energy difference between the Fermi levels for the intrinsic and doped semiconductor:

$$q\,V_{\mathrm{Fi}} = \frac{E_{\mathrm{C}}+E_{\mathrm{V}}}{2} + \frac{1}{2}kT\ \ln\frac{N_{\mathrm{V}}}{N_C} - E_{\mathrm{V}} - kT\ \ln\frac{N_{\mathrm{V}}}{N_{\mathrm{A}}} \approx \frac{E_{\mathrm{G}}}{2} - kT\ \ln\frac{N_{\mathrm{V}}}{N_{\mathrm{A}}}. \quad (2.66)$$

(ii) $q\,\Psi_{\mathrm{s}}$ is the energy difference between the undisturbed semiconductor and the Si/SiO_2 interface.

Using these two potentials, we can easily distinguish four different regions:

(i) $\Psi_{\mathrm{s}} < 0$: The bands 'bend upwards' – *accumulation of holes* at the interface – creation of a positive space charge of mobile carriers there.

(ii) $0 < \Psi_{\mathrm{s}} \le V_{\mathrm{Fi}}$: *Depletion of holes* at the interface – creation of a negative space charge of fixed carriers. The charges are the ionised acceptor atoms. Increase of Ψ_{S} results in an extension of the space charge layer into the semiconductor. In the limit $\Psi_{\mathrm{s}} = V_{\mathrm{Fi}}$, the interface behaves like an intrinsic semiconductor.

(iii) $V_{\mathrm{Fi}} < \Psi_{\mathrm{s}} \le 2\,V_{\mathrm{Fi}}$: As the Fermi level is now closer to the conduction band than to the valence band at the interface, the conduction type converts from p-type to n-type. This condition is called *light inversion*.

The density of *minority electrons* at the interface increases exponentially with Ψ_{s}:

$$n_{\mathrm{p}} = n_{\mathrm{i}}\, e^{q(\Psi_{\mathrm{s}}-V_{\mathrm{Fi}})/kT}.$$

(iv) At $\Psi_{\mathrm{s}} > 2\,V_{\mathrm{Fi}}$ the interface carrier density rises sharply for small changes in Ψ_{s}, which remains almost constant for large changes in the interface charge. This condition is called *strong inversion*. The *width of the depletion region* stays approximately constant at

$$w_{\max} = 2\sqrt{\frac{\epsilon_{\mathrm{Si}}}{q\,N_{\mathrm{A}}}\,V_{\mathrm{Fi}}}, \quad (2.67)$$

for a homogeneously doped semiconductor.

In Figure 2.46, $0 < \Psi_{\mathrm{s}} < V_{\mathrm{Fi}}$, hence the device is in depletion without externally applied voltages and no channel forms. This is typical of n-channel MOSFET transistors – they have positive threshold voltages, in contrast to MESFETs and HEMTs.

As the gate electrode is isolated from the semiconductor by the gate oxide, we can apply large positive gate-channel voltages without creating a gate current, as would be the case in Schottky diodes. This situation is shown in Figure 2.47.

Note that the conduction band forms a *triangular potential well* at the Si/SiO_2 interface. In this respect, the MOSFET is closely related to the HEMT and will equally form a two-dimensional electron gas in the channel.

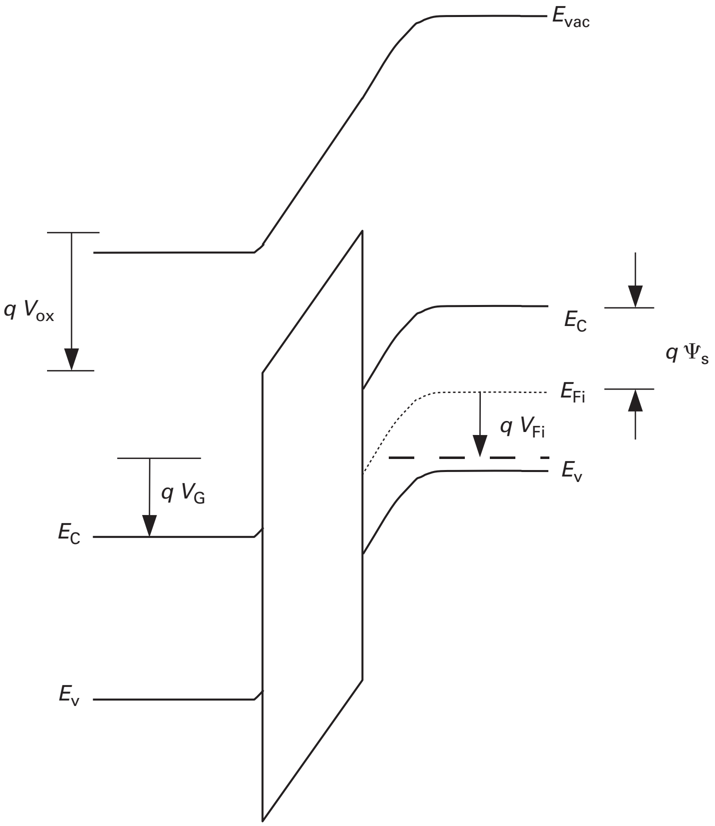


Fig. 2.47 MOS diode band diagram with positive V_G.

The externally applied gate voltage drops partially across the gate oxide, partially across the semiconductor and increases Ψ_s. In the situation drawn in Figure 2.47, the hypothetical intrinsic Fermi level already drops below the Fermi level in the undisturbed p-type semiconductor ($\Psi_s > V_{Fi}$), but $\Psi_s > 2V_{Fi}$ has not been reached – the structure is in light inversion. An even more positive V_G will introduce strong inversion.

If V_{ox} is the voltage drop across the oxide,

$$V_G = V_{ox} + \Psi_s + \frac{1}{q}\left(kT\ \ln\frac{N_V}{N_A} - E_G\right), \tag{2.68}$$

The last term in Equation (2.68) is the *flat-band voltage* V_{FB}:

$$V_{FB} =: \frac{1}{q}\left(kT\ \ln\frac{N_V}{N_A} - E_G\right). \tag{2.69}$$

It is called *flat-band* because for $V_G = V_{FB}$, $\Psi_s + V_{ox} = 0$ and the bands become completely horizontal.

V_{ox} can be calculated from the *gate capacitance* C_{ox} and the *total charge stored under the gate* Q_{s}:

$$V_{\mathrm{ox}} = -\frac{Q_{\mathrm{s}}}{C_{\mathrm{ox}}}.$$

The total gate charge is formed by the *space charge in the depleted region* Q_{B} and the *interface charge* Q_{i}.

Let us now consider the case where $\Psi_{\mathrm{s}} = 2\,V_{\mathrm{Fi}}$ just occurs (onset of strong inversion). Q_{i} is negligible, and hence $Q_{\mathrm{s}} \approx Q_{\mathrm{B}}$ with

$$Q_{\mathrm{B}} = -q\,N_{\mathrm{A}}\,w_{\mathrm{max}}\,L_{\mathrm{G}}\,W_{\mathrm{G}},$$

where w_{max} is the maximum extension of the space charge region, equal to the extension for $\Psi_{\mathrm{s}} = 2\,V_{\mathrm{Fi}}$ (see Equation (2.67)) and $W_{\mathrm{G}}\,L_{\mathrm{G}}$ is the gate footprint, which determines the area of the channel. Hence,

$$Q_{\mathrm{B}} = -2\sqrt{\epsilon_{\mathrm{Si}}\,q\,N_{\mathrm{A}}\,V_{\mathrm{Fi}}}.$$

We can now calculate the *threshold voltage* V_{th} as the necessary V_{G} to reach the onset of strong inversion. Recalling that at this point $\Psi_{\mathrm{s}} = 2\,V_{\mathrm{Fi}}$, Equation (2.68) yields

$$\begin{aligned} V_{\mathrm{th}} &= -\frac{Q_{\mathrm{B}}}{C_{\mathrm{ox}}} + 2\,V_{\mathrm{Fi}} + V_{\mathrm{FB}} \\ &= \frac{2\,W_{\mathrm{G}}\,L_{\mathrm{G}}}{C_{\mathrm{ox}}}\sqrt{\epsilon_{\mathrm{Si}}\,q\,N_{\mathrm{A}}\,V_{\mathrm{Fi}}} + 2\,V_{\mathrm{Fi}} + V_{\mathrm{FB}}. \end{aligned} \tag{2.70}$$

In strong inversion, the oxide capacitance is easy to calculate, because the mobile charge Q_{i} is concentrated as a sheet charge at the Si/SiO$_2$ interface (compare the situation in the HEMT). The sheet charge forms a simple parallel-plate capacitor with the gate electrode:

$$C_{\mathrm{ox}} = W_{\mathrm{G}}\,L_{\mathrm{G}}\frac{\epsilon_{\mathrm{SiO_2}}}{t_{\mathrm{ox}}}, \tag{2.71}$$

where t_{ox} is the gate oxide thickness.

The threshold voltage is then approximately:[5]

$$V_{\mathrm{th}} = \frac{2\,t_{\mathrm{ox}}}{\epsilon_{\mathrm{SiO_2}}}\sqrt{\epsilon_{\mathrm{Si}}\,q\,N_{\mathrm{A}}\,V_{\mathrm{Fi}}} + 2\,V_{\mathrm{Fi}} + V_{\mathrm{FB}}. \tag{2.72}$$

2.4.2 Drain current

So far, we considered only the voltage between gate and channel, assuming that the drain and source electrodes are on equal potentials. Now, we apply external voltages $V_{\mathrm{GS}}, V_{\mathrm{DS}} \neq 0$, as in Figure 2.48. As before in the discussion of MESFET and HEMT, we make certain assumptions for the channel:

- The channel shall be one-dimensional, i.e. the electric field has only a z component.

[5] Because $V_{\mathrm{G}} = V_{\mathrm{th}}$, the MOS diode is not strictly in strong inversion.

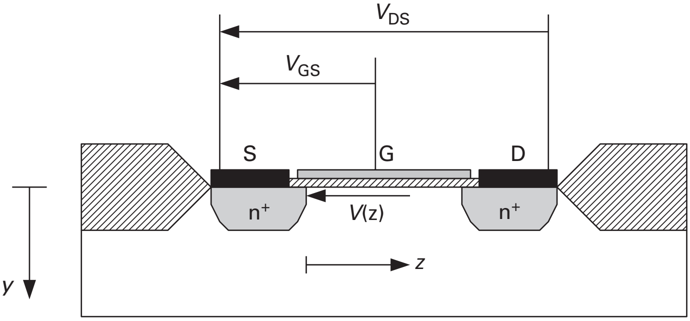


Fig. 2.48 MOSFET structure with externally applied voltages.

- The channel shall be gradual, i.e. the current is driven purely by the electric field and diffusion is neglected.

Now that $V_{DS} > 0$, the voltage between the gate electrode and the semiconductor will depend on the z coordinate along the interface:

$$V_G(z) = V_{GS} - V(z). \tag{2.73}$$

Constant-mobility model

As discussed already, the channel current can be calculated from the local mobile charge and the velocity with which it moves. In case of the MOSFET, the local charge q_i is the mobile interface charge which in strong inversion can be calculated simply from the oxide capacitance and the local gate-channel voltage $V_G(z)$:

$$q_i(z) = \frac{\epsilon_{SiO_2}}{t_{ox}}[V_G(z) - V_{th}]. \tag{2.74}$$

We initially calculate the charge velocity for the low-field case, assuming that $\mu_n =$ const and find

$$I_D(z) = W_G \frac{\epsilon_{SiO_2}}{t_{ox}}[V_G(z) - V_{th}]\,\mu'_n \frac{dV(z)}{dz},$$

where μ'_n is the interface mobility, which is lower than the bulk mobility due to the imperfections of the interface plane.

Applying current continuity, we know that

$$I_D(z) = \text{const} = I_D = \frac{1}{L_G}\int_{z=0}^{z=L_G} I_D(z)dz.$$

Using parameter substitution and noting that $V(z = 0) = 0$, $V(z = L_G) = V_{DS}$, we find

$$I_D(V_{GS}, V_{DS}) = \frac{\epsilon_{SiO_2}}{t_{ox}}\frac{W_G}{L_G}\,\mu'_n \left[(V_{GS} - V_{th})\,V_{DS} - \frac{V_{DS}^2}{2}\right]. \tag{2.75}$$

The above equation only holds as long as the channel is not fully depleted. Because $V(z)$ increases monotonically with z along the channel, depletion of the channel will start at the drain, when V_{DS} reaches the knee voltage V_{k}, in full analogy to the MESFET and HEMT.

$$q_i(z = L_{\mathrm{G}}) = \frac{\epsilon_{\mathrm{SiO_2}}}{t_{\mathrm{ox}}} (V_{\mathrm{GS}} - V_{\mathrm{k}} - V_{\mathrm{th}}) = 0$$

yields

$$V_{\mathrm{k}} = V_{\mathrm{GS}} - V_{\mathrm{th}} \tag{2.76}$$

For $V_{\mathrm{DS}} > V_{\mathrm{k}}$, the channel charge no longer depends on V_{DS} in this simple model. With $V_{\mathrm{DS}} = V_{\mathrm{k}}$ and using Equation (2.76), we find from Equation (2.75)

$$I_{\mathrm{D}}(V_{\mathrm{GS}}) = \frac{\epsilon_{\mathrm{SiO_2}}}{t_{\mathrm{ox}}} \frac{W_{\mathrm{G}}}{2\,L_{\mathrm{G}}} \mu'_{\mathrm{n}} (V_{\mathrm{GS}} - V_{\mathrm{th}})^2, \tag{2.77}$$

for $V_{\mathrm{DS}} > V_{\mathrm{k}}$.

This simple model of the MOSFET static behaviour is often referred to as the *Shockley model* [57].

Backgating

Another parasitic effect influencing the static performance needs to be considered. Because the MOSFET sits on a conducting silicon layer (the 'bulk', p-type for n-channel, n-type for p-channel transistors), the device is essentially a four-terminal device, where the bulk is the fourth terminal. We had implicitly assumed that the bulk layer would have a fixed potential, which is that of the source contact. In an integrated circuit, however, this cannot always be maintained. Therefore, we need to consider a second control voltage, the bulk-source voltage V_{BS}.

Recall that we defined the threshold voltage via the potential difference between the bulk and the Si/SiO$_2$ interface. This suggests a very simple way of accommodating backgating – by modifying the threshold voltage:

$$V_{\mathrm{th}} = V_{\mathrm{t0}} - \gamma\, V_{\mathrm{BS}}. \tag{2.78}$$

Here, V_{t0} is the threshold voltage without backgating (i.e. the one considered so far) and γ is a fitting parameter [49].

Non-ideal effects in short-channel MOSFETs

Channel length modulation.

So far, we neglected another effect: the source and drain regions form n–p junctions with the bulk semiconductor. These n–p junctions necessarily create depletion regions, whose width depends on the voltage across the junction. In an n-channel MOSFET, the drain has a positive potential with respect to source. Assuming that the bulk is held on source potential ($V_{\mathrm{BS}} = 0$), the drain-bulk region is therefore reverse-biased, which leads to an increase in the width of the space charge region there.

Figure 2.49 shows this situation. The effective gate length L_{eff} is shorter than the 'drawn' gate length L_{G}. The difference is V_{DS}-dependent due to the drain space charge region:

$$L_{\mathrm{eff}} = L_{\mathrm{G}} - \Delta L(V_{\mathrm{DS}}). \tag{2.79}$$

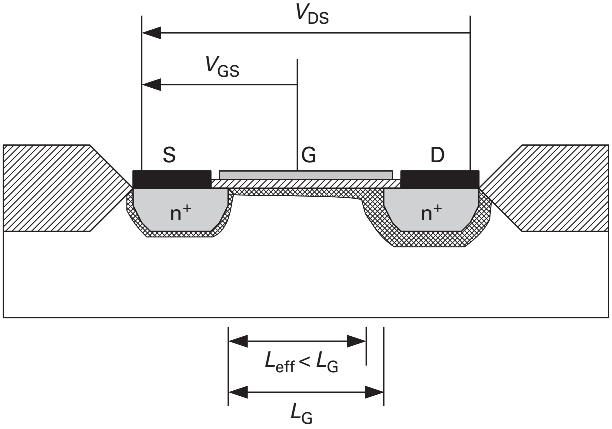


Fig. 2.49 Schematic representation of channel length modulation due to space charge incursion into the channel.

ΔL can be approximated as follows:

$$\Delta L = \frac{1}{2}\sqrt{\frac{2\epsilon}{q\,N_A}(V_{DS} - V_{off})\,\alpha}, \tag{2.80}$$

where α is a factor which depends on the exact geometry of the MOSFET – $\alpha = 0.02\ldots1$. This formula only applies for $V_{DS} > V_{off}$, only then will a part of the channel close to drain be fully depleted.

Recalling Equation (2.77), it is easy to see that the progressive reduction of the effective channel length with increasing V_{DS} will cause the drain current to increase with increasing drain-source voltage. This effect is most pronounced if the 'geometrical' channel length L_G is already small – hence this is a very important effect in high-speed MOSFETs with channel lengths $L_G < 0.5\,\mu$m. This effect is called *channel length modulation* and is conceptually very similar to the *Early effect* which we will introduce for the bipolar transistor (see p. 122).

Short-channel effect.

The calculation of the threshold voltage (Equation (2.72)), assumed that the gate and the mobile sheet charge in the channel form an ideal parallel-plate capacitor: the total interface charge Q_i appears, with opposite sign, at the gate charge Q_G. In reality, some of the field lines emanating from the negative charge in the channel may also terminate on the source and drain areas. Due to the n–p-junctions, the depletion of the channel region progresses more rapidly than predicted from considering the gate potential alone, which leads to a reduction in threshold voltage V_{th}. This is the *short-channel effect* proper, which says that for otherwise unchanged technological parameters, the threshold voltage will decrease with decreasing gate length. The effect is more pronounced, the deeper the source and drain contact regions extend into the bulk material.

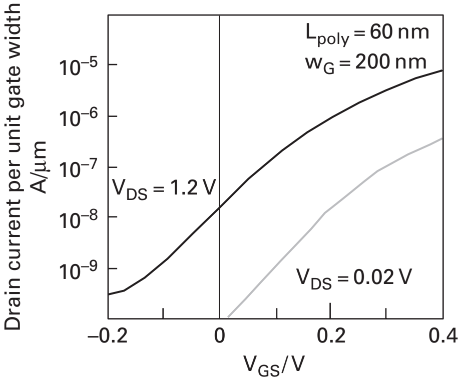


Fig. 2.50 Example for the V_{DS} dependence of subthreshold currents in deep-submicron MOSFETs (T. Sugli, K. Watanabe and S. Sugatani, *Fujitsu Science and Technology Journal*, Vol. 39, No. 6, pp. 9–22, June 2003.).

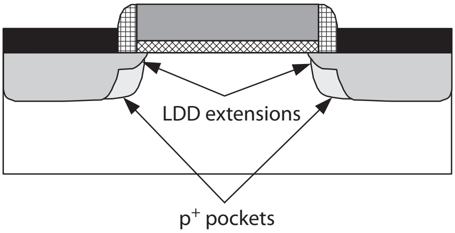


Fig. 2.51 MOSFET structure with a double implant LDD arrangement.

As the shape, specifically of the drain side space charge region, depends on the potential of the drain contact, it is not surprising that V_{DS} also has an effect on V_{th}: as the drain-source voltage is increased, the drain field will deplete the channel more, which leads to a further decrease of the threshold voltage.

The effect of the drain field can best be shown in the *subthreshold regime* (see Figure 2.50). For $V_{GS} < V_{th}$, the channel current does not actually cease to flow, because even before the onset of strong inversion, there are free charge carriers in the channel. Their density and hence the current depend exponentially on $V_{GS} - V_{th}$. The figure shows an example of the subthreshold regime for a deep-submicron MOSFET, for two different values of V_{DS}. We note that even a small increase in V_{DS} increases $I_{off} = I_D(V_{GS} = 0)$ but two orders of magnitude.

A very common modification of the standard MOSFET structure which reduces the short-channel effect and channel length modulation, and also improves the breakdown voltage, is the double implant lightly doped drain (LDD) structure [41] shown in Figure 2.51.

The shallow n-doped drain extensions lower the maximum electric field in the channel and hence increase the breakdown voltage, while the p^+-doped pockets slow the growth of the p–n space charge regions into the channel with increasing V_{DS}.

Mobility degradation.

Short-channel devices benefit from an increase in the bulk doping concentration, because the problem with the high I_{off} and also the channel length modulation can be decreased by increasing the bulk doping. Together with the thin gate oxide, however, this significantly increases the electric field component in y direction. The charge carriers in the channel will then flow, on average, closer to the Si/SiO$_2$ interface where their mobility is reduced by interface scattering, due to imperfections of the interfacial layer. Therefore, the effective mobility will decrease with increasing V_{GS}, approximated by

$$\mu'_{\text{n}}(V_{\text{GS}}) = \frac{\mu'_{\text{n},0}}{1 + m\,(V_{\text{GS}} - V_{\text{th}})}, \tag{2.81}$$

where $\mu'_{\text{n},0}$ is the mobility at threshold and m is a factor describing the degree of *normal-field mobility degradation*.

In summary, in short-gatelength MOSFETs, the simple one-dimensional approach we started out with is no longer adequate, hence two-dimensional effects have to be incorporated into the physical simulation.

Velocity saturation

As in the typically GaAs-based MESFET and HEMT devices, velocity saturation at high electric fields also has to be considered here. The critical field $\mathcal{E}_{\text{sat}}$ for velocity saturation in silicon is $\sim 4 \cdot 10^6\,\text{Vm}^{-1}$ at room temperature, compared to $3\ \cdot 10^5\,\text{Vm}^{-1}$ for GaAs, so the onset of velocity saturation is delayed in Si versus GaAs.

If we assume that the channel is fully velocity saturated, i.e. $v_{\text{n}} = v_{\text{sat}} \neq f(z)$, the drain current becomes

$$I_{\text{D}} = W_{\text{G}} \frac{\epsilon_{\text{SiO}_2}}{t_{\text{ox}}} v_{\text{sat}} (V_{\text{GS}} - V_{\text{th}}), \tag{2.82}$$

where v_{sat} is the drift saturation velocity of silicon, which is approximately 10^5 m/s at room temperature.

In the intermediate region, Lee [34] gives the following approximation for the drain current:

$$I_{\text{D}} = W_{\text{G}} \frac{\epsilon_{\text{SiO}_2}}{t_{\text{ox}}} \frac{v_{\text{sat}}}{1 + \frac{L_{\text{G}}\mathcal{E}_{\text{sat}}}{V_{\text{GS}} - V_{\text{th}}}}, \tag{2.83}$$

for $V_{\text{DS}} > V_{\text{D,sat}}$, where $V_{\text{D,sat}}$ is the necessary field for velocity saturation to occur in the channel, approximated as

$$V_{\text{D,sat}} \approx \frac{(V_{\text{GS}} - V_{\text{th}})\,L_{\text{G}}\,\mathcal{E}_{\text{sat}}}{(V_{\text{GS}} - V_{\text{th}}) + L_{\text{G}}\,\mathcal{E}_{\text{sat}}}. \tag{2.84}$$

Consider a modern MOSFET with $L_{\text{G}} = 0.09\,\mu\text{m}$. $L_{\text{G}}\,\mathcal{E}_{\text{sat}}$ is 0.36 V, and assuming $V_{\text{GS}} - V_{\text{th}} = 0.5$ V, we arrive at $V_{\text{D,sat}} = 0.21$ V – considerably smaller than $V_{\text{k}} = V_{\text{GS}} - V_{\text{th}} = 0.5$ V, as the Shockley model would predict. Velocity saturation is hence a phenomenon with significant importance in deep-submicron MOSFETs.

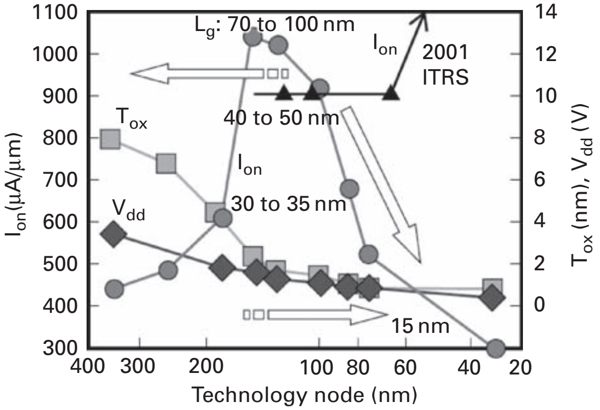


Fig. 2.52 Maximum drain current ('on-current') I_{on}, gate oxide thickness T_{ox} and supply voltage V_{dd} as a function of the technology node (T. Sugli, K. Watanabe and S. Sugatani, *Fujitsu Science and Technology Journal*, Vol. 39, No. 6, pp. 9–22, June 2003.).

Equation (2.82) can also be used to estimate the maximum drain current in a given MOSFET family. In devices with $L_G \leq 130\,\mathrm{nm}$, $t_{\mathrm{ox}} \approx 2\,\mathrm{nm}$ typically. For a gate overtravel $V_{\mathrm{GS}} - V_{\mathrm{th}}$ of 1 V and $V_{\mathrm{sat}} = 10^5\,\mathrm{m\,s^{-1}}$, we find $I_{\mathrm{D}}/W_{\mathrm{G}} = 1.72\,\mathrm{mA\,\mu m^{-1}}$.

In order to increase the current for a given gate over travel, we have only two options:

(i) Decrease the thickness t_{ox} of the gate oxide. However, as t_{ox} is reduced, the electric field in the dielectric increases, and the gate-channel tunnel current increases dramatically.

(ii) Increase the dielectric constant of the gate dielectric. This can be done by replacing the SiO$_2$ with a different dielectric, such as HfO$_2$, which features $\epsilon_{\mathrm{HfO_2}} = 25\,\epsilon_0$ instead of $\epsilon_{\mathrm{SiO2}} = 3.9\,\epsilon_0$, but with limited thermal stability. Additionally, a major advantage of Si, namely its highly stable native oxide, is given up.

Figure 2.52 presents a literature data review [61] of achieved maximum drain current, gate oxide thickness and power supply voltage as a function of the target technology, as of the year 2003. The decrease in the maximum drain current ('on current', I_{on}) and the supply voltage V_{dd} give proof to the problems CMOS designers face due to the reducing gate oxide thickness.

2.4.3 Large-signal modelling

A MOSFET's non-linear circuit (Figure 2.53), is very similar to what we discussed for the MESFET, or HEMT, except that the substrate ('bulk') node needs to be accounted for. The diodes D_{BS} and D_{BD} represent the source and drain p–n diodes, respectively, and include junction capacitance. The series resistances R_{G}, R_{S} and R_{D} are taken as bias-independent.

As discussed in the previous paragraph, the drain current I_{D} will be a function of V_{GS} and V_{DS}. In a more precise model, we need to make the threshold voltage a function of V_{DS} and via backgating also of V_{BS}, so the bulk-source voltage needs to be included as

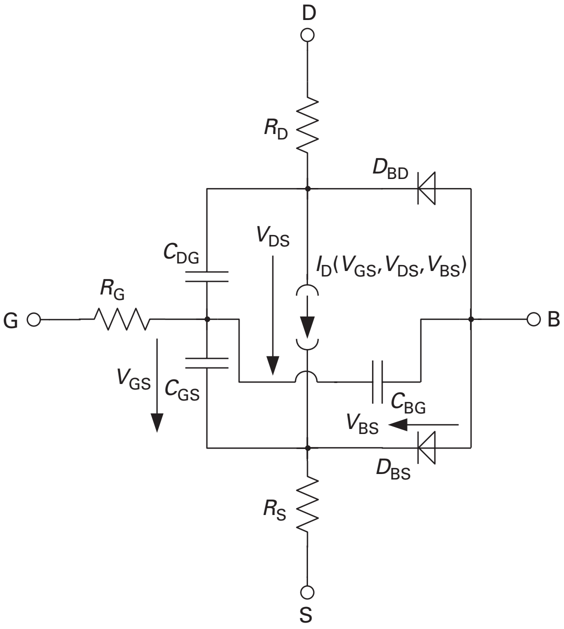


Fig. 2.53 Non-linear equivalent circuit suitable for MOSFETs.

a controlling voltage as well. Please note that the voltages occur between the internal nodes – voltage drops across the series resistances will have to be subtracted.

A particular feature of MOS transistors are the *overlap capacitances*. Referring, for example, to Figure 2.45, note that the gate electrode overlaps the highly doped source and drain regions. Without this overlap, at least on the source side, the channel could not form, as the free charge in the channel is drawn from the source region – unlike in the HEMT, where the free carriers are being introduced through the supply layer on top of the channel. These capacitances will be bias-independent, so that the gate-source and drain-source capacitances can be written as

$$\begin{aligned} C_{GS} &= C_{GSO} + \frac{\delta Q_B}{\delta V_{GS}} \\ C_{GD} &= C_{GDO} + \frac{\delta Q_B}{\delta V_{GD}}, \end{aligned} \tag{2.85}$$

where C_{GSO} and C_{GDO} are the gate-source and gate-drain overlap capacitances and Q_B is the total space charge, which includes the fixed charge ($q\, N_A\, w$) and the mobile interface charge Q_i.

In strong inversion, the change in channel charge is reflected only in the interface charge:

$$\delta Q_B \approx \delta Q_i = W_G \int_{z=0}^{z=L_G} q_i(z) dz.$$

Equation (2.74) indicates that in strong inversion, the total interface charge will only depend on V_{GS}; therefore, C_{GD} will be given only by the overlap capacitance. Non-ideal effects can be included into the capacitance equation by making the threshold voltage V_{th} V_{DS}- and V_{BS}-dependent.

The Meyer capacitance model [38] already introduced for the MESFET is frequently used to model the bias dependence of C_{GS} and C_{GD}. In strong inversion,

- For $V_{\mathrm{DS}} < V_{\mathrm{k}}$,

$$\begin{aligned} C_{\mathrm{GS}} &= C_{\mathrm{GSO}} + \frac{2}{3} C_{\mathrm{GC}} \left[1 - \left(\frac{V_{\mathrm{k}} - V_{\mathrm{DS}}}{2V_{\mathrm{k}} - V_{\mathrm{DS}}} \right)^2 \right] \\ C_{\mathrm{GD}} &= C_{\mathrm{GDO}} + \frac{2}{3} C_{\mathrm{GC}} \left[1 - \left(\frac{V_{\mathrm{k}}}{2V_{\mathrm{k}} - V_{\mathrm{DS}}} \right)^2 \right] \end{aligned} \tag{2.86}$$

- For $V_{\mathrm{DS}} > V_{\mathrm{k}}$,

$$\begin{aligned} C_{\mathrm{GS}} &= C_{\mathrm{GSO}} + \frac{2}{3} C_{\mathrm{GC}} \\ C_{\mathrm{GD}} &= C_{\mathrm{GDO}}, \end{aligned} \tag{2.87}$$

where V_{k} is the drain-source voltage delineating the linear from the saturated regime and (see Equation (2.76)) C_{GC} is the gate-channel capacitance for $V_{\mathrm{DS}} = 0$. In strong inversion, it is simply the total oxide capacitance (see Equation (2.71)).

The gate-bulk capacitance can be similarly expressed; it models the effect of the bulk potential on the channel charge:

$$C_{\mathrm{BG}} = \frac{\delta Q_{\mathrm{B}}}{\delta V_{\mathrm{BG}}}.$$

In strong inversion, it can be neglected.

The model may be extended with additional elements. Particularly, in RF designs the modelling of the impedance connected to the substrate node deserves particular attention.

At the core of most submicron RF CMOS models is the BSIM3 model, developed at University of California, Berkeley.[6] Unlike e.g. the COBRA model introduced for HEMTs, it uses different sets of equations for the MOSFET's different operating regions. It is also more closely related to device physics, i.e. it is not strictly an empirical model, and its input parameters are partly technological and partly empirical fitting parameters. BSIM's complexity, however, is beyond the scope of a book like this.

A simpler model approach, yet useful for many applications, was published by Sakurai and Newton [49]. The threshold voltage is

$$V_{\mathrm{th}} = V_{\mathrm{t0}} + \gamma \left(\sqrt{2\Phi_{\mathrm{F}} - V_{\mathrm{BS}}} - \sqrt{2\Phi_{\mathrm{F}}} \right), \tag{2.88}$$

where V_{t0}, γ and Φ_{F} are model parameters and V_{BS} is the bulk-source voltage – the above equation therefore includes backgating.

[6] Web resource at www-device.eecs.berkeley.edu/bsim3/

The drain-source saturation voltage, V_k, is modelled as

$$V_k = K\,(V_{GS} - V_{th})^m, \tag{2.89}$$

where K and m are model parameters. This formulation is more flexible compared to Equation (2.76) and allows to include velocity saturation effects.

The drain current at $V_{DS} = V_k$ is

$$I_{D,sat} = \frac{W_G}{L_{eff}}\,B\,(V_{GS} - V_{th})^n, \tag{2.90}$$

where B and n are model parameters. With variation of n, we can address both the constant-mobility model ($n = 2$) and the constant velocity model ($n = 1$); however, as n is not V_{DS}-dependent, we cannot move from one regime to the other.

A V_{GS}-dependent n, on the other hand, would allow to include the deterioration of mobility at high electric fields normal to the Si/SiO$_2$ interface (see Equation (2.81)).

The drain current formulation distinguishes between the saturated and non-saturated regions:

- For $V_{DS} > V_k$,

$$I_D = I_{D,sat}\,(1 + \lambda\,V_{DS}), \tag{2.91}$$

where $\lambda = \lambda_0 - \lambda_1\,V_{BS}$.

- For $V_{DS} < V_k$,

$$I_D = I_{D,sat}\,(1 + \lambda\,V_{DS})\left(2 - \frac{V_{DS}}{V_k}\right)\frac{V_{DS}}{V_k}. \tag{2.92}$$

2.4.4 Small-signal model and RF performance

The small-signal equivalent circuit of the MOSFET is very similar to that used for MESFET or HEMT, but must account for the additional substrate node.

Figure 2.54 shows that a fourth terminal (B) has been added, which is capacitively coupled to the internal gate, source and drain nodes. C_{BG} is shown to facilitate comparison with Figure 2.53; in saturation it is neglected.

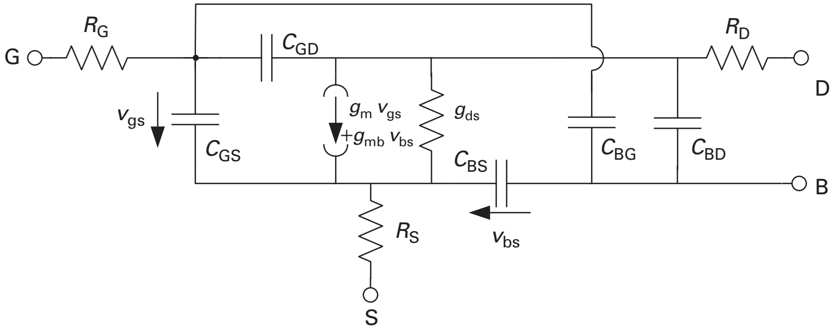


Fig. 2.54 MOSFET small-signal equivalent circuit.

Backgating is included by a special backgating transconductance:

$$g_{\mathrm{mb}} = \frac{\delta I_{\mathrm{D}}}{\delta V_{\mathrm{BS}}}.$$

The transconductance in the constant-mobility limit and for $V_{\mathrm{DS}} > V_{\mathrm{k}}$ is

$$g_{\mathrm{m}} = \frac{d I_{\mathrm{D}}}{d V_{\mathrm{GS}}} = \frac{\epsilon_{\mathrm{SiO_2}}}{t_{\mathrm{ox}}} \frac{W_{\mathrm{G}}\, \mu'_{\mathrm{n}}}{L_{\mathrm{G}}} (V_{\mathrm{GS}} - V_{\mathrm{th}}), \tag{2.93}$$

using Equation (2.77).

In the constant-velocity limit,

$$g_{\mathrm{m}} = \frac{\epsilon_{\mathrm{SiO_2}}}{t_{\mathrm{ox}}} W_{\mathrm{G}}\, v_{\mathrm{sat}}. \tag{2.94}$$

The output conductance is

$$g_{\mathrm{ds}} = \frac{\delta I_{\mathrm{D}}}{\delta V_{\mathrm{DS}}}.$$

In the linear region ($V_{\mathrm{DS}} < V_{\mathrm{k}}$), using the constant-mobility drain current Equation (2.75), we find

$$g_{\mathrm{ds}} = \frac{\epsilon_{\mathrm{SiO_2}}\, W_{\mathrm{G}}\, \mu'_{\mathrm{n}}}{L_{\mathrm{G}}\, t_{\mathrm{ox}}} \left[(V_{\mathrm{GS}} - V_{\mathrm{th}}) - V_{\mathrm{DS}} \right]. \tag{2.95}$$

In the saturated region ($V_{\mathrm{DS}} > V_{\mathrm{k}}$), we use the Sakurai–Newton model Equation (2.91) as our simplified physical regions would predict $g_{\mathrm{ds}} = 0$ there:

$$g_{\mathrm{ds}} = \lambda I_{\mathrm{D,sat}} = \lambda \frac{B\, W_{\mathrm{G}}}{L_{\mathrm{eff}}} (V_{\mathrm{GS}} - V_{\mathrm{th}})^n. \tag{2.96}$$

To reconcile the Sakurai–Newton model with the constant-mobility model, choose $B = (\epsilon_{\mathrm{SiO_2}}\, \mu'_{\mathrm{n}})/t_{\mathrm{ox}}$, $n = 2$, $L_{\mathrm{eff}} = L_{\mathrm{G}}$.

Transit frequency.

We again approximate the transit frequency with

$$f_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{C_{\mathrm{GS}} + C_{\mathrm{GD}}}.$$

In saturation and using the Meyer capacitance equations (2.87),

$$f_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{C_{\mathrm{GSO}} + C_{\mathrm{GDO}} + \frac{2}{3} C_{\mathrm{ox}}}.$$

While recognising the importance of the overlap capacitances, let us assume for simplification that C_{ox} dominates.

For the transconductance, we need to distinguish between the constant-mobility and constant-velocity models. For the constant mobility, using Equation (2.93), we obtain

$$f_{\mathrm{T}} = \frac{3\, \mu'_{\mathrm{n}}}{4\pi\, L_{\mathrm{G}}^2} (V_{\mathrm{GS}} - V_{\mathrm{th}}). \tag{2.97}$$

The transit frequency is predicted to increase linearly with the gate overtravel. However, this does not take the mobility degradation with increasing normal field into account (see Equation (2.81)).

In the constant-velocity limit, the transconductance is given by Equation (2.94) and the transit frequency becomes

$$f_T = \frac{3 v_{sat}}{4\pi \, L_G}. \tag{2.98}$$

We already recognised (page 105) that velocity saturation will dominate in short-channel MOSFETs, so we conclude that, like in MESFETs and HEMTs, f_T will scale inversely proportional to the gate length for short L_G.

For a given L_G, the transit frequency may still be increased by improving the mobility, because velocity saturation will be reached sooner and the average velocity in the channel increases.

Both electron and hole mobilities in silicon are enhanced if the silicon layer experiences a tensile strain in the plane parallel to the Si/SiO$_2$ interface. Semiconductor heterostructures can be used to achieve this: on top of the silicon wafer, first a strain-relaxed Si$_{1-x}$Ge$_x$ buffer is grown. As was discussed in Section 1.20, the addition of Ge lowers the band gap and at the same time increases the lattice constant. The latter effect is used here – if a thin Si layer is grown on top of the SiGe buffer, it experiences a tensile strain.

Figure 2.55 shows an example of this *strained-layer* technique, here combined with silicon-on-insulator [2]. The use of silicon as the channel layer makes this structure fully compatible with existing gate technology modules.

The axial tensile strain has a significant impact on the electron mobility, as is shown in Figure 2.56, at the expense of extra processing steps, and potential yield limitations when using non-lattice-matched materials.

For p-channel MOSFETs, it is advantageous to place the channel into SiGe layers with significant Ge mole fraction. This will not be discussed here further, however.

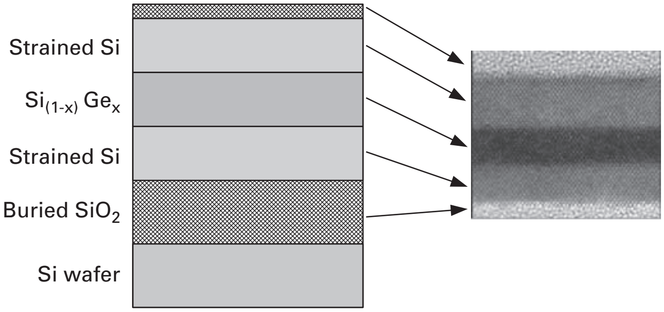


Fig. 2.55 Heterostructure-on-insulator layer stack on a strained-Si MOSFET (left), and corresponding TEM micrograph (TEM micrograph from D. A. Antoniadis, I. Aberg, C. NiCléirigh, O. M. Nayfeh, A. Khakifirooz and J. L. Hoyt, *IBM Journal of Research and Development*, Vol. 50, No. 4/5, pp. 363–377, April–May 2006. ©IBM).

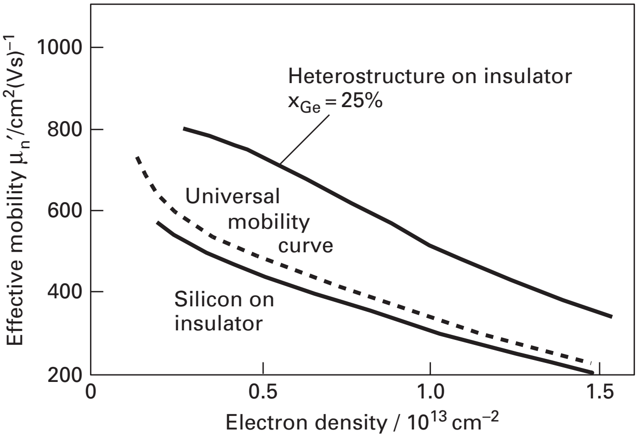


Fig. 2.56 Electron mobility enhancement in strained-Si layer (Data adapted from D. A. Antoniadis, I. Aberg, C. NiCléirigh, O. M. Nayfeh, A. Khakifirooz, J. L. Hoyt, *IBM Journal of Research and Development*, Vol. 50, No. 4/5, pp. 363–377, April–May 2006. ©IBM).

Maximum frequency of oscillation.

As already noted, the maximum frequency of oscillation $f_{\max}$ is the more meaningful figure of merit in analogue high-speed applications. As the equivalent circuit for the MOSFET is very similar to that used for MESFET and HEMT, we can easily adapt the $f_{\max}$ equation used there:

$$f_{\max} = \frac{f_T}{2\sqrt{g_{ds}(R_G + R_S) + 2\pi\, f_T\, R_G\, C_{GDO}}}, \tag{2.99}$$

because in saturation $C_{GD} = C_{GDO}$.

The combination of strained silicon channels and ultrashort gate lengths enables cut-off frequencies above 300 GHz for n-channel CMOS. In a 65 nm technology, a device with $L_G = 29$ nm was reported to have an $f_T = 360$ GHz and an $f_{\max} = 420$ GHz [45].

Microwave noise.

The treatment of noise in MOSFETs traditionally neglects the gate and source series resistances and considers only two noise sources [34]:

- The spectral noise current density generated in the channel:

$$\left\langle |i_d|^2 \right\rangle = 8\,kT\,\gamma\, g_{d0}, \tag{2.100}$$

where $g_{d0} = \delta I_D / V_{DS}$ at $V_{DS} = 0$ and γ is a parameter which varies from a value of 1 at $V_{DS} = 0$ to 2/3 at $V_{DS} = V_k$. This model is valid only in the linear region ($V_{DS} \leq V_k$), and was developed for MOSFETs with long gate lengths. In short-channel FETs and in saturation, the observed spectral noise density can be substantially higher. This can be accommodated by making the temperature T larger than the lattice temperature, to account for the significant kinetic energy of the free charge carriers.

- The induced spectral noise current density of the gate current:

$$\left\langle |i_{\mathrm{g}}|^2 \right\rangle = 8\,kT\,\delta\,g_{\mathrm{g}}, \tag{2.101}$$

where

$$g_{\mathrm{g}} = \frac{\omega^2\,C_{\mathrm{GS}}^2}{5\,g_{\mathrm{d0}}}.$$

In long channel FETs, $\delta = 4/3$.

As discussed for MESFET and HEMT, these two noise sources are partially correlated; in the long gatelength limit, the correlation coefficient is $c = j\,0.395$. From these noise sources, the minimum noise figure can be calculated to be

$$F_{\min} = 1 + \frac{2}{\sqrt{5}}\frac{f}{f_{\mathrm{T}}}\sqrt{\gamma\,\delta\,\left(1 - |c|^2\right)}. \tag{2.102}$$

Because f_{T} in velocity saturation scales $\sim 1/L_{\mathrm{G}}$, we expect the noise figure to vary linearly with the gate length.

An additional noise contribution can come from the substrate. The conducting substrate can be lumped together into a single value, the so-called *spreading resistance*, R_{sub}. This resistance naturally creates thermal noise, with a spectral noise current density:

$$\left\langle |i_{\mathrm{sub}}|^2 \right\rangle = \frac{8\,kT}{R_{\mathrm{sub}}}. \tag{2.103}$$

This noise current can be capacitively coupled into the transistor via the bulk node. It also leads to a voltage drop across C_{BS}, which will create an additional drain current fluctuation via the backgating effect (see Figure 2.54).

The issue of the source and gate series resistances needs to be re-examined. Modern MOS devices have poly-Si gates. Even highly n-doped poly-Si has specific resistivities which are much higher than for metal films. In RF CMOS technologies, the way around this problem is to connect many very short gate fingers in parallel, e.g. 40 gates of 5 µm gate width each, for a total W_{G} of 200 µm. As the gate length is more and more decreased, the gate resistance still needs to be recognised with, as Figure 2.57 [48] demonstrates. In this experiment, the noise figure is no longer decreased for $L_{\mathrm{G}} < 0.5$ µm, due to the increase in gate series resistance. Improved gate processes are therefore an important aspect in RF CMOS technology development.

As the gate oxide thickness t_{ox} decreases, the gate current due to Fowler–Nordheim tunnelling increases strongly. It generates a *shot noise contribution* [43], which will have to be accounted for in future MOSFET noise models. If I_{G} is the gate current, then the spectral noise current density generated is

$$\left\langle |i_{\mathrm{g}}|^2 \right\rangle = 4\,q\,I_{\mathrm{G}}. \tag{2.104}$$

This gate current leads to an additional term in the $F_{\min}$ expression [19], compare Equation (2.102):

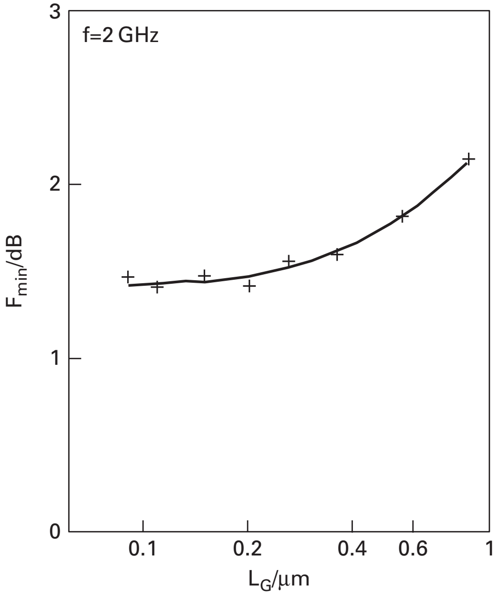


Fig. 2.57 Noise figure versus gate length for n-channel MOSFETs (40 fingers of $W_G = 5\,\mu$m each) (M. Saito, M. Ono, R. Fujimoto, H. Tanimoto, N. Ito, T. Yoshitomi, T. Ohguro, H. S. Momose and H. Iwai, *IEEE Transactions on Electron Devices*, Vol. ED-45, pp. 737–742, March 1998. ©1998 IEEE).

$$F_{\mathrm{min}} = 1 + \frac{f}{f_{\mathrm{T}}}\sqrt{\frac{\delta\gamma}{5}(1 - c_{\mathrm{G}}^2) + \frac{2q I_{\mathrm{G}} g_{\mathrm{d0}}\gamma}{16\pi^2 kT\, f^2\, C_{\mathrm{GS}}^2}}. \tag{2.105}$$

For low frequencies, the second term under the root dominates and the minimum noise figure becomes independent of frequency:

$$F_{\mathrm{min}} \approx 1 + \frac{1}{g_{\mathrm{m}}}\sqrt{\frac{2q I_{\mathrm{G}} g_{\mathrm{d0}}\gamma}{4kT}}. \tag{2.106}$$

The appearance of a frequency-independent component in $F_{\mathrm{min}}(f)$ is a tell-tale sign of gate-related shot noise.

A gate current due to Fowler–Nordheim tunnelling is expected to vary with the normal electric field across the gate oxide as

$$I_{\mathrm{G}} \sim \mathcal{E}_{\mathrm{y,SiO_2}}^2 \exp\left(-\frac{\phi^{\frac{3}{2}}}{\mathcal{E}_{\mathrm{y,SiO_2}}}\right), \tag{2.107}$$

where ϕ is the barrier at the interface. It will therefore be strong function of the gate overtravel $V_{\mathrm{GS}} - V_{\mathrm{th}}$. The occurrence of gate leakage has thus also important implications on the design of low-noise amplifiers using sub-100 nm CMOS technologies, namely in the choice of the bias point.

2.5 Bipolar and hetero-bipolar transistors

Despite the dominance of MOSFETs in digital circuits, and the significance of HFETs in micro- and millimetre-wave ICs, bipolar transistors have made a strong comeback since the 1990s in high-speed analogue electronics, which is particularly due to the arrival of the Si/SiGe heterostructure bipolar transistor (HBT), but also due to widespread use of GaAs-based HBTs in power amplifiers, e.g. of mobile phone handsets.

One important advantage of bipolar devices is that the current flow is vertical rather than lateral in FETs. This means that the critical geometric dimension (the base layer thickness) is defined by epitaxy or ion implantation. In FETs, the speed-limiting geometry is the gate length which, in present commercially available devices, is defined laterally by lithographic means, at a much higher cost.

We will approach the understanding of HBTs by first considering the standard homojunction homojunction bipolar transistor (BJT), which has long been a corner stone of high-speed electronics, but is gradually being replaced by Si/SiGe HBTs. In particular, we will get a grasp of the shortcomings of the homojunction BJT and how they can be solved by the introduction of bandgap engineering. The HBT is then a straightforward extension of the bipolar transistor concept.

2.5.1 Homojunction bipolar transistors

Homojunction bipolar transistors are the ‘classical’ bipolar transistors, where all parts of the device are fabricated from the same semiconductor material. Only silicon devices have any market relevance today; for the discussion of high-speed electronics, we can restrict our considerations to n–p–n type devices for reasons which will become clear shortly.

With a suitable permutation in indices, the discussion of n–p–n homojunction bipolar transistors is also valid for p–n–p devices, however.

Figure 2.58 shows the time-honoured[7] one-dimensional representation of a conceptual bipolar transistor of the n–p–n type: the emitter layer is highly donor-doped

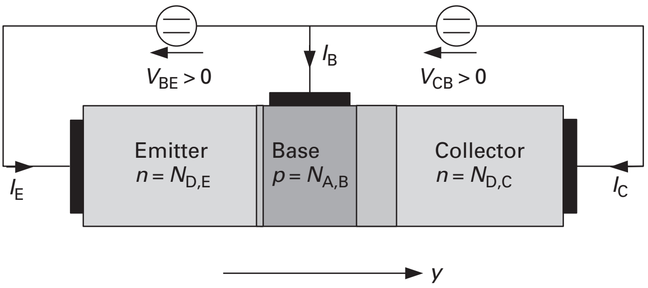


Fig. 2.58 Simplified schematic cross-section of a BJT.

[7] This schematic picture actually dates back to Figure 3 of Shockley’s US patent [56].

(n-type), followed by an acceptor-doped (p-type) base layer of medium doping density and a typically lower doped donor-doped (n-type) collector layer.

The emitter, base and collector contacts are non-blocking (ohmic) contacts and are assumed here to be of the recombination type. As shown, two external voltage sources V_{BE} and V_{CB} are connected to the device in such a way that the

- base–emitter p–n junction is forward-biased and
- base–collector p–n junction is reverse-biased.

This mode of operation is called *active forward operation*.

Diffusion triangle

To understand the way in which we control current in the bipolar transistor, let us concentrate initially on the base layer only. The first parameters to introduce are the *diffusion length* of minority charge carriers in the base; in this case the diffusion length of electrons (the base is p-type) L_{n}, and the thickness of the neutral base layer W_{B}. W_{B} is the thickness of the *neutral* base as we have to subtract the space charge regions first. We calculate the diffusion length from the low-field carrier mobility μ, the carrier lifetime τ_{r} and the absolute temperature T. In the n–p–n transistor, the minority charge in the base are electrons and hence we have to use the electron mobility μ_{n} and the electron lifetime in the base $\tau_{\mathrm{r,n}}$:

$$L_{\mathrm{n}} = \sqrt{\frac{kT}{q}\mu_{\mathrm{n}}\tau_{\mathrm{r,n}}}, \tag{2.108}$$

where k is Boltzmann's constant and q the elementary charge.
The term

$$D_{\mathrm{n}} = \frac{kT}{q}\,\mu_{\mathrm{n}} \tag{2.109}$$

is the *diffusion constant for electrons*, hence

$$L_{\mathrm{n}} = \sqrt{D_{\mathrm{n}}\tau_{\mathrm{r,n}}}.$$

Equation (2.109) is the Einstein equation introduced earlier, in Equation (1.79).

The emitter–base junction is forward-biased. Therefore, the minority (here, electron) concentration in the base immediately adjacent to the emitter–base space charge region (defined as $y = 0$) is elevated according to

$$n_{\mathrm{p}}(0) = \frac{n_{\mathrm{i,b}}^2}{N_{\mathrm{A,B}}} e^{qV_{\mathrm{BE}}/kT}, \tag{2.110}$$

where $n_{\mathrm{i,b}}$ is the intrinsic carrier density in the base.

Provided that the p-layer is infinitely extended in the y direction, the excess minority carrier density decays as

$$n_{\mathrm{p}}(y) = \frac{n_{\mathrm{i,b}}^2}{N_{\mathrm{A,B}}} + \left[n_{\mathrm{p}}(0) - \frac{n_{\mathrm{i}}^2}{N_{\mathrm{A,B}}}\right] e^{-y/L_{\mathrm{n}}}. \tag{2.111}$$

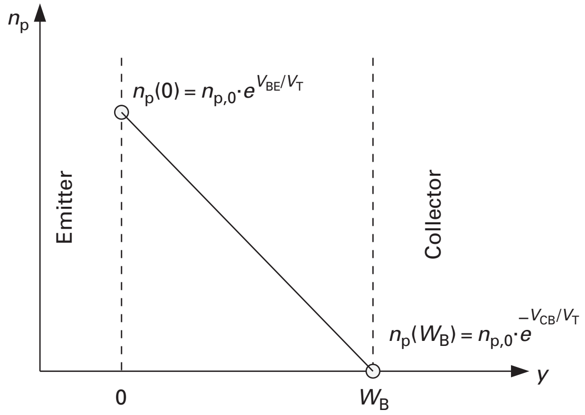

n_p
$n_p(0) = n_{p,0} \cdot e^{V_{BE}/V_T}$
Emitter
Collector
$n_p(W_B) = n_{p,0} \cdot e^{-V_{CB}/V_T}$
y
0
W_B

Fig. 2.59 Minority carrier concentration n_p in the base of an n–p–n bipolar transistor, as a function of coordinate y perpendicular to the surface.

In the bipolar transistor, however, the base width W_B is made much shorter than the diffusion length:[8]

$$W_B \ll L_n.$$

It is instructive to estimate a numeric value for L_n. Both μ_n and $\tau_{r,n}$ are strong functions of doping. Let us assume that for a reasonable base doping concentration, $\tau_{r,n} = 1\,\mu s$ and $\mu_n = 500\,cm^2(Vs)^{-1}$. Then the diffusion length amounts to $L_n = 36\,\mu m$, which is certainly much larger than the base width in any microwave bipolar transistor.

The reverse bias across the collector–base junction will cause the minority carrier at the collector side of the neutral base ($y = W_B$) to be significantly smaller than the minority carrier density in the undisturbed semiconductor:

$$n_p(W_B) = \frac{n_{i,b}^2}{N_{A,B}} e^{-qV_{CB}/kT}. \tag{2.112}$$

Provided that $W_B \ll L_n$, which is equivalent to neglecting recombination in the base, the distribution of minority carriers in the base as a function of the coordinate y (which is perpendicular to the surface of the device) is a linear function (see Figure 2.59). Due to its geometric shape, it is sometimes referred to as the *diffusion triangle*.

Collector current equation

In the classic bipolar transistor, the current is carried through the base layer by diffusion only, because the electric field in y direction in the neutral base can be neglected.

We formulate the electron current density flowing through the base layer as a diffusion current:

$$J_{n,B} = qD_n \frac{dn_p}{dy} = qD_n \frac{n_p(0) - n_p(W_B)}{W_B} \approx qD_n \frac{n_p(0)}{W_B}, \tag{2.113}$$

[8] The so-called *short-base diode condition*.

because $n_\mathrm{p}(0) \gg n_\mathrm{p}(W_\mathrm{B})$. D_n is the diffusion constant of electrons in the base, which has already been defined in Equation (2.109).

Note that due to $dn_\mathrm{p}(y)/dy = \mathrm{const}$, the current is constant throughout the base. We obtain the collector current by inserting Equation (2.110) into (2.113) and multiplying with the emitter area A_E:

$$I_\mathrm{C} = qA_\mathrm{E}D_\mathrm{n}\frac{n_\mathrm{ib}^2}{W_\mathrm{B}N_\mathrm{A,B}}e^{V_\mathrm{BE}/V_\mathrm{T}}, \tag{2.114}$$

still considering $n_\mathrm{p}(W_\mathrm{B}) \ll n_\mathrm{p}(0)$, i.e. under reverse collector–emitter bias. Without this condition, we obtain

$$I_\mathrm{C} = qA_\mathrm{E}D_\mathrm{n}\frac{n_\mathrm{ib}^2}{W_\mathrm{B}N_\mathrm{A,B}}\left(e^{V_\mathrm{BE}/V_\mathrm{T}} - e^{-V_\mathrm{CB}/V_\mathrm{T}}\right). \tag{2.115}$$

The expression in the denominator of Equation (2.115) $W_\mathrm{B}N_\mathrm{A,B}$ is the *Gummel number of the base layer*, G_B. In the above example and also below, we assume that the base doping concentration is constant across the neutral base. If this is not the case, i.e. $N_\mathrm{A,B} = f(y)$, we calculate the Gummel number as the integral sheet charge in the base:

$$G_\mathrm{B} = \int_0^{W_\mathrm{B}} N_\mathrm{A,B}(y)dy. \tag{2.116}$$

Ideal base current

We will now calculate the ideal base current of a bipolar transistor, i.e. the base current without components due to recombination.

For this, we consider the emitter layer as a short-base diode also: $W_\mathrm{E} \ll L_\mathrm{p}$, where W_E is the emitter width and L_p is the diffusion constant of minorities (here, holes) in the emitter:

$$L_\mathrm{p} = \sqrt{D_\mathrm{p}\tau_\mathrm{r,p}}, \tag{2.117}$$

where $\tau_\mathrm{r,p}$ is the carrier lifetime of holes in the emitter and

$$D_\mathrm{p} = \frac{kT}{q}\mu_\mathrm{p} \tag{2.118}$$

is the diffusion constant for holes. μ_p is the hole mobility.

Here, the minority carrier density is assumed to be zero at the emitter contact (ideal recombination contact). The result is a linear dependence of the minority carrier density p_n on y in the emitter (see Figure 2.60).

In analogy to Equation (2.113), we write the hole current in the emitter as a diffusion current:

$$J_\mathrm{pE} = qD_\mathrm{p}\frac{n_\mathrm{ie}^2}{N_\mathrm{D,E}W_\mathrm{E}}e^{V_\mathrm{BE}/V_\mathrm{T}}, \tag{2.119}$$

where n_ie is the intrinsic carrier density in the emitter. Multiplication with the emitter area A_E yields the base current:

$$I_\mathrm{B} = qD_\mathrm{p}A_\mathrm{E}\frac{n_\mathrm{ie}^2}{N_\mathrm{D,E}W_\mathrm{E}}e^{V_\mathrm{BE}/V_\mathrm{T}}. \tag{2.120}$$

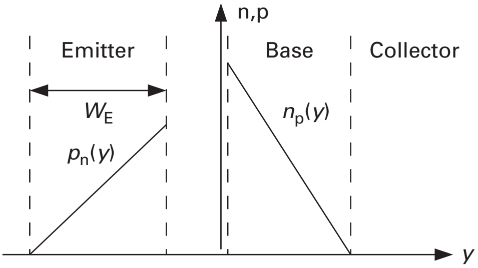


Fig. 2.60 Minority carrier concentration in a bipolar transistor, assuming short-base condition in emitter and base layer.

Ideal current gain

We can now calculate the ideal large-signal forward current gain of the bipolar transistor, dividing Equation (2.115) by (2.120).

$$B_F = \frac{I_C}{I_B} = \frac{D_n W_E}{D_p W_B} \frac{N_{D,E}}{N_{A,B}} \frac{n_{i,b}^2}{n_{i,e}^2}. \tag{2.121}$$

Let us consider the third term in Equation (2.121) first: in a homojunction transistor, where emitter and base are composed of the same material, we can assume $n_{i,b} \approx n_{i,e}$. This is not exactly true because the band gap and hence the intrinsic carrier density also depend weakly on the doping concentration, but it is a useful simplification.

However, we also note that provided we can fabricate the emitter from a different material, it should have a larger band gap (and correspondingly a smaller intrinsic carrier density) such that $n_{i,b} \gg n_{i,e}$. This *wide-gap emitter* is the fundamental idea behind the HBT which we will treat in the next section. It was already included in Shockley's original transistor patent [56] and theoretically expanded upon by Kroemer as early as 1957 [31].

If we consider the second term, we recognise that we cannot arbitrarily increase the base doping concentration unless we also increase the emitter doping concentration, without hurting the current gain. This will lead us to the fundamental limitation of the homojunction bipolar transistor – the inability to lower the base resistance sufficiently for excellent microwave operation.

Non-ideal current contributions

In certain bias conditions, we will have to include additional currents in our considerations. For this, it is instructive to view Figure 2.61.

Especially at low collector currents, *recombination currents* can frequently not be neglected: in bipolar transistors fabricated in direct bandgap semiconductors such as GaAs, the carrier lifetime may be so short that the short-base diode condition ($W_B \ll L_n$ in case of an n–p–n transistor) is never quite fulfilled, so that *volume recombination* in the neutral base has to be accounted for. In other devices, *surface recombination* near the emitter–base p–n junction may play a significant role. These effects are all lumped together in a current contribution J_R, which in an n–p–n transistor exists as an electron

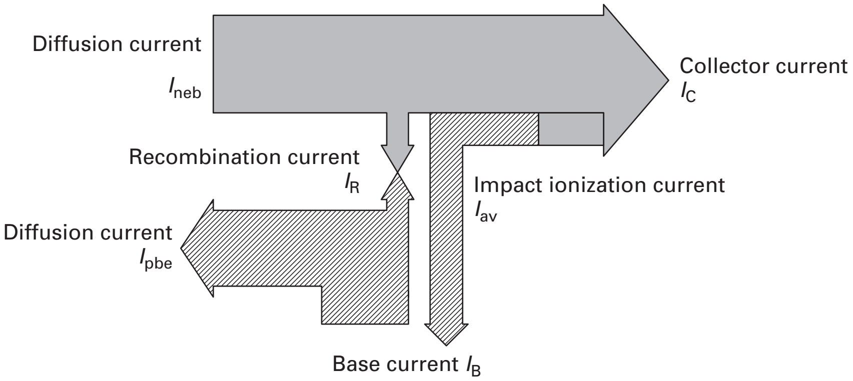


Fig. 2.61 Schematic representation of electron (solid shaded) and hole (hatched) currents in a bipolar transistor.

current (contributing to the emitter current) and a hole current (contributing to the base current).

To account for the recombination current, we add an additional term to the base current Equation (2.120). It has the form of a diode current term with its typical exponential voltage dependence:

$$I_B = \frac{I_C}{B_F} + I_{SR}(\exp^{q \cdot V_{BE}/(N_R \cdot kT)} - 1). \tag{2.122}$$

The emission coefficient N_R is larger than 1. $N_R = 2$ is a typical value for many recombination processes and I_{SR} is the saturation current of the non-ideal base current term.

For high V_{CB}, electrons in the collector space charge region may gain sufficient kinetic energy to elevate a valence electron into the conduction band when colliding with a lattice atom – *impact ionisation* occurs. The charge carriers created by the impact ionisation may again gain sufficient kinetic energy in the strong electric field to cause impact ionisation themselves, leading to a strong increase in current. This is called *avalanching* and is an important breakdown mechanism in bipolar transistors.

Avalanching is accounted for through an additional current term I_{av}, which exists as an electron current adding to the collector current, and a hole current, which *subtracts* from the base current. This is again shown in Figure 2.61.

Non-ideal current gain

The non-ideal current gain can now be written as

$$B = \frac{I_{\text{neb}} - I_R + I_{\text{av}}}{I_{\text{pbe}} + I_R - I_{\text{av}}}. \tag{2.123}$$

Even more insight is provided if we express Equation (2.123) in the form of a *common base current gain*.

$$A = \frac{-I_E}{I_C} = \frac{B-1}{B+1}$$

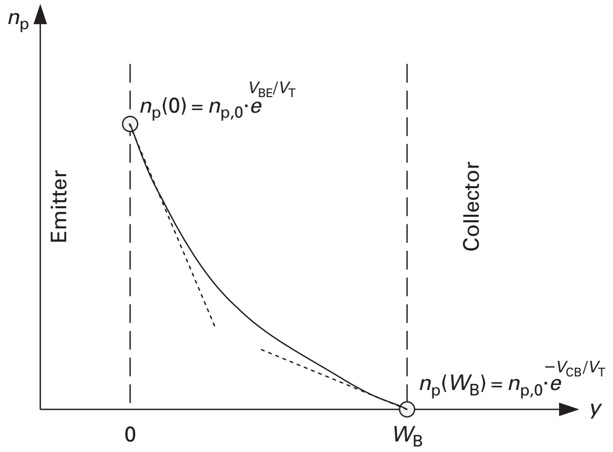


Fig. 2.62 Minority carrier distribution in the base layer in the presence of recombination.

$$A = \frac{I_{\mathrm{neb}}}{I_{\mathrm{neb}} + I_{\mathrm{pbe}}} \cdot \frac{I_{\mathrm{neb}} - I_{\mathrm{R}}}{I_{\mathrm{neb}}} \cdot \left(1 + \frac{I_{\mathrm{av}}}{I_{\mathrm{neb}} - I_{\mathrm{R}}}\right). \tag{2.124}$$

In Equation (2.124), the first product term is the *emitter efficiency* γ_{E}, which describes the electron current from the emitter to the base normalised to the overall current across the emitter–base junction. The second term is the *base transport factor* α_{T}, which describes the ratio of electron currents across the base–collector and emitter–base junctions. The last term is the *impact ionisation factor* α_{M}.

Equation (2.124) can hence be rewritten as

$$A = \gamma_{\mathrm{E}} \cdot \alpha_{\mathrm{T}} \cdot \alpha_{\mathrm{M}}.$$

Let us dwell on α_{T} for a moment. If recombination in the base cannot be neglected, then the minority carrier concentration in the base becomes

$$n_{\mathrm{p}}(y) = \frac{n_{\mathrm{iB}}^2}{N_{\mathrm{A,B}}} \left[\frac{\sinh\left(\frac{W_{\mathrm{B}}-y}{L_{\mathrm{n}}}\right)}{\sinh(\frac{W_{\mathrm{B}}}{L_{\mathrm{n}}})} \left(e^{qV_{\mathrm{BE}}/kT} - 1\right) + \frac{\sinh\ (\frac{y}{L_{\mathrm{n}}})}{\sinh(\frac{W_{\mathrm{B}}}{L_{\mathrm{n}}})} e^{-qV_{\mathrm{CB}}/kT} \right]. \tag{2.125}$$

This is depicted in Figure 2.62.

α_{T} can be interpreted as the ratio of the minority carrier gradients at $y = W_{\mathrm{B}}$ and $y = 0$:

$$\alpha_{\mathrm{T}} = \frac{\frac{dn_{\mathrm{p}}}{dy}(y = W_{\mathrm{B}})}{\frac{dn_{\mathrm{p}}}{dy}(y = 0)}. \tag{2.126}$$

In the presence of recombination, the base transport factor is therefore always less than one.

Saturation

In the above discussion, we assumed that the base–collector space charge region was reverse-biased – no carriers were injected into the base from the collector. We will

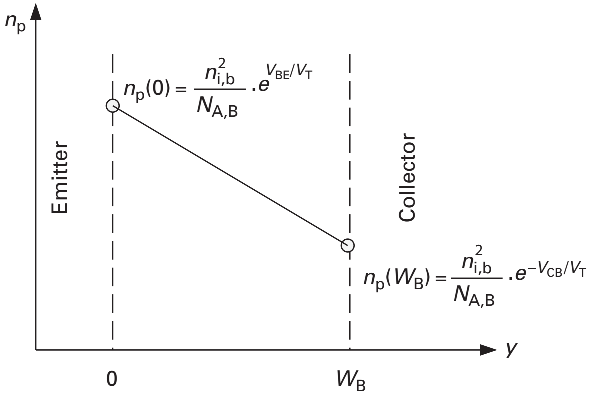

n_p
$n_p(0) = \frac{n_{i,b}^2}{N_{A,B}} . e^{V_{BE}/V_T}$
Emitter
Collector
$n_p(W_B) = \frac{n_{i,b}^2}{N_{A,B}} . e^{-V_{CB}/V_T}$
y
0
W_B

Fig. 2.63 Base minority concentration under saturation conditions.

now drop this condition and allow the base–collector junction to also become forward-biased. In the n–p–n transistor considered here, this means $V_{CB} < 0$. The corresponding bias condition is called *saturation*.

We can conveniently treat this condition again using the diffusion triangle concept introduced in Figure 2.59. Again, recombination across the base is neglected.

Because charge carriers are being injected into the base across the forward-biased collector–base junction, the minority carrier density in the neutral base adjacent to the collector–base space charge region is elevated. For the n–p–n transistor,

$$n_p(W_B) = \frac{n_{i,b}^2}{N_{A,B}} e^{-q \cdot V_{CB}/kT}. \tag{2.127}$$

The resulting diffusion triangle in the base is shown in Figure 2.63.

The ideal collector current under saturation conditions can once again be calculated using a pure diffusion current *ansatz* in the base:

$$I_C = q D_n \frac{dn_p(y)}{dy} = q D_n \frac{n_{i,b}^2}{N_{A,B} W_B} \left(e^{q \cdot V_{BE}/kT} - e^{-q \cdot V_{CB}/kT} \right). \tag{2.128}$$

For inhomogeneous base doping profiles, replace $W_B N_{A,B}$ by the integral Gummel number G_B according to Equation (2.116).

For a fixed base–emitter voltage, the collector current will now strongly decrease with decreasing collector–base voltage, while in the initial discussion of the active forward regime (see Equation 2.115), I_C did not depend on V_{CB}.

In high-speed circuits, the saturation regime has to be carefully avoided due to charge-storage effects whose detailed discussion is beyond the scope of this book.

Early effect

Upon closer examination, the collector current will show a dependence on V_{CB} even in the active forward regime. When deriving Equation (2.115) and the following equations, we had assumed that W_B was constant.

However, consider that W_B is defined as the width of the undepleted (neutral) region bordered by the emitter–base and the base–collector space charge regions. Because the collector current depends exponentially on the base–emitter voltage, $V_{BE} \approx$ const across a wide range of collector currents, and the width of the base–emitter space charge region can be considered constant.

As we change the voltage across the base–collector p–n junction, however, the width of its space charge region will be modulated, which leads to a variation in W_B.

If for simplicity we assume homogeneous doping profiles in base and collector ($N_{A,B}$ = const, $N_{D,C}$ = const), the extension of the base–collector space charge region into the base layer is

$$\delta y_B = \sqrt{2\frac{\epsilon_B}{q} \cdot \frac{N_{D,C}}{N_{A,B}} \cdot \frac{V_D + V_{CB}}{N_{A,B} + N_{D,C}}}, \tag{2.129}$$

where ϵ_B is the dielectric constant of the base layer material and V_D the built-in voltage of the base–collector p–n junction.

The diffusion triangle representation in Figure 2.64 may be helpful again. The dark-shaded areas denote the initial space charge regions. If V_{CB} is increased (with V_{BE} = const), the base–collector space charge region will expand as indicated by the light-shaded areas. Correspondingly, the minority carrier gradient in the neutral base will increase as the neutral base width shrinks.

Hence, the collector current will increase with increasing V_{CB}. This is the *Early effect* [15].

More quantitatively, revisit Equation (2.115), as we consider only $V_{CB} \gg kT/q$. However, now $W_B = f(V_{CB})$:

$$I_C = qA_E D_n \frac{n_{ib}^2}{W_B(V_{CB})N_{A,B}}. \tag{2.130}$$

We differentiate Equation (2.130) with respect to V_{CB}:

$$\frac{dI_C}{dV_{CB}} = -\frac{I_C}{W_B} \cdot \frac{dW_B}{dV_{CB}}. \tag{2.131}$$

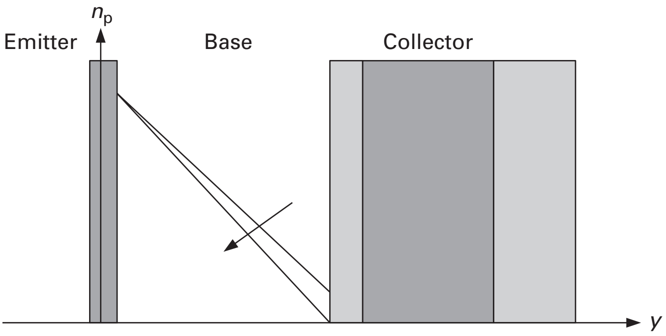


Fig. 2.64 Modification of the base diffusion triangle due to the modulation of the base–collector space charge region.

W_{B} can be written as $W'_{\mathrm{B}} - \delta y_{\mathrm{B}}$, where W'_{B} is the layer thickness of the base, i.e. without subtracting the space charge regions.[9] Hence for small changes in the neutral base width ($W'_{\mathrm{B}} \approx W_{\mathrm{B}}$),

$$\frac{dI_{\mathrm{C}}}{dV_{\mathrm{CB}}} = -\frac{I_{\mathrm{C}}}{W_{\mathrm{B}}} \cdot \frac{dW_{\mathrm{B}}}{dV_{\mathrm{CB}}} = \frac{I_{\mathrm{C}}}{W_{\mathrm{B}}} \frac{d\delta y_{\mathrm{B}}}{dV_{\mathrm{CB}}} = \frac{I_{\mathrm{C}}}{N_{\mathrm{A,B}} W_{\mathrm{B}}} \sqrt{\frac{\epsilon_{\mathrm{B}}}{2q} \frac{N_{\mathrm{D,C}} N_{\mathrm{A,B}}}{N_{\mathrm{D,C}} + N_{\mathrm{A,B}}} \frac{1}{V_{\mathrm{D}} + V_{\mathrm{CB}}}}. \tag{2.132}$$

With the Gummel number for homogeneous doping concentration in the base,

$$G_{\mathrm{B}} = N_{\mathrm{A,B}} W_{\mathrm{B}}$$

and the base–collector capacitance per unit area

$$C'_{\mathrm{j,BC}} = \sqrt{\frac{\epsilon_{\mathrm{B}} q}{2} \frac{N_{\mathrm{A,B}} N_{\mathrm{D,C}}}{N_{\mathrm{A,B}} + N_{\mathrm{D,C}}} \frac{1}{V_{\mathrm{D}} + V_{\mathrm{CB}}}}$$

it follows that

$$\frac{dI_{\mathrm{C}}}{dV_{\mathrm{CB}}} = \frac{I_{\mathrm{C}}}{V_{\mathrm{A}}}, \tag{2.133}$$

where V_{A} is the Early voltage:

$$V_{\mathrm{A}} = \frac{q \cdot G_{\mathrm{B}}}{C'_{\mathrm{jCB}}}. \tag{2.134}$$

The Early voltage is therefore directly proportional to the base Gummel number.

In microwave electronics, the Early voltage is an important factor because it affects the linearity of power amplifiers. For highly linear power amplifiers, a high Early voltage is desired, as it reduces the dependence of the collector current on the collector–emitter voltage. Due to the exponential dependence of the collector current on the base–emitter voltage, V_{BE} is approximately constant even for large variations of I_{C}, so that $\delta V_{\mathrm{CE}} \approx \delta V_{\mathrm{CB}}$. A high Early voltage V_{A} hence reduces the dependence of the collector current on the collector–emitter voltage.

Kirk effect

The last intrinsic effect to be discussed here is the Kirk effect. It can be once again explained using the diffusion triangle (see Figure 2.65).

Despite the fact that the minority carriers traversing the base are being injected into the base–collector space charge region, we had so far assumed that the space charge itself remains unaffected. This is true as long as the density of mobile charge is much smaller than the density of fixed charge.

If we increase the collector current sufficiently to create a density of mobile charge comparable to the fixed space charge density, the mobile charge will start to compensate the space charge and the space charge region will shrink.

[9] Well, actually we have to subtract the extension of the emitter–base space charge region into the base, but that can be neglected here.

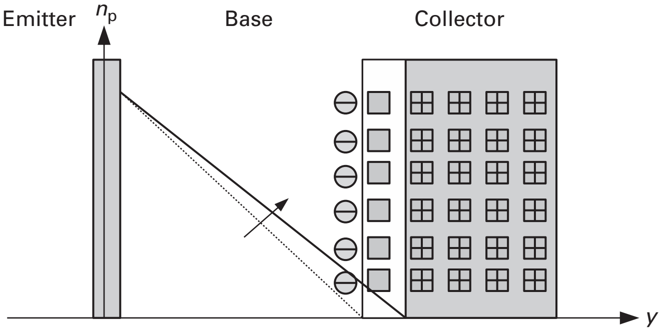


Fig. 2.65 Schematic representation of the Kirk effect influence on the diffusion triangle.

If the width of the space charge region decreases, the neutral base zone will expand – this is called *base push out*. Correspondingly, the minority carrier gradient and with it the collector current for a given V_{BE} will decrease.

To get an estimate of the collector current necessary to cause the Kirk effect, let us assume that the free charge (electrons in our case) will be accelerated to their drift saturation velocity v_{sat} immediately after they enter the base–collector space charge region. Then the electron density n_{C} corresponding to a collector current density J_{C} is

$$n_{\mathrm{C}} = \frac{J_{\mathrm{C}}}{q \cdot v_{\mathrm{sat}}}.$$

If we now require that the fixed ionised donors in the collector of density $N_{\mathrm{D,C}}$ shall be fully compensated by the mobile charge ($n_{\mathrm{C}} = N_{\mathrm{D,C}}$), we find for the critical current for the onset of the Kirk effect:

$$J_{\mathrm{C,Kirk}} = q \cdot N_{\mathrm{DC}} \cdot v_{\mathrm{sat}}. \tag{2.135}$$

The onset of Kirk effect hence scales proportionally with the collector doping concentration.

2.5.2 Small-signal dynamic behaviour

Next, we will discuss the dynamic behaviour of the bipolar transistor for the small-signal case, where the non-linear relationships between the terminal currents and the voltages can be described as linear relationships between the deviations of these entities from a given bias point, i.e.

$$i_{\mathrm{c}} = \delta I_{\mathrm{C}}, i_{\mathrm{b}} = \delta I_{\mathrm{B}}, v_{\mathrm{be}} = \delta V_{\mathrm{be}}, v_{\mathrm{ce}} = \delta V_{\mathrm{CE}}.$$

The forward-biased emitter–base junction will now be described by a conductance:

$$\frac{d(I_{\mathrm{C}} + I_{\mathrm{B}})}{dV_{\mathrm{BE}}} = -\frac{dI_{\mathrm{E}}}{dV_{\mathrm{BE}}} = g_{\mathrm{e}}, \tag{2.136}$$

where g_{e} will be referred to as the *dynamic emitter conductance*. The minus sign is due to the convention that all currents (I_{C}, I_{B}, I_{E}) are counted positive flowing into the transistor.

From Equations (2.115), (2.120) and (2.121), we conclude that

$$I_{\mathrm{E}} = -\left(I_{\mathrm{C}} + I_{\mathrm{B}}\right) = -q A_{\mathrm{E}} D_{\mathrm{n}} \frac{n_{\mathrm{ib}}^2}{G_{\mathrm{B}}} e^{V_{\mathrm{BE}}/V_{\mathrm{T}}} \left(1 + \frac{1}{BF}\right) \tag{2.137}$$

in the active forward regime, neglecting any non-ideal current contributions.

Differentiating Equation (2.137) with respect to V_{BE}, we find the dynamic emitter conductance to be

$$g_{\mathrm{e}} = -\frac{I_{\mathrm{E}}}{V_{\mathrm{T}}}. \tag{2.138}$$

It can be written as the dynamic emitter resistance:

$$r_{\mathrm{e}} = -\frac{V_{\mathrm{T}}}{I_{\mathrm{E}}}. \tag{2.139}$$

The modulation of stored charge in the neutral base results in the *diffusion capacitance*. In the short-base diode limit, for a homogeneously doped base layer, and if we neglect carrier injection from the collector into the base (i.e. for V_{CB} sufficiently high), the stored minority charge in the base can be easily calculated, refer again to Figure 2.59:

$$Q_{\mathrm{B}} = q A_{\mathrm{E}} \frac{W_{\mathrm{B}}}{2} \frac{n_{\mathrm{ib}}^2}{N_{\mathrm{A,B}}} e^{V_{\mathrm{BE}}/V_{\mathrm{T}}}. \tag{2.140}$$

Differentiating Q_{B} with respect to V_{BE} results in the diffusion capacitance C_{d}:

$$C_{\mathrm{d}} = \frac{dQ_{\mathrm{B}}}{dV_{\mathrm{BE}}} = q A_{\mathrm{E}} \frac{W_{\mathrm{B}}}{2} \frac{n_{\mathrm{ib}}^2}{N_{\mathrm{A,B}}} \frac{1}{V_{\mathrm{T}}} e^{V_{\mathrm{BE}}/V_{\mathrm{T}}}. \tag{2.141}$$

Considering Equation (2.137) and recalling that in the homogeneously doped case, the base Gummel number is $G_{\mathrm{B}} = N_{\mathrm{A,B}} W_{\mathrm{B}}$, we find that Equation (2.141) can be written as

$$C_{\mathrm{d}} = \frac{I_{\mathrm{E}}}{V_{\mathrm{T}}} \frac{W_{\mathrm{B}}^2}{2D_{\mathrm{n}}} = g_{\mathrm{e}} \frac{W_{\mathrm{B}}^2}{2D_{\mathrm{n}}}. \tag{2.142}$$

The ratio of diffusion capacitance to dynamic emitter conductance is bias-independent (at least in the active forward regime) and is called the *base transit time* τ_{B}:

$$\tau_{\mathrm{B}} = \frac{C_{\mathrm{d}}}{g_{\mathrm{e}}} = \frac{W_{\mathrm{B}}^2}{2D_{\mathrm{n}}}. \tag{2.143}$$

Recall that $D_{\mathrm{n}} = (kT/q)\mu_{\mathrm{n}}$ Equation (2.109); therefore, the base transit time is inversely proportional to the minority carrier mobility in the base. As the electron mobility μ_{n} is significantly larger than the hole mobility μ_{p}, in Si as well as the most common compound semiconductors, this justifies the restriction of our discussions to n–p–n-type transistors.

The *output conductance* is related to the Early effect. We find it by differentiating I_{C} with respect to V_{CE}:

$$g_{\mathrm{ce}} = \frac{dI_{\mathrm{C}}}{dV_{\mathrm{CE}}} = \frac{dI_{\mathrm{C}}}{dV_{\mathrm{CB}}}, \tag{2.144}$$

for constant V_{BE}. Therefore,

$$g_{\mathrm{ce}} = \frac{I_{\mathrm{C}}}{V_{\mathrm{A}}}. \tag{2.145}$$

The emitter–base and base–collector p–n junctions also present *junction capacitances* which we have to take into account. The calculation of the junction capacitance can be reduced to the problem of calculating the width of the depletion layer w, because the capacitance is that of a parallel plate capacitor with area A_{J} and plate separation w:

$$C_{\mathrm{J}} = \epsilon_{\mathrm{S}} \frac{A_{\mathrm{J}}}{w}. \tag{2.146}$$

The depletion layer width is calculated from Poisson's equation:

$$\frac{d^2\Phi}{dy^2} = -\frac{\rho(y)}{y}, \tag{2.147}$$

where Φ is the potential across the junction. A simple analytic solution can be found assuming that within the space charge region the mobile charge can be neglected compared to the fixed charge (i.e. the ionised donor density $N_{\mathrm{D}}^{+}(y)$ on the n side and the ionised acceptor density $N_{\mathrm{A}}^{-}(y)$ on the p side), and that within the depletion layer all doping atoms are ionised, while outside all doping atoms are neutral. The total potential difference across the depletion region must be equal to the sum of built-in (or diffusion) voltage V_{D} and the externally applied voltage V_{ext}.

For the simple case of homogeneous doping on both p and n sides ($N_{\mathrm{D}}(y) =$ const, $N_{\mathrm{A}}(y) =$ const), we obtain

$$w = \sqrt{2\frac{\epsilon_{\mathrm{S}}}{q}\frac{N_{\mathrm{A}} + N_{\mathrm{D}}}{N_{\mathrm{A}}N_{\mathrm{D}}}(V_{\mathrm{D}} + V_{\mathrm{ext}})}. \tag{2.148}$$

Note that V_{ext} is defined as a reverse (depleting) voltage here.

In active forward operation, the emitter–base diode is forward-biased ($V_{\mathrm{ext}} = -V_{\mathrm{BE}}$), while the base–collector diode is reverse-biased ($V_{\mathrm{ext}} = V_{\mathrm{CB}}$). In practical transistors, the emitter–base junction area A_{E} will also be different from the base–collector junction area A_{C}, so that we obtain for the n–p–n transistor:

$$C_{\mathrm{BE}} = A_{\mathrm{E}}\sqrt{\frac{q \cdot \epsilon_{\mathrm{S}}}{2}\frac{N_{\mathrm{A,B}}N_{\mathrm{D,B}}}{N_{\mathrm{A,B}} + N_{\mathrm{D,E}}}\frac{1}{V_{\mathrm{D}} - V_{\mathrm{BE}}}} \quad \text{for } V_{\mathrm{BE}} < V_{\mathrm{D}} \tag{2.149}$$

$$C_{\mathrm{CB}} = A_{\mathrm{C}}\sqrt{\frac{q \cdot \epsilon_{\mathrm{S}}}{2}\frac{N_{\mathrm{A,B}}N_{\mathrm{D,C}}}{N_{\mathrm{A,B}} + N_{\mathrm{D,C}}}\frac{1}{V_{\mathrm{D}} + V_{\mathrm{CB}}}}. \tag{2.150}$$

Finally, we have to account for the time lag associated with the transit of free charge carriers through the base–collector space charge region – the *collector transit time*.[10]

The calculation of the collector transit time is not as straightforward as it may seem, as it in principle needs both diffusive (at the base side edge of the space charge region)

[10] Because the emitter–base diode is forward-biased, its depletion region is very thin and the associated transit time can be safely neglected.

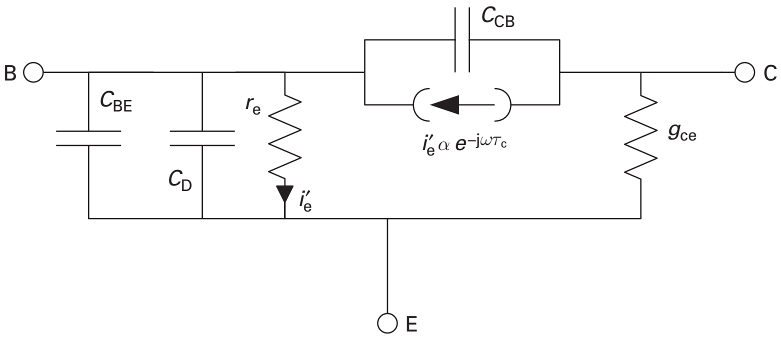


Fig. 2.66 Small-signal equivalent circuit of the intrinsic transistor.

and drift transport components. The field-dependent velocity also needs to be taken into account, as well as the displacement current created by charge moving within the space charge region.

A common assumption is that the carriers reach their drift saturation v_{sat} instantaneously after entering the space charge region. Then, the collector transit time τ_C is

$$\tau_C = \frac{w_C}{2 \cdot v_{sat}}. \tag{2.151}$$

The small-signal components of the intrinsic transistors, which we considered above, can be combined in the *small-signal equivalent circuit* shown in Figure 2.66. α is the small-signal common base current gain in the quasi-static limit.

Figure 2.66 applies only to the intrinsic transistor and will have to be extended by extrinsic parasitic components at microwave frequencies. Most importantly, we have to account for the series resistances. In order to appreciate the problem, please refer to Figure 2.67, which presents a more realistic cross-section of the bipolar transistor, compared to Figure 2.58.

The *emitter resistance* R_E (not to be confused with the dynamic emitter resistance r_e) is composed of the emitter contact resistance and the vertical resistance of the emitter layer, which in homojunction transistors typically is a poly-Si plug. The *collector resistance* R_C is formed by the collector contact resistance, the vertical resistance of the collector 'sinker' implant and the lateral resistance of the subcollector layer. The *base resistance* R_B finally is formed by the base contact resistance, the lateral resistance of the extrinsic base and the lateral resistance of the intrinsic base layer. Of these individual series resistance contributions, the lateral resistance of the intrinsic base layer is the most problematic, as the thickness of this layer is given by the neutral base width W_B and has to be very thin to minimise the base transit time, (see Equation (2.143)).

The series resistances have been added to the small-signal equivalent circuit in Figure 2.68. From an application point of view, R_B and R_E are the most significant.

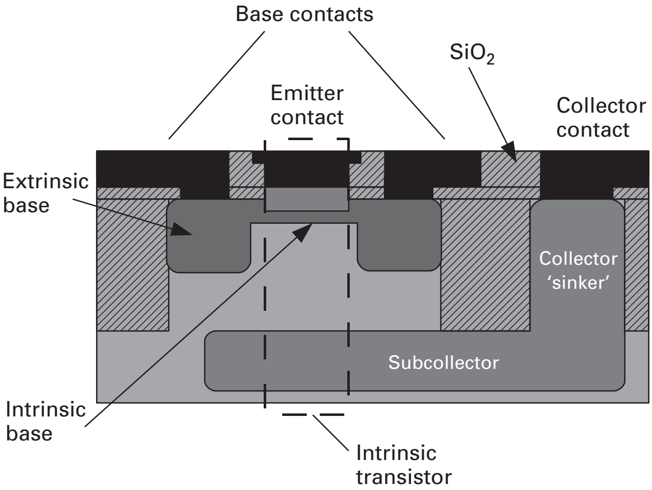


Fig. 2.67 A more realistic schematic cross-section of a typical bipolar transistor with planar contact arrangements. The dashed box indicates the intrinsic transistor – compare with Figure 2.58.

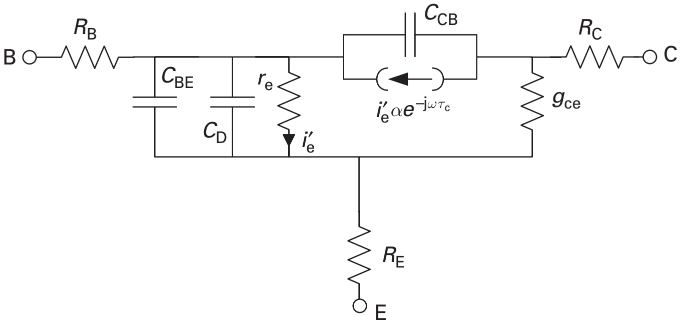


Fig. 2.68 Small-signal equivalent circuit of the bipolar transistor, including the series resistances.

Transit frequency.

The total transit time through the bipolar transistor is calculated from the *transit frequency* f_T, which is the frequency where the magnitude of the short-circuit current gain ($h_{21} = i_c/i_b$ for $v_{ce} = 0$) becomes one: $\tau_T = 1/(2\pi f_T)$.

$$\tau_T = \tau_B + \tau_C + r_e\,(C_{BE} + C_{BC}) + (R_E + R_C)\,(C_{BE} + C_{BC})\,. \tag{2.152}$$

The total transit time can be separated as follows:

- τ_B and τ_C are intrinsic time constants which do not depend on the emitter current (neglecting the Kirk effect).
- $r_e\,(C_{BE} + C_{BC})$ is the intrinsic emitter charging time which is inversely proportional to the emitter current (see Equation (2.139)).
- $(R_E + R_C)\,(C_{BE} + C_{BC})$ is the parasitic charging time due to the emitter and collector series resistances, which is frequently neglected.

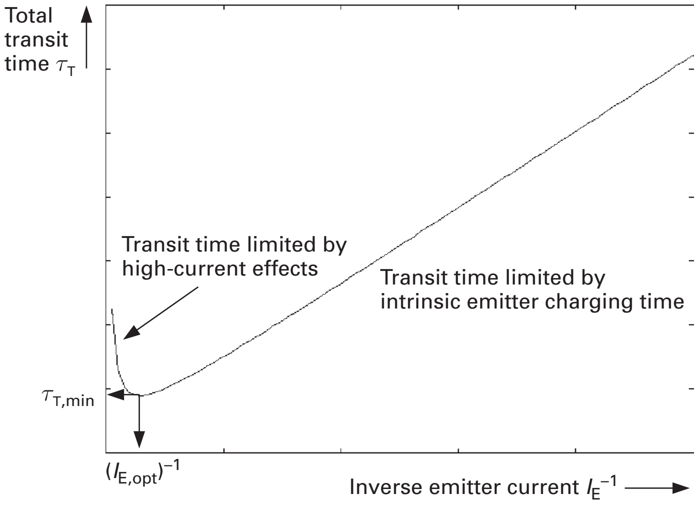


Fig. 2.69 Schematic representation of the total transit time in a bipolar transistor as a function of the inverse emitter current.

The effect of the intrinsic emitter charging time, which depends linearly on the emitter current (or, as $\alpha \approx 1$ in technical transistors, in good approximation on the collector current), leads to a strong bias dependence of the total transit time (and hence f_T), which is shown in Figure 2.69. For $I_E > I_{E,opt}$, high-current effects such as the Kirk effect will again prolong the transit time.

One of the key issues in designing high-speed bipolar circuits is therefore to choose the emitter current as close as possible to the optimum emitter current.

Note that neither the base resistance nor the output conductance have an influence on the total transit time – this is an effect of the definition via f_T and hence h_{21}. The definition of h_{21} assumes an ideal current source at the input and a short circuit at the output. g_{ce} has no effect as it is short-circuited (neglecting R_C here), and R_B is in series with an ideal current source and hence also has no effect.

Maximum frequency of oscillation.

As already discussed, the maximum frequency of oscillation f_{max} is a measure of the power gain cutoff frequency (f_T measures only the current gain behaviour): the frequency where the MAG of a two-port becomes one.

A common approximation of f_{max} for the bipolar transistor is

$$f_{max} = \sqrt{\frac{f_T}{8\pi R_B C_{BC}}}. \tag{2.153}$$

This equation is equivalent to the one introduced for FETs; see Equation (2.27) for very low output conductances and replacing $R_G \to R_B$, $C_{GD} \to C_{BC}$. For an in-depth treatment, see M. B. Das [11]. As explained there, the simplification neglects the distributed nature of the base resistance (as we did in this introductory text) and is only valid if the emitter series resistance R_E and output conductance g_{ce} are sufficiently small.

In practice, however, Equation (2.153) is useful even for today's HBTs with several hundred GHz $f_{\max}$.

Note that now R_B has a strong influence on the maximum frequency of oscillation.

2.5.3 Microwave noise performance of bipolar transistors

We will investigate the microwave noise performance of bipolar transistors using a simplified noise equivalent circuit (see Figure 2.70).

The series resistance R_E and R_C will be neglected, as will be the output conductance g_{ce} and the base–collector capacitance C_{BC}. This leaves three different noise sources to be included:

(i) the thermal noise associated with the base resistance R_B: $\langle |v_{nb}|^2 \rangle$;
(ii) the shot noise associated with the emitter–base p–n junction: $\langle |v_{ne}|^2 \rangle$;
(iii) the shot noise associated with the base–collector p–n junction: $\langle |i_{nc}|^2 \rangle$.

Due to the short base transit time, the emitter–base and base–collector shot noise sources are strongly correlated.

The rationale for the omission of the collector resistance is that its contribution would be divided by the gain of the transistor; further the value of R_C is typically much smaller than R_B. The thermal noise source of the emitter resistance with a squared spectral voltage density of $8\,kT\ R_E$[11] would be in series with the shot noise source of the emitter current, whose spectral voltage density is

$$\left\langle |v_{ne}|^2 \right\rangle = 4\,q\ I_E\,r_e^2 = 4\,kT\,r_e, \tag{2.154}$$

as $r_e = kT/(q\ I_E)$. As long as $r_e \gg 2R_E$, the thermal noise contribution of the emitter resistance can be neglected. Because low-noise bias points for bipolar transistors occur at small I_E, this can generally be assumed.

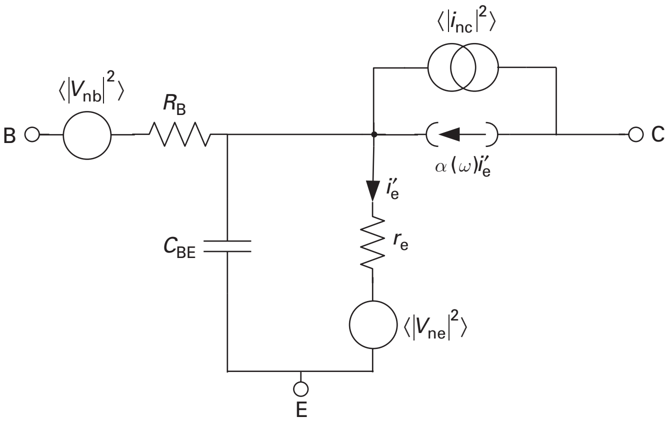


Fig. 2.70 Strongly simplified T-type equivalent noise circuit of a bipolar transistor.

[11] Magnitude of the complex phasor.

Using the equivalent circuit in Figure 2.70, Hawkins [23] derived for the bipolar transistor:

$$F_{\mathrm{min}} = a\,\frac{R_{\mathrm{B}} + R_{\mathrm{opt}}}{r_{\mathrm{e}}} + \frac{\alpha_0}{|\alpha(\omega)|^2}, \tag{2.155}$$

where $\alpha(\omega)$ is the frequency-dependent common base current gain and α_0 its quasi-stationary value:

$$\alpha(\omega) = \frac{\alpha_0}{1 + j\omega/\omega_{\mathrm{b}}}.$$

with the base cutoff frequency $\omega_{\mathrm{b}} = \tau_{\mathrm{B}}^{-1}$ and τ_{B} the base transit time Equation (2.143).

The parameter a is

$$a = \frac{1}{\alpha_0}\left[1 + \left(\frac{\omega}{\omega_{\mathrm{e}}}\right)^2\right]\left[1 + \left(\frac{\omega}{\omega_{\mathrm{b}}}\right)^2\right] - 1.$$

The cutoff frequency ω_{e} represents the emitter charging time:

$$\omega_{\mathrm{e}} = \frac{1}{C_{\mathrm{BE}} r_{\mathrm{e}}} = \frac{q\, I_{\mathrm{E}}}{kT\, C_{\mathrm{BE}}}.$$

R_{opt} is the real part of the noise-optimum generator impedance:

$$R_{\mathrm{opt}} = \sqrt{R_{\mathrm{B}}^2 - X_{\mathrm{opt}}^2 + \frac{\alpha_0}{|\alpha(\omega)|^2}\,\frac{r_{\mathrm{e}}(2R_{\mathrm{B}} + r_{\mathrm{e}})}{a}},$$

and X_{opt} the imaginary part:

$$X_{\mathrm{opt}} = \omega \frac{\alpha_0}{|\alpha(\omega)|^2}\,\frac{C_{\mathrm{BE}}\, r_{\mathrm{e}}^2}{a}.$$

It is instructive to consider the quasi-static case, $\omega \to 0$. In this case,

$$a = \frac{1 - \alpha_0}{\alpha_0}$$
$$X_{\mathrm{opt}} = 0$$
$$R_{\mathrm{opt}} = \sqrt{R_{\mathrm{B}}^2 + \frac{r_{\mathrm{e}}\,(2R_{\mathrm{B}} + r_{\mathrm{e}})}{1 - \alpha_0}}.$$

We obtain therefore

$$F_{\mathrm{min}}(\omega \to 0) = \frac{1}{\alpha_0} + \frac{R_{\mathrm{B}}}{\beta_0 r_{\mathrm{e}}} + \sqrt{\frac{R_{\mathrm{B}}^2}{(\beta_0\, r_{\mathrm{e}})^2} + (1 - \alpha_0)\frac{2\,R_{\mathrm{B}} + r_{\mathrm{e}}}{r_{\mathrm{e}}}}, \tag{2.156}$$

where $\beta_0 = \alpha_0/(1 - \alpha_0)$ is the common-emitter small-signal current gain.

We conclude that for $\omega \to 0$, the minimum noise figure does not converge towards 1, as in FETs, (see e.g. Equation (2.29) for the MESFET), but a higher value which depends on the current gain and the base resistance. If we further assume small $R_B/(\beta_0\, r_e)$, Equation (2.156) reduces to

$$F_{\min}(\omega \to 0) \approx \frac{1}{\alpha_0} + \sqrt{(1-\alpha_0)\frac{2R_B + r_e}{r_e}}.$$

It is obvious that the current gain has a very important influence on the noise performance of a bipolar transistor.

Let us now investigate a medium frequency range $\omega_e \ll \omega \ll \omega_b$. To simplify matters, we assume an ideal current gain $\alpha_0 = 1$. In this case,

$$\begin{aligned} a &= \left(\frac{\omega}{\omega_e}\right)^2 \\ X_{\text{opt}} &= \frac{\omega_e}{\omega} r_e \\ R_{\text{opt}} &= \sqrt{R_B^2 + 2\, R_B\, r_e \left(\frac{\omega_e}{\omega}\right)^2}, \end{aligned}$$

and finally for $F_{\min}(\alpha_0 = 1, \omega_e \ll \omega \ll \omega_b)$:

$$F_{\min} = 1 + \frac{\omega^2}{\omega_e^2}\frac{R_B}{r_e} + \frac{\omega}{\omega_e}\sqrt{\frac{R_B^2}{r_e^2}\frac{\omega^2}{\omega_e^2} + 2\frac{R_B}{r_e}}. \qquad (2.157)$$

We find that in this case the increase with frequency is determined by the base resistance R_B, which is therefore a very important parameter for the microwave noise behaviour of bipolar transistors.

Equation (2.155) contains an implicit bias dependence via $r_e = V_T/I_E$, where $V_T = kT/q$, as usual. r_e also determines f_e. As long as $I_E \gg V_T\, \omega\, C_{BE}$, $F_{\min}$ will increase proportionally with increasing I_E. For very small I_E, however $F_{\min}$ will increase inversely proportional to I_E. We note that here will be an optimum emitter current with respect to the noise performance. This current is usually much lower than the current required for optimum f_T (Figure 2.69) and hence requires a trade-off between device speed and noise in circuit design. Figure 2.71 shows an example calculation. We note that the optimum emitter current is frequency-dependent and moves to higher currents with increasing frequency.

An important noise parameter not considered in Hawkins' theory is the equivalent noise resistance. It determines the sensitivity of the noise figure on deviations from the noise-optimum generator impedance. Therefore, a small R_n facilitates circuit design as it makes exact noise match less critical (Section 5.3). Using Hawkins' equivalent circuit, an expression for R_n was introduced by Pucel and Rohde:

$$\begin{aligned} R_n = R_B\left(D - \frac{1}{\beta_0}\right) + \frac{r_e}{2}\Bigg\{ D + \left(\frac{R_B}{r_e}\right)^2 \cdot \Bigg[1 - \alpha_0 + \left(\frac{f}{f_b}\right)^2 \\ + \left(\frac{f}{f_e}\right)^2 + \left(\frac{1}{\beta_0} - \left(\frac{f}{f_b}\right)\left(\frac{f}{f_b}\right)\right)^2 \Bigg]\Bigg\}. \qquad (2.158) \end{aligned}$$

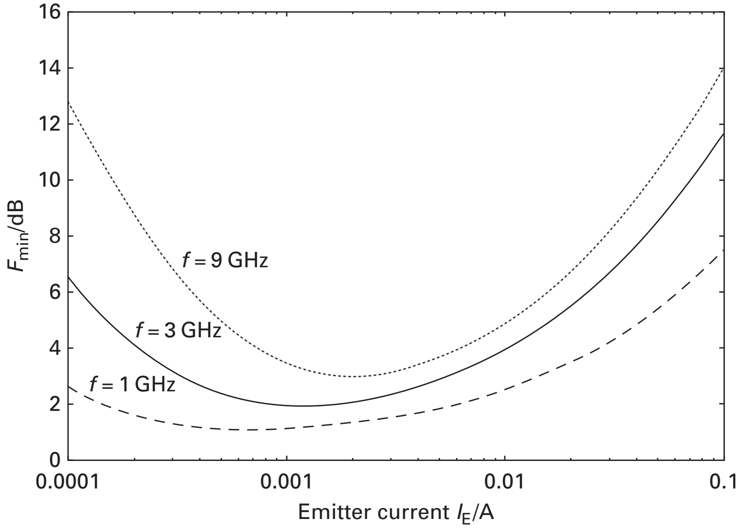


Fig. 2.71 Example calculation of bipolar noise figure dependence on the emitter current. Parameters chosen are: $f_b = 50\,\mathrm{GHz}$, $C_{\mathrm{BE}} = 0.2\,\mathrm{pF}$, $\alpha_0 = 0.99$ and $R_B = 10\,\Omega$.

The newly introduced parameter D is

$$D = \frac{1}{\alpha_0}\left[1 + \left(\frac{f}{f_b}\right)^2\right].$$

We note the importance of a low base resistance to achieve a small R_n.

2.5.4 Transit time optimisation

Drift field in the base

We had explicitly assumed that charge carriers traverse the base by diffusion only, that any electric field in the neutral base can be neglected. This is reasonably true provided that the base layer is highly and uniformly doped.

Any significant variation in doping concentration will lead to a built-in electric field which will either enhance or impede the carrier transport in the base. Advantageously, we make the base doping concentration higher at the base–emitter junction than at the base–collector junction, introducing an accelerating field for charge carriers travelling from emitter to collector.

Figure 2.72 shows the schematic band diagram of such a structure. The conduction and valence bands in the neutral base are titled due to the doping variation, adding a drift field force acting upon both electrons and holes. A constant electric field results if the doping concentration is exponentially varied:

$$N_{\mathrm{A,B}}(y) \sim e^{-\mathrm{a}\,\mathrm{y}}.$$

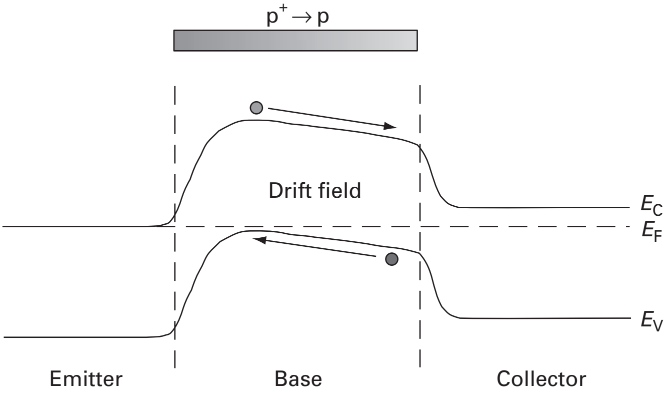


Fig. 2.72 Band diagram of a bipolar transistor with variation of the doping concentration in the base, creating a drift field.

If the base doping is adjusted using ion implantation from the emitter side, as is commonly the case in today's bipolar technologies, a suitable doping profile automatically results.

The built-in field can be easily calculated provided that the Boltzmann approximation is assumed to be valid:

$$\mathcal{E}_{\mathrm{y,bi}} = \frac{kT}{2\,q\,W_{\mathrm{B}}} \ln \frac{N_{\mathrm{A,B,max}}}{N_{\mathrm{A,B,min}}}.$$

The base transit time under the influence of this built-in field is then [65]:

$$\tau_{\mathrm{B}} = \frac{W_{\mathrm{B}}^2}{2\left[1 + \left(q\,\frac{\mathcal{E}_{\mathrm{y,bi}}\,W_{\mathrm{B}}}{kT}\right)^{3/2}\right] D_{\mathrm{n}}}. \tag{2.159}$$

Even modest variations of the base doping concentrations can result in substantial reductions in base transit time.

Collector transit time optimisation

The collector transit time Equation (2.151) can become a significant part of the total transit time, especially in devices with high breakdown voltages. Optimising the collector design involves important design compromises:

- For high f_{T}, the device needs to be driven to high collector currents, minimising the emitter charging time constant. Therefore, the Kirk effect needs to be pushed to higher currents, demanding a larger collector doping concentration. Equally, the collector transit time needs to be reduced by reducing the depleted collector width W_{C}. For a given collector–base voltage, this agrees with the demanded increase in collector doping concentration.
- However, a high maximum frequency of oscillation needs a low C_{BC} which, for a given collector–base voltage, demands a decrease in the collector doping concentration.

- Equally, an increase in collector–base breakdown voltage needs a decrease in collector doping concentration and a larger W_{C}.

The link between breakdown voltage and transit frequency is frequently expressed in terms of the *Johnson limit* [17]. To derive the Johnson limit, let us assume that the collector transit time fully dominates the total transit time. Using Equation (2.151), we find

$$f_{\mathrm{T}} \approx \frac{1}{2\pi\ \tau_{\mathrm{C}}} = \frac{v_{\mathrm{sat}}}{\pi\ W_{\mathrm{C}}} \rightarrow W_{\mathrm{C}} = \frac{v_{\mathrm{sat}}}{\pi\ f_{\mathrm{T}}}.$$

If $\mathcal{E}_{\mathrm{crit}}$ is the critical field for breakdown and if we assume homogeneous doping in the collector, the collector–base breakdown voltage (open emitter terminal) is

$$BV_{\mathrm{CBO}} = \mathcal{E}_{\mathrm{crit}} \frac{W_{\mathrm{C}}}{2}.$$

The product of transit frequency and breakdown voltage will then only depend on the material properties V_{sat} and BV_{CBO}:

$$f_{\mathrm{T}} \cdot BV_{\mathrm{CBO}} = \frac{\mathcal{E}_{\mathrm{crit}} v_{\mathrm{sat}}}{2\pi}.$$

The collector–emitter breakdown voltage BV_{CEO} is lower than BV_{CBO} because the impact ionisation current is amplified by the current gain B when entering the base:

$$BV_{\mathrm{CEO}} = \frac{BV_{\mathrm{CBO}}}{\sqrt[m]{B}}, \tag{2.160}$$

where m is a parameter which depends on the exact geometry and doping of the transistor.

We find for the Johnson limit:

$$f_{\mathrm{T}} \cdot BV_{\mathrm{CEO}} = \frac{\mathcal{E}_{\mathrm{crit}} v_{\mathrm{sat}}}{2\pi\ \sqrt[m]{B}}. \tag{2.161}$$

For Si, $\mathcal{E}_{\mathrm{crit}} \approx 5 \cdot 10^5\,\mathrm{V\,cm^{-1}}$. The drift saturation velocity at room temperature is $v_{\mathrm{sat}} \approx 10^7\,\mathrm{cm\,s^{-1}}$. Assuming a typical $B = 250$ and $m = 4$, we find $f_{\mathrm{T}} \cdot BV_{\mathrm{CEO}} = 200\,\mathrm{GHz}$. This is the frequently quoted ‘Johnson Limit’ for silicon bipolar devices. We readily recognise from Equation (2.161) that it is not a constant and can be significantly different for other values of B and m.

2.5.5 Heterojunction bipolar transistors

The base design dilemma

In our discussion of homojunction bipolar transistors, three main parameters with crucial impact on the high frequency performance were identified:

(i) the base transit time which sets the ‘intrinsic speed’ of the transistors;
(ii) the base resistance which affects the maximum frequency of oscillation and the noise performance;
(iii) the current gain which not only influences the noise performance, but also has to be typically $\geq$100 to simplify circuit design.

The dilemma is that the base design parameters W_B, $N_{A,B}$ influence these parameters in different ways:

- $\tau_B \sim W_B^2$ and increases, albeit weakly, with increasing N_A due to the reduction in the minority carrier mobility;
- $R_B \sim \frac{1}{W_B N_{A,B}}$;
- $\beta \sim \frac{1}{W_B} \frac{N_{D,E}}{N_{A,B}}$.

A transistor with high f_T and f_{max} would therefore have a thin, highly doped base layer. However, the high base doping concentration will decrease the current gain if the emitter doping concentration cannot be proportionally increased.

On the other hand, there are limits to the increase in emitter doping. The main limitation is *bandgap narrowing*. With increasing doping concentration, the band gap in the emitter will decrease. In silicon,

$$\Delta E_G \approx 22.5 \left(\frac{N_D}{10^{18}\,\mathrm{cm}^{-3}} \frac{300\,\mathrm{K}}{T} \right)^{0.5} . \tag{2.162}$$

The decrease in band gap in the emitter will increase the intrinsic carrier concentration there ($n_i{\sim}e^{-E_G/2kT}$), which in turn lowers the current gain because the base current due to injection of holes from the base into the emitter is

$$J_B \sim \frac{n_{i,E}^2}{N_{D,E}}.$$

We therefore conclude that the base doping concentration cannot be be strongly increased while keeping a high current gain. Therefore, thin-base microwave bipolar transistors have a problem with rather high base resistances.

Two approaches can be taken to solve the base design dilemma:

(i) We can increase the carrier velocity in the base so that a target transit frequency can be met with a larger base width W_B. This has been discussed already in the context of doping variations in the base layer; we will see further down that the effect can be achieved much more elegantly using bandgap variations.

(ii) We can search for a way to increase the base doping concentration while maintaining a sufficiently high current gain. This we will discuss first.

The wide-gap emitter

The first approach to solving the base design dilemma had already been discussed when deriving the equation for the maximum current gain, Equation (2.121). It was noted that it would be advantageous to fabricate the emitter from a material with a lower intrinsic carrier concentration, $n_{i,E}$. Because

$$n_i = \sqrt{N_C N_V} e^{E_G/(2kT)},$$

the bandgap in the emitter needs to be increased. Then the maximum current gain becomes

$$B_{\max} = \frac{J_C}{J_B} = \frac{J_{\text{neb}}}{J_{\text{pbe}}} = \frac{D_{\text{n,B}}}{D_{\text{p,E}}} \frac{W_E}{W_B} \frac{N_{\text{D,E}}}{N_{\text{A,B}}} \frac{n_{\text{i,b}}^2}{n_{\text{i,E}}^2} \sim \frac{N_{\text{D,E}}}{N_{\text{A,B}}} e^{\Delta E_G / kT} \tag{2.163}$$

with $\Delta E_G = E_{g,E} - E_{g,B}$, neglecting secondary effects like the differences in effective densities of states in the two materials.

The enhancement factor $e^{\Delta E_G/kT}$ can have very large values. Consider as an example:

Emitter: $Al_{0.25}Ga_{0.75}As$ with $E_G = 1.74\,\text{eV}$
Base: GaAs with $E_G = 1.42\,\text{eV}$
This results in $e^{\Delta E_G/kT} = 2.2 \times 10^5$ at room temperature!

In a technical transistor, this current gain enhancement is traded for a dramatic increase in base doping with a simultaneous decrease in emitter doping. For a Si homojunction transistor, a typical doping combination is $N_{\text{D,E,typ}} = 10^{20}\,\text{cm}^{-3}$, $N_{\text{A,B,typ}} = 10^{18}\,\text{cm}^{-3}$, whereas for an AlGaAs/GaAs HBT, $N_{\text{D,E,typ}} = 5 \times 10^{17}\,\text{cm}^{-3}$, $N_{\text{A,B,typ}} = 4 \times 10^{19}\,\text{cm}^{-3}$. The reduction in emitter doping in the HBT is necessary to maintain an adequate reverse breakdown voltage of the base–emitter junction – the effect of current gain on the emitter–collector breakdown voltage BV_{CEO} was already discussed (see Equation (2.160)).

Figure 2.73 shows the band diagram of a wide-gap emitter transistor with a graded emitter–base heterostructure. In a graded heterostructure, the two materials used for the emitter and the base are allowed to intermix over a certain distance. We note that the injection of holes from the base into the emitter, which constitutes a major part of the base current, now faces a much larger potential wall than the injection of electrons from the emitter into the base. This provides the intuitive explanation for the potentially huge increase in current gain.

In an abrupt heterostructure, such as considered for the HEMT (Figure 2.17), we need to consider the effect of the conduction band and valence band discontinuities resulting

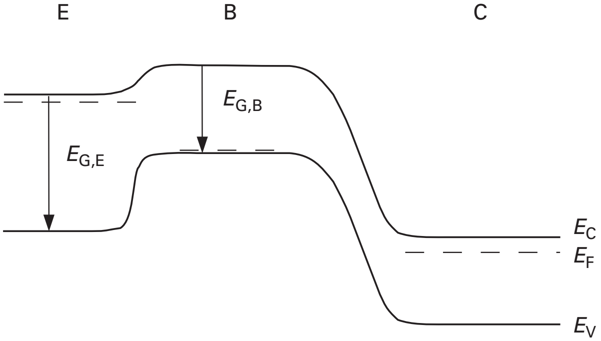


Fig. 2.73 Band diagram of a wide-gap emitter HBT under bias ($V_{BE} > 0$, $V_{CB} > 0$) with graded heterojunction.

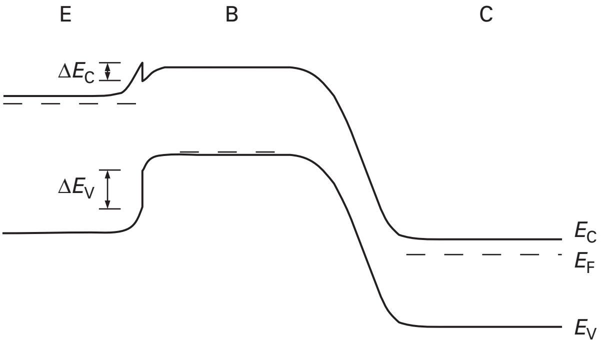


Fig. 2.74 Band diagram of an abrupt heterojunction wide-gap emitter HBT under bias ($V_{BE} > 0$, $V_{CB} > 0$).

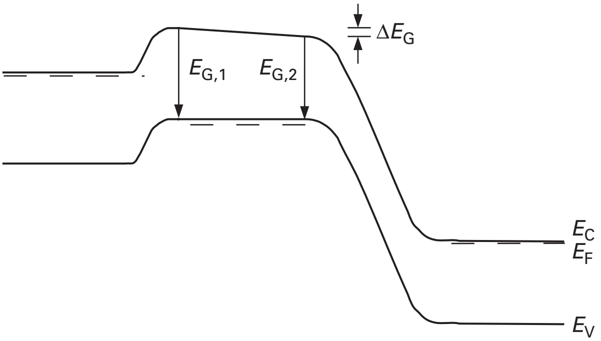


Fig. 2.75 Introduction of a drift field in the base using bandgap variation.

from Anderson's rule (Section 1.20.1). The conduction band discontinuity will lead to an additional energy barrier for electrons (see Figure 2.74).

Assuming purely thermionic emission of electrons over the conduction band barrier, we can write in first order for the maximum current gain of the abrupt HBT, compare Equation (2.163):

$$\begin{aligned} B_{\text{max,abrupt}} &\approx B_{\text{max,graded}}\, e^{-\Delta E_C/kT} \\ &\sim \frac{N_{D,E}}{N_{A,B}} e^{\Delta E_V/kT}. \end{aligned} \tag{2.164}$$

The enhancement factor in this case is only related to the part of ΔE_G which occurs in the valence band. In the above example, $Al_{0.25}Ga_{0.75}As$/GaAs, the enhancement factor is now only 107, because $\Delta E_V = 0.38\,\Delta E_C$.

Drift base

Using compositional grading in the base, we can also introduce a drift field in the base, as shown in Figure 2.75. The emitter has now the same band gap $E_{G,1}$ as the base immediately adjacent to the emitter–base junction. The band gap is then reduced towards the collector to $E_{G,2} < E_{G,1}$. The resulting bandgap difference ΔE_G reduces the base transit time [9].

$$\tau_{B,\text{graded}} = \tau_{B,\text{ungraded}} \frac{2}{\Delta E_G/kT} \left(1 - \frac{1 - e^{-\Delta E_G/kT}}{\frac{\Delta E_G}{kT}}\right). \tag{2.165}$$

The bandgap reduction is very efficient in reducing the transit time – a modest $\Delta E_G = 4kT$ results in a 62% reduction of the transit time.

HBT implementations

Group III–V HBTs.

HBTs fabricated from group III–V materials such as AlGaAs/GaAs, GaInP/GaAs or InP/InGaAs typically have multiple-mesa structures such as the cross-section shown in Figure 2.76. Due to its cross-sectional shape, it is frequently referred to as a *wedding cake* structure. The structure can be fabricated with a minimum number of masks; the base contacts are usually self-aligned to the emitter mesa using a deliberate undercut of the emitter contact.

While cost-effective in production, this structure has three major drawbacks:

(i) The topology is strongly non-planar. This makes realisations of sub-micron lateral feature sizes difficult, as well as the implementation of multi-level interconnect systems.

(ii) The necessary area for the base contacts and allowances for alignment accuracy necessarily lead to a base–collector area which is substantially larger than the base–emitter junction area. This leads to a larger-than-necessary base–collector capacitance C_{BC}, which in turn lowers the maximum frequency of oscillation (see Equation (2.153)).

(iii) The base–emitter junction is not embedded in semiconductor material, but reaches the less-than-ideal interface with the passivation layer. This gives rise to enhanced surface recombination currents, which increase the non-ideal portion of the base current (see Equation (2.122)). As a consequence, III–V HBTs have a current gain which is strongly dependent on the collector current.

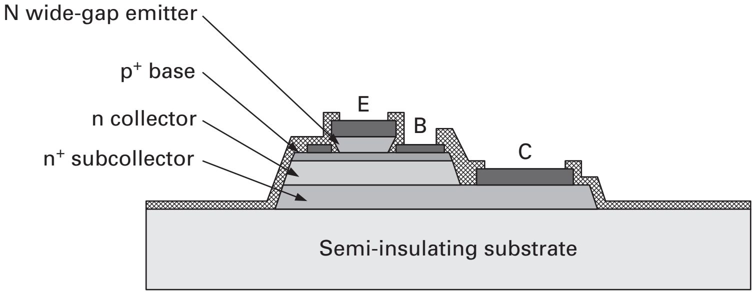


Fig. 2.76 Generic HBT structure typical of III–V semiconductor materials.

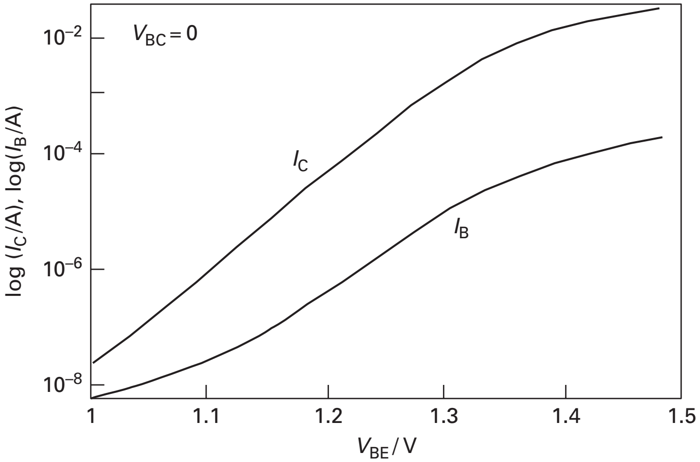


Fig. 2.77 Gummel plot of a typical AlGaAs/GaAs HBT.

Figure 2.77 shows the Gummel plot representation of the collector and base currents of a typical AlGaAs/GaAs HBT. The Gummel plot displays the currents in a semi-logarithmic way as a function of V_{BE}, for $V_{BC} = 0$. Ideally, base current and collector current both have an emission factor (ideality factor) of 1 (see Equation (2.121)). This would result in perfectly parallel curves for $\log I_C$ and $\log I_B$, which is not the case here. The non-ideal base currents, with their emission factor >1, are seen predominantly for very low base–emitter voltages and reduce the current gain there. This is a problem especially for low-noise operation. Furthermore, the surface recombination currents give rise to low-frequency noise ($1/f$ or generation-recombination type noise), with negative impact e.g. on the phase noise of microwave oscillators.

For the wide-gap emitter, AlGaAs was long the material of choice. It was more recently largely replaced with $Ga_{1-x}In_xP$, which for an In mole fraction of 0.5 is lattice-matched to GaAs. GaInP as the emitter material has several advantages:

- The bandgap difference at the GaInP/GaAs junction occurs predominantly in the valence band.
- GaInP can be selectively etched with respect to GaAs, allowing for an automatic etch stop on the base layer when structuring the emitter mesa.
- The reliability of the base–emitter junction under current stress was shown to be substantially higher.

For optoelectronic integration and millimetre-wave applications, HBTs are also being fabricated in InP substrates. The base is now $In_{0.53}Ga_{0.47}As$. As was discussed for the pseudomorphic HEMT structure, the electron mobility is substantially higher than for GaAs, leading to a much shorter base transit time (see Equation (2.143)). The hole mobility in InGaAs, however, is lower than in GaAs, leading to an increased base resistance. The emitter material is either InP or $In_{0.52}Al_{0.48}As$.

HBTs with InGaAs base and collector regions have a major problem with low collector–base breakdown voltages, because the lower band gap of InGaAs lowers

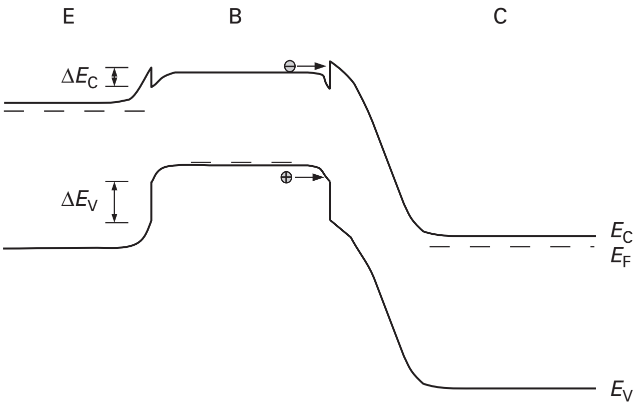


Fig. 2.78 Band diagram of a DHBT under active forward bias.

the threshold for impact ionisation. Therefore, double-heterojunction bipolar transistors (DHBTs) are frequently used in this material system, with either InAlAs or InP as the collector material. DHBTs, however, introduce another problem, which is illustrated in Figure 2.78. The heterostructure at the base–collector interface introduced additional energy barriers in the conduction and valence bands. The conduction band barrier impedes the collection of electrons and lowers the current gain.

The valence band barrier is important especially after the onset of the Kirk effect (see Section 'Kirk effect'). After all fixed donors in the collector have been neutralised, charge neutrality requires that with a further increase in current, free holes from the base are injected into the collector region. Due to the valence band barrier, however, this is restricted in the DHBT. As a result, holes accumulate at the base–collector interface, the bands bend upwards, and the electron barrier in the conduction band becomes higher. This leads to a much more severe deterioration of transistor parameters in the high-current regime. The problem can be avoided either by compositionally grading the base–collector junction, or by introducing a composite collector structure where the heterojunction is offset away from the base–collector p–n junction into the collector [18].

Another problem related to the collector region is the aforementioned substantial C_{BC} due to the triple mesa structure of III–V HBTs (see Figure 2.76). One solution is to fabricate the subcollector in a buried fashion by ion implantation beneath a semi-insulating layer, and to connect it to the collector contacts and to the collector proper via 'sinker' implants.

An example for such a structure is shown in Figure 2.79 [44]. The subcollector is implanted into the semi-insulating InP substrate; the layer above has a drastically increased conductivity due to an Fe implant. Heavily n-doped local implants connect the buried subcollector to the collector contacts and to the collector itself, which is grown together with the base and emitter layers subsequently. While the base–collector area is not changed here, C_{BC} is still reduced because the collector layer itself is depleted in normal operation and therefore the reduced collector–subcollector interface area diminishes the capacitance. Further, C_{BC} is less V_{CE}-dependent, which enhances linearity.

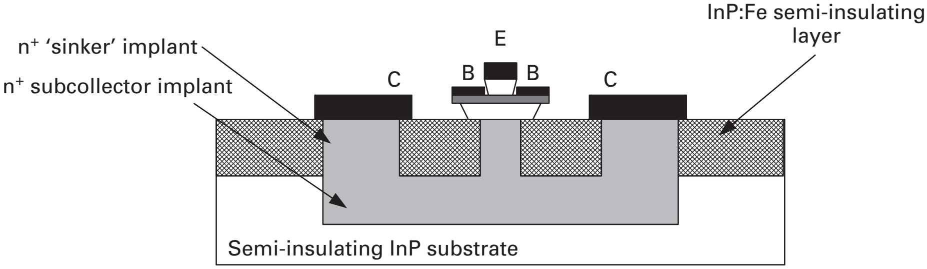


Fig. 2.79 InP/InGaAs HBT structure with a buried subcollector.

Finally, burying the subcollector improves the planarity of the structure. In the example shown, the InP/InGaAs DHBT demonstrated $f_T = 350\,\text{GHz}$ and $f_{\max} = 400\,\text{GHz}$.

Another method to restrict C_{BC} is by damage implant through the base contact window prior to metal deposition.

C_{BC} can be further reduced by eliminating the subcollector altogether and attaching the contact directly to the collector layer. To achieve this, the HBT structure must be inverted, i.e. the emitter contact is now at the bottom [50]. The collector must be accurately aligned to the buried emitter structure. A new problem which arises in classical collector-up HBTs is that now the emitter contact must be made laterally through a 'sub-emitter' layer, increasing the crucial emitter series resistance.

The latter problem is addressed in the very aggressive 'transferred substrate' device, where the HBT structure is grown emitter-up. The emitter and base/collector structures are fabricated first. The structure is then flipped around and the emitter is attached to a Au metal structure which provides for the low-resistivity lateral emitter contact. The InP substrate is then removed and the collector contact is structured. The collector layer outside of the contact area is fully depleted and does not add extra capacitance.

A schematic cross-section is shown in Figure 2.80 [47]. Together with submicron scaling ($0.4\,\mu\text{m}\cdot 6\,\mu\text{m}$ emitter area, $0.7\,\mu\text{m}\cdot 10\,\mu\text{m}$ collector area), an InP/InGaAs HBT with a transferred-substrate structure exhibited $f_T = 204\,\text{GHz}$ and $f_{\max} = 1080\,\text{GHz}$.

Si/SiGe HBTs.

Unlike III–V HBTs, which are usually fabricated from lattice-matched heterostructures, HBTs in the $\text{Si}_{1-x}\text{Ge}_x$ material system are necessarily pseudomorphic (Section 1.20), which delayed their practical realisation until the late 1980s. They are commercially available since 1998 and have enjoyed an unparalleled technical and commercial success.

Due to the large difference in lattice constant between Si ($a = 5.43\,\text{Å}$) and Ge ($a = 5.66\,\text{Å}$), an elastically strained SiGe layer will necessarily be very thin, as was shown in Figure 1.35. The use of SiGe compounds is therefore restricted to the base layer – everything else is silicon, making Si/SiGe transistors necessarily DHBTs.

The SiGe alloy can be used in two different ways:

(i) The base may start with a zero Ge mole fraction at the emitter–base junction, and be increased towards the base–collector junction. The corresponding decrease in

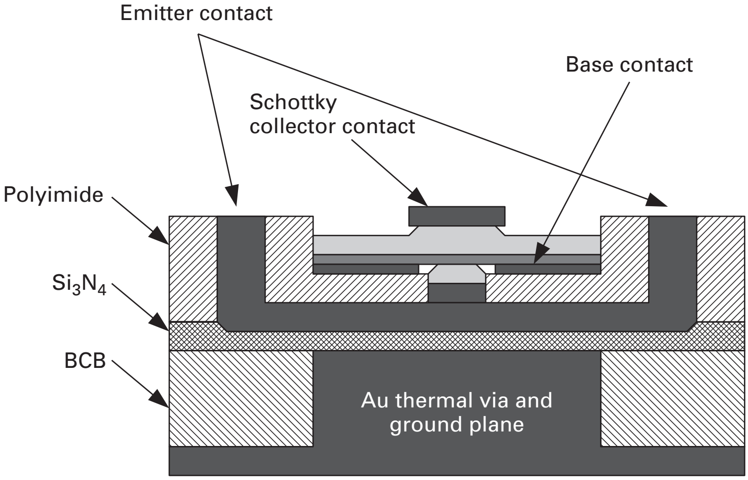


Fig. 2.80 Schematic cross-section of a transferred-substrate collector-up HBT [47].

band gap creates a built-in field for electrons in n–p–n transistors – a drift-base transistor results, with a band diagram similar to the one shown in Figure 2.75, except that the base–collector interface is now a hetero-interface. Pseudomorphic SiGe layers sandwiched between relaxed Si layers have an interesting property: the bandgap difference is almost exclusively in the valence band (see Figure 1.35). Hence, there is no parasitic conduction band barrier, at least not until high-current effects set in and the hole pile-up against the base–collector valence band barrier makes the bands buckle upwards.

The major benefit of the built-in drift field is the reduction in base transit time given by Equation (2.165). Due to the emitter–base interface being a homojunction, it is bound by similar base doping limitations as the homojunction bipolar transistor.

This Si/SiGe drift-base concept has the significant advantage that the average Ge mole fraction in the base, and with it the built-in mechanical strain, is very low. In terms of processing temperatures, these transistors are fully CMOS-compatible. The drift-base heterostructure transistor is therefore the approach of choice in most Si/SiGe BiCMOS processes.

(ii) Si/SiGe heterostructures can, of course, also be used to fabricate a wide-gap emitter structure. In this case, the Ge mole fraction is already significant at the emitter–base junction, leading to a significant valence band discontinuity, which allows to dramatically increase the base doping concentration (see Equation (2.164)). In these transistors, the Ge mole fraction is typically constant across the base.

The major benefit of the wide-gap emitter structure is the high base doping concentration and resulting low base sheet resistance, which allows to achieve high cutoff frequencies despite very relaxed lateral scaling rules, e.g. f_T, f_{max} = 80 GHz with 0.8 μm design rules [55] .

The two approaches may be combined, of course – the Ge mole fraction profile may start with a moderate non-zero value at the emitter–base interface and increase towards the collector to a higher value at the base–collector junction, combining a hole-blocking effect towards the emitter with a built-in drift field towards the collector. This is called a *trapezoidal* Germanium profile in the base. The current gain in this case is [26]:

$$B_{\mathrm{SiGe}} = B_{\mathrm{Si}}\,\eta\,\gamma\,\frac{E_{\mathrm{G}}(y=0)-E_{\mathrm{G}}(y=W_{\mathrm{B}})}{kT}\,\frac{e^{\Delta E_{\mathrm{G}}(y=0)/\mathrm{kT}}}{1-e^{-[E_{\mathrm{G}}(y=0)-E_{\mathrm{G}}(y=W_{\mathrm{B}})/kT]}}, \quad (2.166)$$

where B_{Si} is the current gain of a homojunction transistor with the same geometry, η is the ratio of the position-averaged minority mobilities in the base of the two transistors, and γ is the position-averaged ratio of the density of states product ($N_{\mathrm{V}} \cdot N_{\mathrm{C}}$) across the base. The emitter–base junction is at $y = 0$, and the base–collector interface at $y = W_{\mathrm{B}}$.

Since

$$\lim_{\mathrm{x}\to 0} \frac{x}{1-e^{-x}} = 1,$$

Equation (2.166) reverts to Equation (2.163) for $E_{\mathrm{G}}(y=0) = E_{\mathrm{G}}(y=W_{\mathrm{B}})$. On the other hand, we see that having a pure drift-base profile ($E_{\mathrm{G}}(y=0)=0$) also results in a certain increase in the current gain.

Irrespective of the Ge profile in the base, a major advantage of the Si/SiGe HBTs is that they can harness the full potential of silicon technology, especially aggressive lateral scaling developed predominantly for CMOS process, different isolation techniques, and $\mathrm{SiO_2}$ as a highly stable native oxide.

A typical SiGe HBT in a commercially available technology has a structure similar to the schematic in Figure 2.81. Note the very planar structure compared to III–V HBTs, and the extensive use of $\mathrm{SiO_2}$ isolation. The $\mathrm{n^+}$ subcollector is created by ion implantation, after which a low n-doped Si layer is epitaxially grown and converted to $\mathrm{SiO_2}$ by local oxidisation, except in the areas below the collector contact and where the transistor structure will be. The collector area is doped using selective ion implantation, which allows for several collector doping concentrations on one chip, with different f_{T} versus BV_{CEO} trade-offs. The transistor structure is then grown selectively in the transistor window.

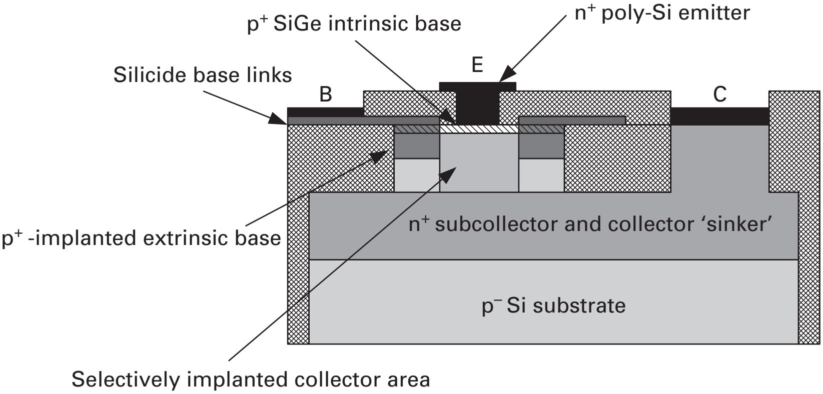

p⁺ SiGe intrinsic base
n⁺ poly-Si emitter
Silicide base links
E
B
C
p⁺ -implanted extrinsic base
n⁺ subcollector and collector ‘sinker’
p⁻ Si substrate
Selectively implanted collector area

Fig. 2.81 Planar Si/SiGe HBT with implanted extrinsic base region.

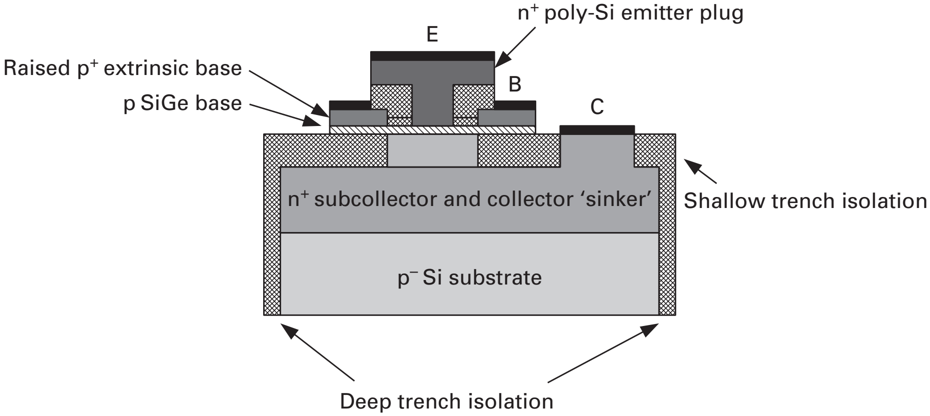


Fig. 2.82 Planar Si/SiGe HBT with a raised extrinsic base structure.

The extrinsic base resistance is reduced by heavy p^+ implantation. This works well if the transistor is not aggressively scaled laterally. It does create, however, crystal faults immediately adjacent to the intrinsic base, which leads to enhanced diffusion of the p-dopant in the base and is a major obstacle to fabricating deep submicron lateral emitter widths. Additionally, the close proximity of the p^+ extrinsic base and the selectively implanted collector increases the base–collector capacitance.

The latter problems are solved using a raised base structure, where a p^+ extrinsic layer is grown selectively on top of the base adjacent to the emitter, as shown schematically in Figure 2.82 [14]. A combination of these techniques with deep submicron scaling led production Si/SiGe HBT technologies to achieve f_T and f_{max} values above 200 GHz.

III–V versus Si/SiGe HBTs – a brief comparison.

Si/SiGe HBTs displaced III–V HBTs in many applications primarily due to their supreme potential for large-scale integration, owing to their technological proximity to very mature Si processes. In terms of raw speed, as measured from f_T and f_{max}, record values are still scored by InP/InGaAs devices, but Si/SiGe HBTs are competitive, because they compensate for material deficiencies (e.g. the much lower electron mobility versus InGaAs), by aggressive lateral scaling and superior suppression of parasitic capacitances. Further, Si has a significantly higher thermal conductivity than either GaAs or InP, which makes the thermal management of dense transistor arrays easier.

In the area of microwave power amplification, however, III–V-based HBTs have an inherent advantage. When deriving the Johnson limit, Equation (2.161), we recognised the importance of the product of drift saturation velocity v_{sat} and the electrical field necessary for impact ionisation $\mathcal{E}_{crit}$. Taking v_{sat} at an electric field of 10 kV cm^{-1}, and $\mathcal{E}_{sat}$ at a donor doping concentration of 10^{17} cm^{-3}, this product is shown for Si, GaAs and InP in Table 2.1.

When comparing practical transistors, the ratio in $f_T\,BV_{CEO}$ between Si- and GaAs-based HBTs may appear even larger than the factor of 1.6 suggested by Table 2.1; but

Table 2.1 $v_{\text{sat}} \cdot \mathcal{E}_{\text{crit}}$ product for Si, GaAs and InP

Si	5,000	GHz V
GaAs	8,000	GHz V
InP	22,000	GHz V

this is due to the generally lower current gain in the GaAs devices, which increases BV_{CEO}.

2.5.6 Large-signal modelling

Bipolar transistor models have become increasingly complex. An exhaustive description of popular large-signal formulations is beyond the scope of this book. The following will concentrate on emphasising the major differences between the models, with respect to active forward operation of the transistor, quasi-static non-linear equations and avoiding extreme areas of operation.

The Ebers–Moll model

The Ebers–Moll equivalent circuit model was historically the first compact model of the bipolar transistor [16]. It approximates the intrinsic transistor as a network of two junction diodes and two current-controlled current sources (see Figure 2.83(a)). The parameter A_{F} is the common-base current gain in forward operation. A_{R} is the common-base current gain in reverse operation (emitter and base interchanged), which is not being considered here. For active forward operation, the base–collector diode is reverse-biased. Further, $A_{\text{R}}\, I_{\text{C}}$ is much smaller than the forward current through the base–emitter diode and can hence be neglected. The resulting simplified equivalent circuit is shown in Figure 2.83(b).

The emitter current in active forward operation is

$$I_{\text{E}} = -I_{\text{SBE}} \left(e^{V_{\text{BE}}/(N_{\text{E}} V_{\text{T}})} - 1 \right), \tag{2.167}$$

where I_{SBE} is the base–emitter saturation current, N_{E} the base–emitter ideality factor and $V_{\text{T}} = kT/q$ the thermal voltage.

Using the full equivalent circuit, the Ebers–Moll equivalent circuit can account for saturation (both diodes are forward-biased), but cannot model Early and Kirk effects. Further, the current dependence of the current gain at low V_{BE} can also not be included.

The Gummel–Poon model

An improved model of the bipolar transistor which is capable of including more of the non-ideal effects of bipolar transistors was introduced by Gummel and Poon in 1970 [22].

The equivalent circuit (Figure 2.84) uses a voltage-controlled current source for the collector current – the transistor is seen in common-emitter configuration here. The

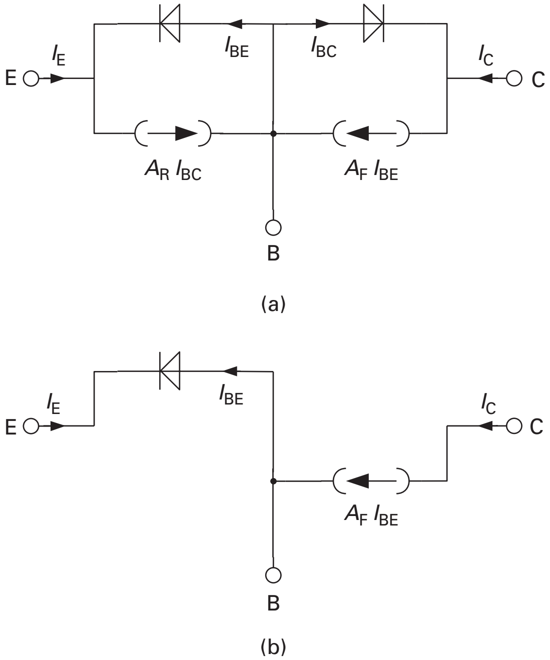


Fig. 2.83 Quasi-static Ebers–Moll equivalent circuit of the intrinsic bipolar transistor: (a) for forward and reverse operation; (b) simplified for forward active operation only.

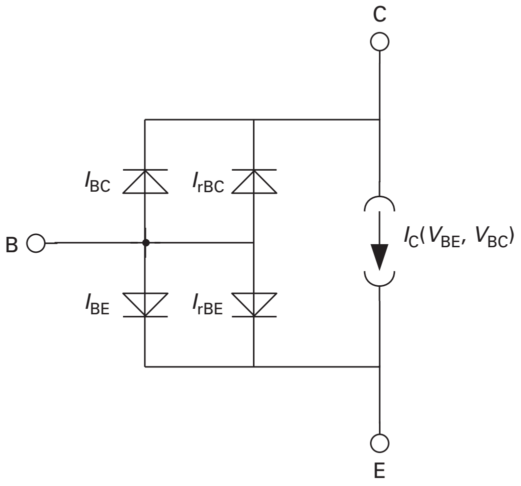


Fig. 2.84 Gummel–Poon quasi-static equivalent circuit of the intrinsic bipolar transistor.

use of two parallel diodes for the base–emitter and base–collector junctions allows to include both the ideal (I_{BE}, I_{BC}) and non-ideal (I_{rBE}, I_{rBC}) current contributions in forward and reverse operations of the transistor, with different emission factors. Hence, the current gain reduction at low V_{BE} (or V_{BC} in reverse operation) can be easily included.

The collector current formulation (shown here for forward operation only) uses a saturation current I_S and a charge control parameter Q_B:

$$I_C = \frac{I_S}{Q_B}\left(e^{V_{BE}/(N_F\,V_T)} - e^{V_{BC}/(N_R\,V_T)}\right), \tag{2.168}$$

where N_F is the ideality factor in forward direction and N_R the ideality factor in reverse direction. Early and Kirk effects are modelled through the charge control parameter:

$$\begin{aligned} Q_B &= \frac{Q_1}{2}\left(1+\sqrt{1+4Q_2}\right) \\ Q_1 &= \left(1-\frac{V_{CB}}{V_{AF}}\right)^{-1} \\ Q_2 &= \frac{I_S}{IKF}\left(e^{V_{BE}/(N_F\,V_T)}-1\right), \end{aligned} \tag{2.169}$$

where V_{AF} is the Early voltage in forward direction and IKF is the knee current for the onset of high-current effects in the forward direction.

In the active forward regime, I_{BC} and I_{rBC} can be neglected and the base current becomes:

$$\begin{aligned} I_B &= I_{BE} + I_{rBE} \\ &= \frac{I_S}{BF}\left(e^{V_{BE}/(N_F\,V_T)}-1\right) + I_{SE}\left(e^{V_{BE}/(N_E\,V_T)}-1\right), \end{aligned} \tag{2.170}$$

where BF is the ideal forward current gain, I_{SE} the saturation current of the non-ideal base current and N_E the emission factor of the non-ideal base current.

Extension of the equivalent circuit to the dynamic case is shown in Figure 2.85. In active forward operation, C_{BE} contains both the diffusion capacitance Equation (2.141) and the junction capacitance of the base–emitter junction, while C_{BC} is a junction capacitance only. The capacitance C_{CS} models the reverse-biased junction between the (sub-)collector region and the substrate node. On semi-insulating substrates, it is not necessary.

The Gummel–Poon model also deals with the bias dependence of the base resistance which is frequently observed – R_B decreases from a higher value at low collector current to a much lower value at high collector current. This effect is due to a concentration of the emitter current towards the emitter periphery with increasing current – due to the lateral voltage drop in the base layer, the local base–emitter voltage is higher and closer to the base contact. As the local current depends exponentially on the local V_{BE}, even a small voltage change can lead to substantial redistributions in current. The base resistance decreases because the inner parts of the emitter–base area get increasingly detached. Due to the much lower base sheet resistance, this effect is less pronounced in wide-gap emitter HBTs. The Gummel–Poon model describes this effect using the parameters RB, RBM and IRB:

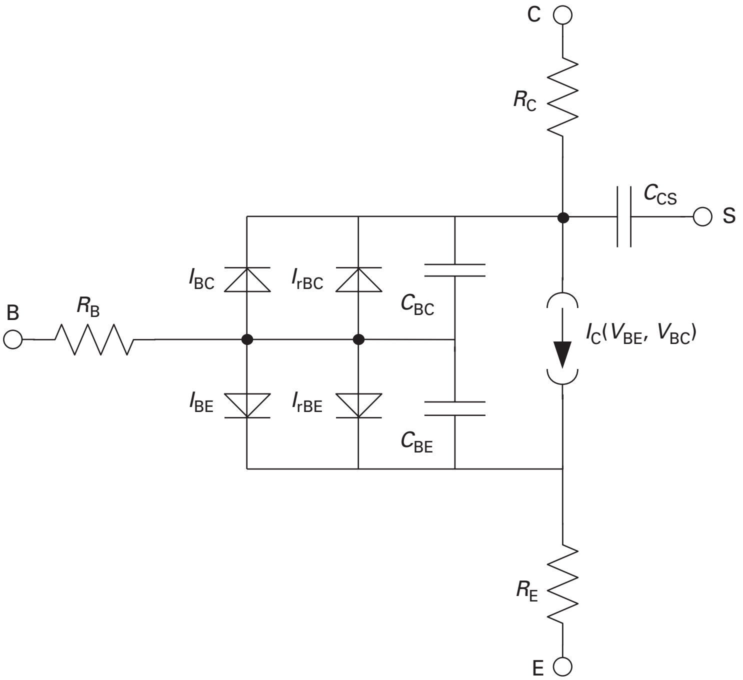


Fig. 2.85 Gummel–Poon equivalent circuit with parasitic elements and substrate node.

$$R_B(I_B) = \mathrm{RBM} + 3\,(\mathrm{RB} - \mathrm{RBM})\frac{\tan(z) - z}{z \cdot \tan^2(z)}$$

$$z = \frac{\sqrt{1 + \left(\frac{12}{\pi}\right)^2 \frac{I_B}{IRB}} - 1}{\left(\frac{24}{\pi^2}\right)\sqrt{\frac{I_B}{IRB}}}. \tag{2.171}$$

In total, the Gummel–Poon model implemented in SPICE contains 42 model parameters. A full discussion is therefore beyond the scope of this book.

The VBIC95 model

The VBIC95 model [37] is an extension of the Gummel–Poon model. Among others, the following problems are being addressed:

- The description of base width modulation using a constant Early voltage is a simplification which only applies to small V_{CE}.
- Self-thermal effects are not included in Gummel–Poon, yet play an important role especially for power amplifiers.
- The collector resistance is not a constant, but depends on V_{CB}, because the undepleted part of the collector increases the series resistance.
- Avalanche breakdown in the collector space charge region needs to be included.

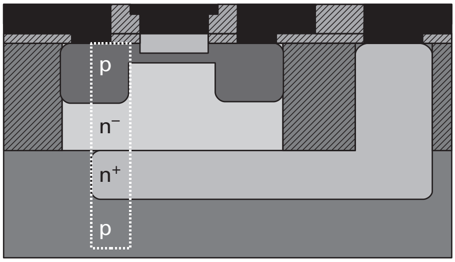


Fig. 2.86 Generic n–p–n BJT cross-section highlighting the parasitic p–n–p transistor.

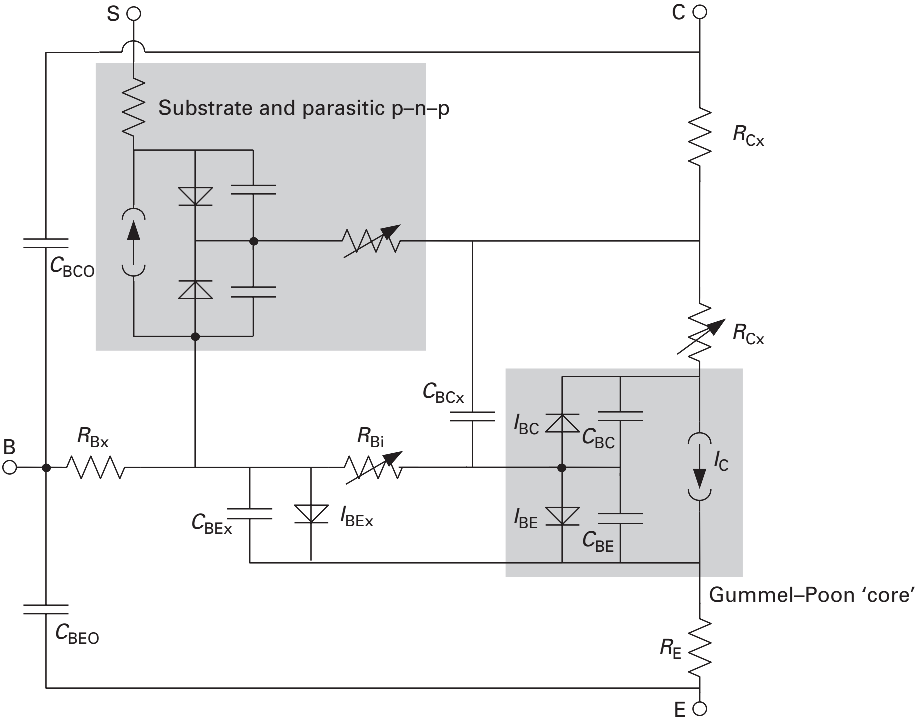


Fig. 2.87 VBIC95 equivalent circuit.

- A major addition has been the implementation of a subcircuit for the parasitic p–n–p transistor, which is formed in Si-based bipolar transistors between the base, the collector and the substrate. This parasitic p–n–p can be easily recognised in Figure 2.86. Under certain bias conditions, it may sink an unexpectedly large current between the base terminal and the substrate node.

Figure 2.87 shows the VBIC95 in a representation which emphasises its Gummel–Poon heritage. The distributed nature of the base resistance is accounted for. The collector resistance is now separated into a bias-dependent part which symbolises the contact, sinker and subcollector resistances, and a bias-dependent internal part modelling the undepleted part of the collector proper. The substrate network is now much more complex and includes the parasitic p–n–p as a separate Gummel–Poon type equivalent circuit. Additional capacitances C_{BEO} and C_{BCO} have been added to account for overlap capacitances between the poly-Si emitter plug and the base and collector contacts, respectively.

Note that these are the only linear capacitances – all other capacitances are bias-dependent, even though this has not been noted in the equivalent circuit to enhance readability.

The VBIC95 model implemented in newer versions of SPICE has 85 parameters, which also hints at the complexity of setting up such a model from measurements.

The MEXTRAM model

The MEXTRAM bipolar transistor model was created by Philips N. V. [42] and released into the public domain in 1993. It has been implemented in several industry standard simulation environments, such as several versions of SPICE and Agilent Advanced Design System (ADS).

The equivalent circuit (Figure 2.88) shows stronger deviations from the Gummel–Poon topology. The main current equation, however, shows the similarity:

$$I_{\mathrm{N}} = \frac{I_{\mathrm{S}}}{q_{\mathrm{b}}}\left(e^{V_{\mathrm{B2E1}}/V_{\mathrm{T}}} - e^{V^{*}_{\mathrm{B2C2}}/V_{\mathrm{T}}}\right). \tag{2.172}$$

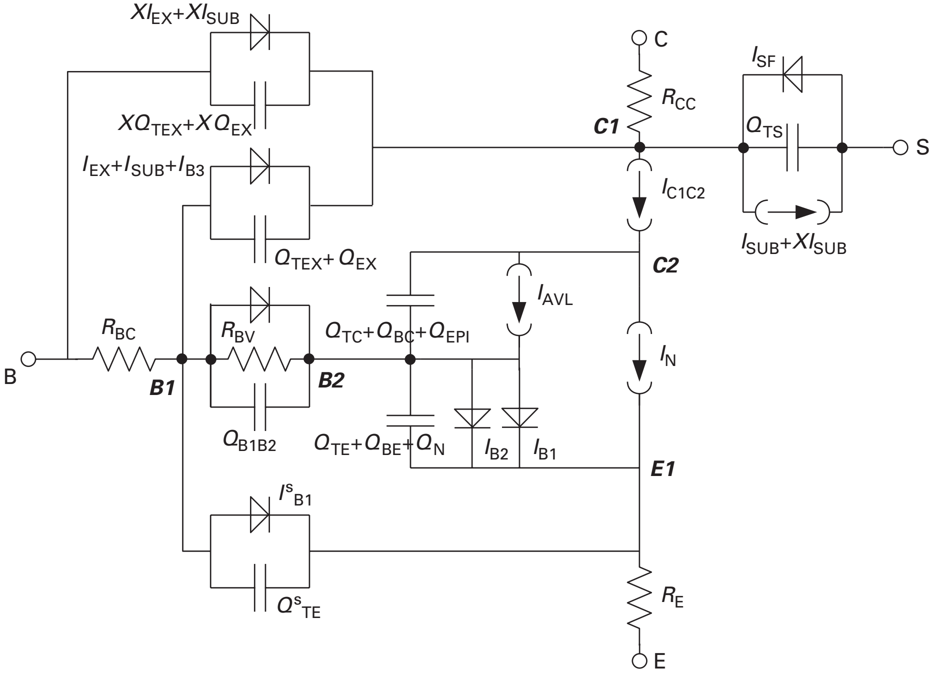


Fig. 2.88 MEXTRAM model equivalent circuit topology.

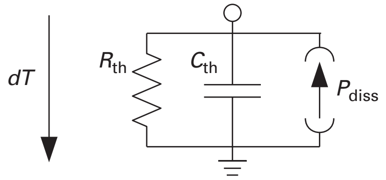


Fig. 2.89 Thermal equivalent circuit used by VBIC95 and MEXTRAM.

Here, V_{B2E1} is the voltage between nodes B2 and E1, while V^*_{B2C2} is a calculated quantity which corresponds to the voltage drop between nodes B2 and C2 – for an explanation of this and other intricacies, please refer to the MEXTRAM documentation [42]. q_b is the normalised base charge, which is used to model both Early and high-current effects. This is conceptually as in Gummel–Poon, but the implemented equations provide a higher level of accuracy, for example in the bias dependence of the Early voltage.

The base is modelled as a distributed structure – this is a must for accurate modelling at elevated frequencies. The base–collector capacitance is split into three partial capacitances. The model does not only distinguish between an external and an internal part of the base, but models the sidewall interface between base and emitter separately (I^S_{B1} and Q^S_{RE}). Two diodes are used to model the ideal and non-ideal parts of the base current.

The parasitic p–n–p transistor is also modelled here, even though this is less obvious – the current source between nodes $C1$ and S is exponentially controlled by the intrinsic base–collector voltage, V_{B1C1}. The reverse behaviour of the parasitic p–n–p, however, is not modelled.

The major claimed advantage over VBIC95 is related to the modelling of high-current effects [30]. This is especially important for double-heterostructure transistors, such as Si/SiGe HBTs (see Figure 2.78 and its associated discussion).

MEXTRAM has also been extended to account for neutral base recombination and base drift fields introduced through bandgap variations, again in an effort to make this model especially useful for Si/SiGe HBTs.

Self-thermal effects are being simulated in VBIC95 and MEXTRAM in the same way (see Figure 2.89). The model calculates the sum of all powers dissipated in resistors and space charge regions as P_{diss}. In the electric equivalent circuit, P_{diss} is treated as a current which creates a voltage drop dT of the parallel connection R_{th}, C_{th}, which establishes the thermal time constant τ_{th}. dT is analogous to the temperature difference between the device (taken to be at one single temperature – a simplification) and the ambient. It is then used as an additional control voltage for the bias-dependent current sources. This is a very common technique to include self-thermal effects, but it neglects the fact that the thermal conductivity of semiconductors, and with it the thermal resistance R_{th}, is temperature-dependent. With increasing temperature T, R_{th} will also increase.

The HICUM model

The acronym of the last model to be discussed here already indicates its major claimed advantage – HICUM [51] stands for **Hi**gh-**C**urrent **M**odel. Aside from being a general purpose non-linear bipolar model, with special emphasis on high-speed applications, it

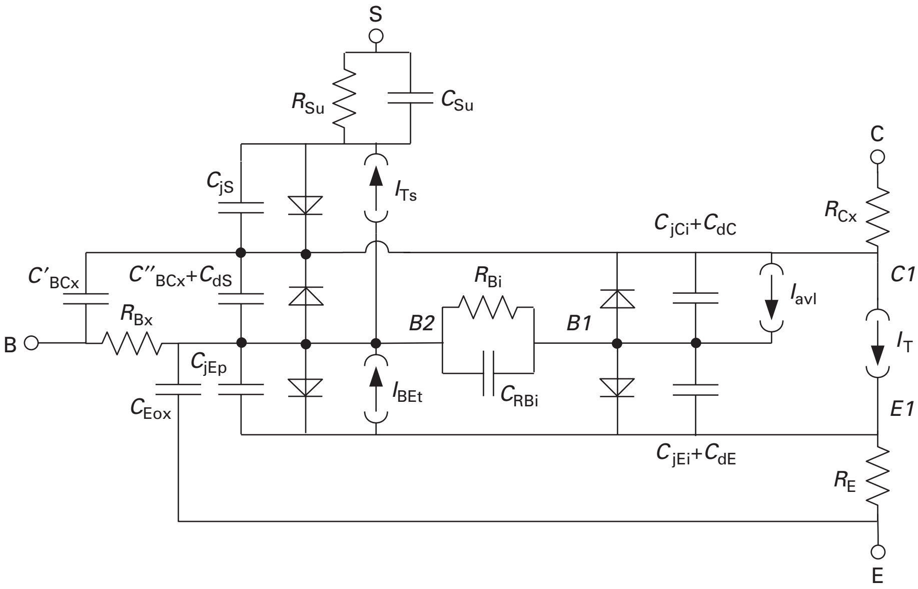


Fig. 2.90 HICUM equivalent circuit (adapted from [52]).

concentrates especially on an accurate prediction of high-current effect. Remember that high-speed bipolar operation will occur at high collector current densities, minimising the emitter charging time (see Figure 2.69). Accurate assessment of high-current effects is therefore a must for any simulator with high-speed emphasis. HICUM's development started in 1980 and it has been implemented in commercial simulation environments since 1994. Its model equations take a semi-physical approach to allow scalability and a certain degree of prediction. A companion program, TRADICA, facilitates the latter two issues.[12]

Figure 2.90 shows the equivalent circuit used in HICUM. The modelling of self-thermal effects is done analogous to Figure 2.89 and is not shown here again.

The HICUM model does not have an equivalent element to the Gummel–Poon ideal current gain BF, but calculates the collector and base currents independently and treats the current gain as a derived entity. The main current in the HICUM model is the transfer current I_{T}m, which can be compared to the intrinsic collector current I_{C} (see Equation (2.168)):

$$I_{\mathrm{T}} = I_{\mathrm{S}} \frac{e^{V_{\mathrm{B1E1}}/VT} - e^{V_{\mathrm{B1C1}}/V_{\mathrm{T}}}}{\frac{Q_{\mathrm{p,T}}}{Q_{\mathrm{p0}}}}, \tag{2.173}$$

where Q_{p0} is the total hole charge in the base at zero bias. Note that unlike the Gummel–Poon expression, the exponential function does not have ideality factors. The deviation from non-ideal diode characteristics is handled in the bias-dependent hole charge $Q_{\mathrm{p,T}}$.

[12] An introduction to TRADICA is available at www.iee.et.tu-dresden.de/s̃chroter/Trad/features.pdf.

The formulation for $Q_{\mathrm{p,T}}$ in HICUM allows to include the effect of strongly varying intrinsic carrier densities across the base layer, as necessary for the simulation of drift-base HBTs [53]. In the model, this is done through different weighting factors being applied to the depletion charges at the base–emitter and base–collector junctions.

The model can also accommodate the hole accumulation at the base–collector heterojunction, with current gain roll-off and transit time deterioration, as needed in DHBTs [54]. For an in-depth treatment of HICUM parameters, refer to [52].

Differences between BJTs and HBTs relevant to large-signal modelling

In general, the standard bipolar models discussed above are also applicable to HBTs, with appropriately chosen parameters.

An important deviation concerns self-heating effects. In homojunction bipolar transistors, both the collector saturation current and the current gain have positive temperature coefficients. The saturation current (see Equation (2.115)) increases because the intrinsic carrier concentration in the base depends exponentially on temperature. The current gain is limited by bandgap narrowing in the heavily doped emitter (see e.g. Equation (2.162)). The bandgap narrowing effect has a negative temperature coefficient, which lets the current gain increase with increasing temperature.

In a wide-gap emitter HBT, the saturation current equally has a positive temperature coefficient. The current gain, however, decreases with increasing temperature. To understand this, investigate again Equation (2.163):

$$B_{\mathrm{max}} \sim e^{\Delta E_{\mathrm{G}}/kT}$$

The enhancement factor therefore decreases with increasing temperature, because the valence band barrier gets less and less efficient.

A frequently observed self-thermal effect in HBTs is the current crush in multi-finger HBTs (see Figure 2.91). At moderate dissipated powers, all fingers will have approximately the same temperature and the current distribution is equal. With increasing V_{CE} and under constant base current, the collector current will gradually decrease due to the negative temperature coefficient of B. However, due to the strongly positive temperature coefficient of the saturation current, a finger which is slightly hotter than the others will draw more and more current, deviating it away from the others, and increase its temperature. This strongly non-linear positive feedback will lead to a situation where the hottest finger takes on all the current, increases dramatically in temperature, with a resulting sudden decrease in current gain. This effect cannot be modelled with the standard bipolar models.

Another important effect which cannot be simulated using the standard models is the rapid onset of high-current effects in DHBTs, discussed in the context of Figure 2.78.

A minor effect, but worth mentioning, is that the non-ideal base current, see Equation (2.122), may have a different V_{BE} dependence in HBTs. This is due to space charge region recombination associated with the conduction band discontinuity at an abrupt emitter–base heterojunction (see Figure 2.74). The potential well on the base side of the junction leads to an increased recombination, which depends in turn on the voltage across the junction.

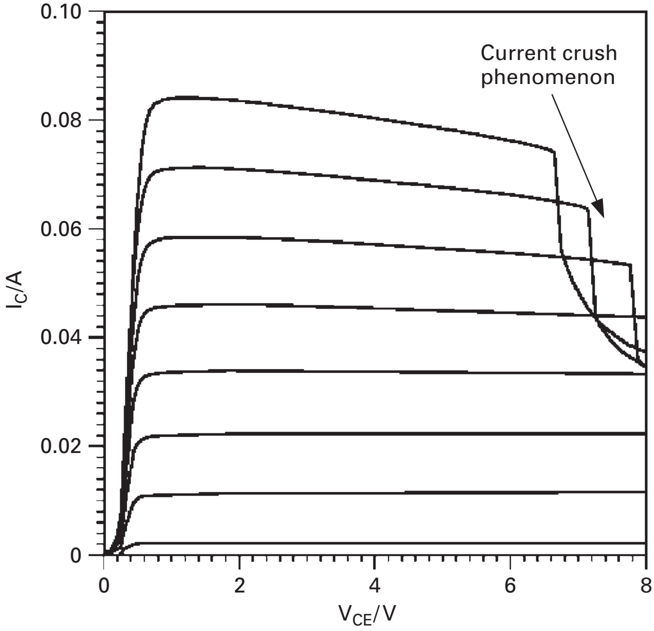


Fig. 2.91 Current crush phenomenon in a multi-finger HBT (http://parts.jpl.nasa.gov/mmic/mmic_complete.pdf). Courtesy NASA/JPL-Caltech.

2.6 Problems

(1) In a MESFET device, the following parameters are known from the fabrication process:

Gate	Channel	Drain/source
Ti, $\Phi_B = 0.7\,\text{eV}$,	$N_D = 1 \cdot 10^{17}\,\text{cm}^{-3}$,	$R_S = R_D = 10\,\Omega$
$L_G = 1\,\mu\text{m}$,	layer thickness	
$W_G = 100\,\mu\text{m}$,	$a = 0.3\,\mu\text{m}$,	
$R_G = 1\,\Omega$	$v_{sat} = 1.2 \cdot 10^7\,\text{cm}\,\text{s}^{-1}$	

a) Draw the qualitative band diagram under the gate in thermodynamic equilibrium.

b) Calculate the Schottky gate built-in voltage and the pinch-off voltage.

c) Assuming constant velocity throughout the channel, calculate the drain current for an applied bias of $V_{GS} - V_P = 2\,\text{V}$ and $V_{DS} = 3\,\text{V}$.

d) Calculate the device transconductance.

e) What is the expected transit frequency?

f) Calculate the expected minimum noise figure F_{min} at a frequency of 2 GHz.

(2) A GaAs MESFET with a gate width of $W_G = 100\,\mu\text{m}$ is specified with a transit frequency $f_T = 35\,\text{GHz}$ and a maximum frequency of oscillation $f_{max} = 50\,\text{GHz}$. The transconductance is $g_m = 21\,\text{mS}$, and the gate resistance is $R_G = 2\,\Omega$.

Estimate the gate-source capacitance and the gate-drain capacitance, assuming that the output conductance is negligible. Can you provide an estimated value for the minimum noise figure?

Due to a fabrication error, the gate resistance is increased to 5 Ω. What is the impact on f_T, $f_{\max}$ and $F_{\min}$?

(3) Why does a MOSFET require an overlap between the gate and the source implantation region? What does this imply for device capacitances?

(4) In an n-channel MOSFET with a *metal* gate electrode, the original gate metal is replaced by a metal with a smaller work function. Explain qualitatively the effect on the threshold voltage.

(5) In a MOSFET technology, the thickness of the field oxide is chosen such that under the highest possible voltage between metallisation and substrate, no inversion channel can form at the SiO_2/Si interface. Considering an Al metallisation with a work function of 4.1 eV and a bulk doping concentration of $N_A = 5 \cdot 10^{17}$, calculate the minimum thickness of the field oxide, if the maximum voltage between Al metallisation and the substrate is 5 V.

(6) A silicon-on-insulator n-channel MOSFET has a p-doped 'bulk' layer above the oxide layer with a doping concentration of $N_A = 2 \cdot 10^{16}\,\text{cm}^{-3}$. The gate is heavily n-doped poly-Si. The gate oxide thickness is 5 nm.

Calculate the thickness of the doped layer such that it will be fully depleted in active device operation. What is the purpose of the buried oxide layer? Explain its effect(s) on device performance.

(7) In a HEMT, what is the purpose of the spacer layer? Would the device still function without it?

(8) A HEMT device has the following layer structure:

Function	Composition	Thickness	Doping concentration
Supply	$Al_{0.3}Ga_{0.7}As$	80 nm	$N_D = 3 \cdot 10^{17}\,\text{cm}^{-3}$
Spacer	$Al_{0.3}Ga_{0.7}As$	5 nm	Nominally undoped
Buffer	GaAs	100 nm	$N_A = 1 \cdot 10^{15}\,\text{cm}^{-13}$
Substrate	GaAs	150 μm	Intrinsic

Calculate the threshold voltage of this device at room temperature.

Let now $V_{GS} - V_{\text{off}} = 0.5$ V. Calculate the sheet density charge of the 2DEG.

(9) You want to optimise the gain and low-noise behaviour of a HEMT by changing the position of the gate electrode between source and drain. You observe the best performance if the gate is placed

a) in the middle between the contacts

b) closer to the drain contact

c) closer to the source contact

One or none of the statements is true – explain your choice!

(10) Draw the small-signal equivalent circuit of an FET for $V_{DS} = 0$, making appropriate simplifications. What is the relationship between C_{GS} and C_{GD} in this bias point? Would you expect a noticeable difference between MESFETs and HEMTs in this mode of operation?

(11) Explain the two major advantages of a pseudomorphic HEMT structure, compared to the classic AlGaAs/GaAs HEMT. How do they relate to f_T, f_max and F_min? Is there a potential disadvantage of the lower band gap in the channel?

(12) In order to reduce the series resistance of the gate, FETs (MESFETs, HEMTs and MOSFETs alike) are typically constructed with several gate fingers in parallel. Which effect(s) on device performance will result from this measure? Will this affect the transit frequency f_T? Explain your answer.

(13) A HEMT technology has $f_\mathrm{T} = 80\,\mathrm{GHz}$ and $g_\mathrm{m} = 600\,\mathrm{mS\,mm^{-1}}$. For a device with $W_\mathrm{G} = 2 \cdot 60\,\mathrm{m}$, the gate resistance is measured to be $R_\mathrm{G} = 10\,\Omega$, and the minimum noise figure at 24 GHz is $F_\mathrm{min} = 1.7\,\mathrm{dB}$. For a device with $W_\mathrm{G} = 6 \cdot 60\,\mu\mathrm{m}$, what is the expected noise figure at 30 GHz? You may neglect the source and drain resistances.

(14) Only one of the following answers is correct: in a bipolar transistor in active forward operation, the base transit time is

a) not a function of V_CE
b) a weak function of V_CE
c) a strong function of V_CE.
Explain your choice!

(15) Explain the following observations on high-speed bipolar transistors:

a) In devices optimised for record f_T, the maximum frequency of oscillation is often quite low, and the breakdown voltage is also low.
b) Devices optimised for record f_max often have low f_T, higher breakdown voltage, and need relatively high V_CE for optimum operation.
c) The latter devices suffer from significant Kirk effect.

(16) In n–p–n bipolar transistors, a drift field for electrons in the base can reduce the base transit time. This can be done in two ways:

(a) introduce a continuously varying material composition;
(b) vary the doping concentration in the base.

Explain how to achieve a constant field strength in the neutral base using either of the two methods.

(17) Note that Si/SiGe HBTs are always double-heterostructure devices. Why?

References

[1] Allam R., Kolanowski C., Theron D., Crosnier Y. (1994). Large signal model for analysis and design of HEMT gate mixer. *IEEE Microwave Guided Wave Lett. MGWL-4*, 12 (December), 405–407.

[2] Antoniadis D. A., Aberg I., NiCléirigh C., Nayfeh O. M., Khakifirooz A., Hoyt J. L. (2006). Continuous MOSFET performance increase with device scaling: the role of strain and channel material innovations. *IBM J. Res. & Dev. 50*, 4/5 (April–May), 363–377.

[3] Belache A., Vanoverschelde A., Salmer G., Wolny M. (1991). Experimental analysis of HEMT behavior under low-temperature conditions. *IEEE Trans. Electron Devices ED-38*, 1 (January), 3–13.

[4] Benkhelifa F., Chertouk M., Dammann M., Massler M., Walther H., Weimann G. (2001). High performance metamorphic HEMT with 0.25 μm refractory metal gate on 4" GaAs substrate. In *International Conference on Semiconductor Manufacturing Technology GaAs MANTECH 2001 Digest of papers*. Las Vegas, NV: GaAs MANTECH, 230–233. http://www.csmantech.org/Digests/2001/PDF/11_3_Benkhelifa_V2.pdf.

[5] Bourgoin J., Mauger A. (1988). Physical origin of the DX center. *Appl. Phys. Lett. 53*, 9 (August), 749–751.

[6] Cappy A. (1988). Noise modeling and measurement techniques. *IEEE Trans. Microw. Theory Tech. MTT-36*, 1 (January), 1–10.

[7] Chan Y.-J., Pavlidis D., Razeghi M., Omnes F. (1990). $Ga_{.51}In_{.49}P$/GaAs HEMT's exhibiting good electrical performance at cryogenic temperatures. *IEEE Trans. Electron Devices ED-37*, 10 (October), 2141–2147.

[8] Cojocaru V. I., Brazil T. J. (1997). Scalable general-purpose model for microwave FET's including DC/AC dispersion effects. *IEEE Trans. Microw. Theory Tech. MTT-45*, 12 (December), 2248–2255.

[9] Cressler J. (2003). http://extenv.jpl.nasa.gov/presentations/SiGe_HBT_BiCMOS.pdf.

[10] Curtice W. (1980). A MESFET model for use in the design of GaAs integrated circuits. *IEEE Trans. Microw. Theory Tech. MTT-28*, 5 (May), 448–456.

[11] Das M. B. (1988). High-frequency performance limitations of millimeter-wave heterojunction bipolar transistors. *IEEE Trans. Electron Devices ED-35*, 5 (May), 604–614.

[12] Delagebeaudeuf D., Chevrier I., Laviron M., Delescluse P. (1985). A new relationship between the Fukui coefficient and optimal current value for low noise operation of field effect transistors. *IEEE Electron Device Lett. EDL-6*, 9 (September), 444–445.

[13] Dingle R. (1984). New high-speed III–V devices for integrated circuits. *IEEE Trans. Electron Devices ED-31*, 11 (November), 1662–1667.

[14] Dunn J. S., Ahlgren D. C., Coolbaugh D. D., *et al.* (2003). Foundation of RF CMOS and SiGe BiCMOS technologies. *IBM J. Res.&Dev. 47*, 2/3 (February/March), 101–138.

[15] Early J. M. (1952). Effects of space-charge layer widening in junction transistors. *Proc. IRE 40*, 11 (November), 1401–1406.

[16] Ebers J. J., Moll J. L. (1954). Large-signal behavior of junction transistors. *Proc. IRE 42*, 12 (December), 1761–1772.

[17] Johnson E. O. (1965). Physical limitations on frequency and power parameters of transistors. *RCA Rev. 26*, 6 (June), 163.

[18] Feygenson A., Ritter D., Hamm R. A., *et al.* (1992). InGaAs/InP composite collector heterostructure bipolar transistors. *Electron. Lett. 28*, 7 (March), 607–609.

[19] Fiegna C. (2003). Analysis of gate shot noise in MOSFETs with ultrathin gate oxides. *IEEE Electron Device Lett. EDL-24*, 2 (February), 108–110.

[20] Folkes P. A. (1985). Thermal noise measurements in GaAs MESFETs. *IEEE Electron Device Lett. EDL-6*, 12 (December), 620–622.

[21] Fukui H. (1979). Design of microwave GaAs MESFETs for broad-band low noise amplifiers. *IEEE Trans. Microw. Theory Tech. MTT-27*, 7 (July), 643–650.

[22] Gummel H. K., Poon H. C. (1970). An integral charge-control model for bipolar transistors. *Bell Syst. Tech. J. 49*, 827–852.

[23] Hawkins R. J. (1977). Limitations of Nielsen's and related noise equations applied to microwave bipolar transistors and a new expression for the frequency and current dependent noise figure. *Solid-State Electron. 20*, 3 (March), 191–196.

[24] Hooge F. N. (1969). 1/f noise is no surface effect. *Phys. Lett. A 29*, 3 (April), 139–140.

[25] Hsia H., Tang Z., Caruth D., Becher D., Feng M. (1999). Direct ion-implanted 0.12 μm GaAs MESFET with f_t of 121 GHz and $f_{\max}$ of 160 GHz. *IEEE Electron Device Lett. 20*, 5 (May), 245–247.

[26] Joseph A., Cressler J. D., Richey D. M., Jaeger R. C., Harame D. L. (1997). Neutral base recombination and its influence on the temperature dependence of Early voltage and current gain-Early voltage product in UHV/CVD SiGe heterojunction bipolar transistors. *IEEE Trans. Electron Devices ED-44*, 3 (March), 404–413.

[27] Kallfass I. (2005a). Comprehensive Nonlinear Modelling of Dispersive Heterostructure Field Effect Transistors and their MMIC Applications. Ph.D. thesis, Ulm Universität, Ulm, Germany.

[28] Kallfass I. (2005b). Comprehensive Nonlinear Modelling of Dispersive Heterostructure Field Effect Transistors and their MMIC Applications. Ph.D. thesis, Ulm Universität, Ulm, Germany. Chapter 3.4.1.

[29] Kallfass I., Schumacher H., Brazil T. J. (2006). A unified approach to charge-conservative capacitance modelling in HEMTs. *Microwave and Wireless Components Letters 16*, 12 (December), 678–680.

[30] Kloosterman W. J. (1996). Comparison of Mextram and the VBIC95 Bipolar Transistor Model. Tech. Rep. NL-UR 034/96, Philips Electronics N. V. `http://www.nxp.com/acrobat_download/other/philipsmodels/ur034_96.pdf`.

[31] Kroemer H. (1957). Theory of a wide-gap emitter for transistors. *Proc. IRE 45*, 11 (November), 1535–1537.

[32] Ladbrooke P. H. (1985). The theory and practice of the GaAs microwave MESFET. *GEC J. Res. 3*, 191–200.

[33] Lee K., Shur M., Drummond T., Morkoc H. (1984). Parasitic MESFET in (Al,Ga)As/GaAs modulation doped FET's and MODFET characterization. *IEEE Trans. Electron Devices ED-31*, 1 (January), 29–35.

[34] Lee T. H. (2004). *The design of CMOS Radio-Frequency Integrated Circuits*. Cambridge University Press.

[35] Lilienfeld J. E. (1930). Method and apparatus for controlling electric currents. USA Patent 1,745,175.

[36] Long S. I. (1989). A comparison of the GaAs MESFET and the AlGaAs/GaAs heterojunction bipolar transistor for power microwave amplification. *IEEE Trans. Electron Devices ED-36*, 5 (May), 1274–1279.

[37] McAndrew C. C., Seitchik J. A., Bowers D. F., *et al.* (1996). VBIC95, the vertical bipolar inter-company model. *IEEE J. Solid-State Circ. 31*, 10 (October), 1476–1483.

[38] Meyer J. E. (1971). MOS models and circuit simulations. *RCA Rev. 32*, 3 (March), 42–63.

[39] Mimura T., Hiyamizu S., Fujii T., Nanbu K. (1980). A new field-effect transistor with selectively doped GaAs/n-Al$_x$Ga$_{1-x}$As heterojunctions. *Jp. J. Appl. Phys. 19*, 5 (May), L225–L227.

[40] Nakajima S., Otobe K., Kuwata N., Shiga N., Matsuzaki K., Hayashi H. (1990). Pulse-doped GaAs MESFETs with planar self-aligned gate for MMIC. *IEEE MTT-S Int. Microwave Symp. Dig. 3*, 1081–1084.

[41] Ogura S., F. Codella C., Rovedo N., Shepard J. F., Riseman J. (1982). A half-micron MOSFET using double implanted LDD. In *Int'l Electron Devices Mtg. Proceedings*, Vol. 28. Piscataway, NJ: IEEE, 718–722.

[42] Paaschens J. C. J., v. d. Toorn R., Kloosterman W. J. (1995). The Mextram Bipolar Model. Tech. Rep. NL-UR 2000/811, Philips Electronics N. V. http://www.nxp.com/acrobat_download/other/philipsmodels/nlur2000811.pdf.

[43] Pailloncy G., Iniquez B., Dambrine G., Danneville F. (2004). Influence of a tunneling gate current on the noise performance of SOI MOSFETs. In *Proceedings 2004 IEEE International SOI Conference*. Piscataway, NJ: IEEE, 55–57.

[44] Parthasarathy N., Griffith Z., Kadow C. *et al.* (2006). Collector-pedestal InGaAs/InP DHBTs fabricated in a single-growth, triple-implant process. *IEEE Electron Device Lett. EDL-27*, 5 (May), 313–316.

[45] Post I., Akbar M., Curello G., *et al.* (2006). A 65 nm CMOS SOC technology featuring strained silicon transistors for RF applications. In *Int'l Electron Devices Mtg. Proceedings*. Piscataway, NJ: IEEE, 1–3.

[46] Pucel R. A., Haus H. A., Statz H. (1975). Signal and noise properties of gallium arsenide microwave field effect transistors. In *Advances in Electronics and Electron Physics*, L. Marton, ed. Vol. 38. Academic Press, 195–265.

[47] Rodwell M. J. W., Urteaga M., Mathew T., *et al.* (2001). Submicron scaling of HBTs. *IEEE Trans. Electron Devices ED-48*, 11 (November), 2606–2624.

[48] Saito M., Ono M., Fujimoto R., *et al.* (1998). 0.15 μm RF CMOS technology compatible with logic CMOS for low-voltage operation. *IEEE Trans. Electron Devices ED-45*, 3 (March), 737–742.

[49] Sakurai T., Newton A. R. (1991). A simple MOSFET model for circuit analysis. *IEEE Trans. Electron Devices ED-38*, 4 (April), 887–894.

[50] Sato H., Vlcek J. C., Fonstad C. G., Meskoob B., Prasad S. (1990). InGaAs/InAlAs/InP collector-up microwave heterojunction bipolar transistors. *IEEE Electron Device Lett. EDL-11*, 10 (October), 457–459.

[51] Schröter M. (2002). Staying current with HICUM. *IEEE Circ. Dev. Mag. 18*, 3 (May), 16–25.

[52] Schröter M. (2007). RF Modeling of Bipolar Transistors with HICUM. http://www.iee.et.tu-dresden.de/~schroter/Conf/hic_ovw.pdf.

[53] Schröter M., Friedrich M., Rein H.-M. (1993). A generalized integral charge-control relation and its application to compact models for silicon-based HBT's. *IEEE Trans. Electron Devices ED-40*, 11 (November), 2036–2046.

[54] Schröter M., Lee T. Y. (1999). Physics-based minority charge and transit time modeling for bipolar transistors. *IEEE Trans. Electron Devices ED-46*, 2 (February), 288–300.

[55] Schüppen A., Berntgen J., Maier P., Tortschanoff M., Kraus W., Averweg M. (2001). An 80 GHz SiGe production technology. *III–V Review 14*, 6 (August), 42–46.

[56] Shockley W. (1951). Circuit element utilizing semiconductive material. USA Patent 2,569,347.

[57] Shockley W. (1952). A unipolar 'field effect' transistor. *Proc. IRE 40*, 11, 1365–1376.

[58] Statz H., Newman P., Smith I. W., Pucel R. A., Haus H. A. (1987). GaAs FET device and circuit simulation in SPICE. *IEEE Trans. Electron Devices ED-34*, 2 (February), 160–169.

[59] Stillman G., Wolfe C., Dimmock J. (1970). Hall coefficient factor for polar mode scattering in N-type GaAs. *J. Phys. Chem. Solids 31*, 6, 1199–1204.

[60] Südow M., Andersson K., Billström N., *et al.* (2006). An SiC MESFET-based MMIC process. *IEEE Trans. Microw. Theory Tech. MTT-54*, 12 (December), 4072–4078.

[61] Sugli T., Watanabe K., Sugatani S. (2003). Transistor design for 90 nm-generation and beyond. *Fujitsu Sci. Tech. J. 39*, 6 (June), 9–22.

[62] van der Ziel A. (1962). Thermal noise in field effect transistors. *Proc. IRE 50*, 8 (August), 1808–1812.

[63] van der Ziel A. (1963). Gate noise in field effect transistors at moderately high frequencies. *Proc. IEEE 51*, 3 (March), 461–467.

[64] Yngvesson S. (1991). *Microwave Semiconductor Devices*. Kluwer Academic Publishers.

[65] Zeghbroeck B. V. (2004). Principles of Semiconductor Devices. http://ece-www.colorado.edu/~bart/book/book/chapter5/pdf/ch5_5.pdf.

3 Optimisation and parameter extraction of circuit models

3.1 Executive summary

Optimised device models are important in the design of electronic devices for specific performance. They help the designer to predict the behaviour of the device in an analogue circuit. However, standard methods of optimisation do not lend themselves to fast computation and may present problems with convergence. The simulated annealing, genetic and structured genetic algorithms are alternative optimisation methods that help to solve the convergence problems by avoiding entrapment in local minima of the solution space. These methods are used for the extraction of the parameters of the equivalent circuit model of the device or to construct its neural network model. The neural network models are black box models that determine device outputs from known inputs after appropriate weighting values have been calculated. Another method of parameter extraction is the semi-analytical procedure that uses an analytic approach together with empirical optimisation methods. Basic expressions and approximations to extract small-signal equivalent circuit parameters are developed so that accurate device models can be obtained.

3.2 Optimisation of device models

Accurate device models are developed to predict behaviour that is in good agreement with experimental observations. The optimisation of the model parameters can be considered to be the curve-fitting of the computed device characteristics to experimental data. Traditional gradient methods are computationally intensive and there are problems with convergence and entrapment in local minima. Optimisation only involves the "trial" of a number of initial solutions to minimise the error.

Combinatorial optimisation overcomes this problem of entrapment in local minima. To set up the device modelling problem as a combinatorial optimisation problem, the limits on each parameter are specified. Parameter values are then discretised so that a large but finite number of solutions is possible. An objective function, also referred to as the *cost function* is given by

$$F(V_\mathrm{i}) = \sum_{j=1}^{m}(M_\mathrm{j} - M'_\mathrm{j})^2, \tag{3.1}$$

where V_i ($i = 1, 2, \ldots, n$; $n =$ number of parameters) = the model parameters to be determined; M_j = a measured characteristic for the jth data point; M'_j = a calculated characteristic which is a function of the parameters V_i for the jth data point; and m = the total number of characteristics to be fitted. Note that $F(V)$ is a least square difference function. The number of possible solutions is extremely large and an exhaustive search for the optimum solution is practically impossible. Hence, it is necessary to employ a heuristic method. Such methods are well-established in CAD tools since they enable the designer to find a feasible solution in a finite period of time. A typical heuristic optimisation process utilises an iterative improvement strategy. At each step of the iteration, the algorithm generates a new solution and tests if it reduces the value of the objective function. If so, it accepts the new solution. If not, another new solution is generated and tested and eventually the global minimum of $F(V)$ in Equation (3.1) should be reached. However, the size of the solution space, as defined by all possible and physically meaningful parameters, is very large and grows exponentially with the number of variables in the model. An exhaustive search for the best solution cannot be performed in a finite period of time.

3.3 Simulated annealing

Most heuristic algorithms search for a solution only in the directions that improve the objective function. This type of heuristic search has a major drawback: it can be easily trapped into the local minima of an objective function. Figure 3.1 demonstrates the problem.

The curve shown in Figure 3.1 is assumed to be the objective function of an iterative improvement process; the circles indicate the costs, i.e. $F(V_\mathrm{i})$ in the least squares objective function given in Equation (3.1) calculated from certain parameter sets. Since

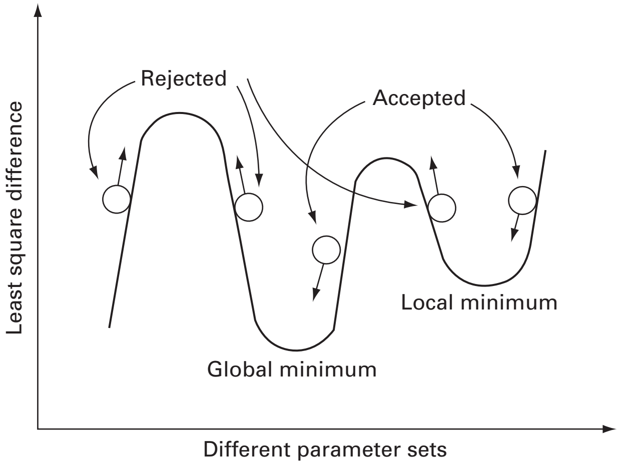


Fig. 3.1 Local minimum trapping in iterative optimisation.

a new set of parameters is generated by introducing small modifications to the model, its corresponding location on the curve is most likely to be somewhere near that of the original configuration. The traditional iterative improvement algorithms only accept parameter sets that have reduced the cost. This criterion of set acceptance implies that the process can only go downhill into a minimum and any uphill movement is forbidden. Depending on the starting point of the search, it is possible that this minimum is only a local minimum of the objective function. This search process generally does not have the capability of climbing over a peak of the curve to reach the global minimum.

The drawback of this process can be overcome by using simulated annealing (SA)[12, 22, 23, 30]. This is a method of finding a near optimal solution for combinatorial optimisation problems. The SA algorithm has the advantage of asymptotically producing the global optimal solution with a probability of one. This is achieved by making the following important modification to conventional heuristic methods. A cost-increasing solution may still be accepted. The probability of acceptance depends on: (1) a parameter called a *pseudo-temperature* T_{k}, which is artificially decreased as the iteration proceeds, and (2) the value of

$$\Delta F(V) = F(V_{\mathrm{k}}) - F(V_{\mathrm{k-i}}) \tag{3.2}$$

where V_{k} and $V_{\mathrm{k-1}}$ are the values of the parameter vector V at iteration steps k and $k-1$. Specifically, a law similar to Boltzmann statistics is used to determine the probability P of accepting a certain cost-increasing solution V_{k} at the kth iteration step. The probability function P is given by

$$P(V_{\mathrm{k}}, T_{\mathrm{k}}) = \exp\left[\frac{-\Delta F(V_{\mathrm{k}})}{T_{\mathrm{k}}}\right]. \tag{3.3}$$

A careful choice of the initial pseudo-temperature T_0 and a rule for decreasing the pseudo-temperature T are necessary to save computation time while being able to escape from the local minima. The temperatures are related by the equation

$$T_{\mathrm{k}} = \alpha T_{\mathrm{k-1}}, \tag{3.4}$$

where α is a constant between 0.8 and 0.95. Its value can be gradually increased from the lowest to the highest value. $T_0 \geq 500$ is a satisfactory choice for device modelling.

At each iteration, new parameter values are generated by first choosing one device parameter V_{i} at random. A user-defined base value $V_{\mathrm{i-base}}$ is multiplied by a random number R, such that $0 \leq R \leq 1$, and a variation $\Delta V_{\mathrm{i}} = RV_{\mathrm{i-base}}$ is introduced into the parameter V_{i}. The new parameter value so obtained is used in the next iteration unless it exceeds prescribed limits, in which case it will be set to the maximum or minimum allowable value.

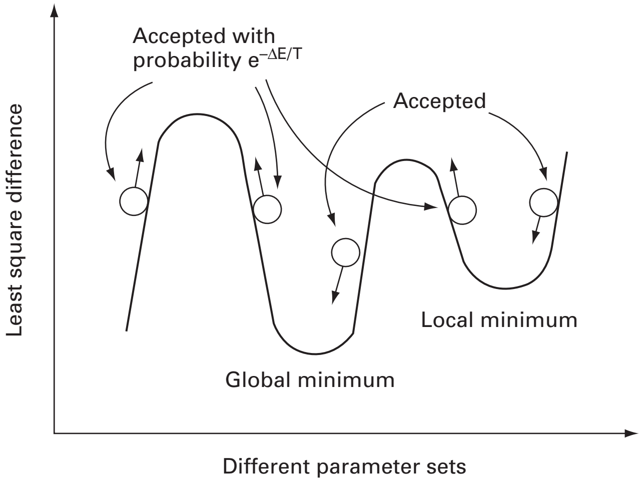


Fig. 3.2 Hill-climbing capability of SA.

A relative stopping criterion is used, since there is no guarantee that the device model can approximate experimental data closely. The optimisation process is stopped when the value of the objective function has remained virtually unchanged for several consecutive iterations (e.g. $\Delta F \leq 0.001$ for 10 consecutive iterations).

Figure 3.2 shows the same objective function as in Figure 3.1 with the hill-climbing capability of SA.

Example: Application of SA to the modelling of a HEMT

Optimisation by SA has been applied to the HEMT [30]. Three test cases are given as examples to demonstrate the optimisation process and evaluate its performance. Parameters are allowed to vary within ±90% of their initial values in test cases A and B. In test case C, the limits are reduced to ±20% for realistic parameters.

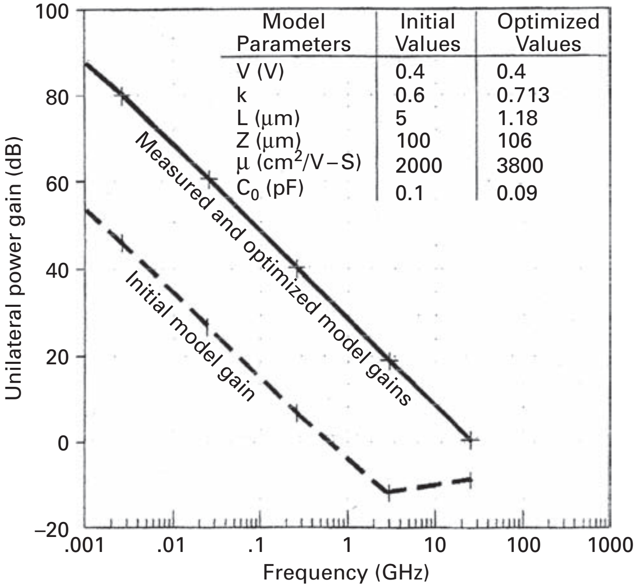
Test case A: AC Model of a HEMT

The unilateral power gain U of the HEMT (also known as the MODFET as explained in Chapter 2) was determined by Roblin *et al.* [21]. The values of U are sampled at various frequencies and used as the measured characteristic M_j to be matched to the model. The Mason unilateral power gain (defined in Chapter 5) is expressed in terms of the admittance (Y) parameters as

$$U = \frac{|\ (Y_{21} - Y_{12})\ |^2}{4[\mathrm{Re}(Y_{11})\mathrm{Re}(Y_{22}) - \mathrm{Re}(Y_{12})\mathrm{Re}(Y_{21})]}. \tag{3.5}$$

The Y-parameters are determined by primitive model parameters, i.e. the gate capacitance C_0, the gate length L, the gate width Z, the bias voltage $V = V_{\mathrm{GS}} - V_{\mathrm{T}}$ and the parameter k which is given by

$$K = \frac{V_{\mathrm{DS}}}{(V_{\mathrm{GS}} - V_{\mathrm{T}})}, \tag{3.6}$$



Model Parameters	Initial Values	Optimized Values
V (V)	0.4	0.4
k	0.6	0.713
L (μm)	5	1.18
Z (μm)	100	106
μ (cm^2/V−S)	2000	3800
C_0 (pF)	0.1	0.09

Fig. 3.3 Test case A: unilateral power gain curves and optimisation results [+, measured; solid line, optimised model; dotted line, initial solution] (M-K. Vai, S. Prasad, N. C. Li and F. Kai, *IEEE Transactions on Electron Devices*, Vol. 36, No. 4, pp. 761–762, April 1989. ©1989 IEEE).

where V_{DS} is the drain to source voltage, V_{GS} is the gate to source voltage and V_T is the threshold voltage.

Figure 3.3 compares the measured gain curve with curves generated from an initial model and the final optimised models. The objective function is reduced from an initial value of 4421.83 to 0.161095, virtually a perfect match.

Test case B: Equivalent circuit of a HEMT

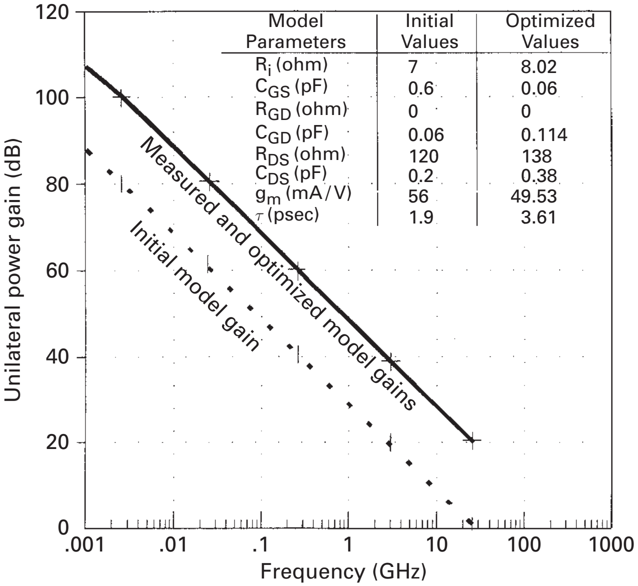
The small-signal equivalent circuit of an intrinsic HEMT is used to deduce the unilateral power gain using Equation (3.5). The Y-parameters are expressed in terms of the circuit elements R_i, C_{GS}, R_{DS}, C_{DS}, g_m and τ. The objective function is reduced from 1783.78 to 0.01789 when optimisation is completed. Figure 3.4 shows the agreement between measured and computed values using the optimised model.

Test case C: Device design parameters

In this test, a set of primitive device parameters (gate length, gate width, bias voltage and mobility) are obtained by optimisation such that the device has the highest value of f_{max}, the maximum frequency of oscillation. The analytical expression for f_{max} has been obtained by Roblin *et al.* [21] by setting the unilateral gain equal to 1. The primitive device parameters are related to f_{max} and the objective function is formulated. The initial design gives a value of f_{max} equal to 29.57 GHz. The optimised HEMT has f_{max} equal to 165.81 GHz. Such an optimisation is an aid to the device designer. Table 3.1 gives the optimised values. The initial value of K is 0.6 and the optimised value is 0.95.

Table 3.1 Test case C: optimised design parameters for the highest $f_{\max}$

Model parameters	Initial values	Optimised values
V_T(V)	0.4	0.4
L(μm)	1	0.8
Z(μm)	250	300
Mobility μ(cm^2)/(V-s)	4000	4800
C_0(pF)	0.1	0.108



Model Parameters	Initial Values	Optimized Values
R_i (ohm)	7	8.02
C_{GS} (pF)	0.6	0.06
R_{GD} (ohm)	0	0
C_{GD} (pF)	0.06	0.114
R_{DS} (ohm)	120	138
C_{DS} (pF)	0.2	0.38
g_m (mA/V)	56	49.53
τ (psec)	1.9	3.61

Fig. 3.4 Test case B: unilateral power gain curves and optimisation results [+, measured; solid line, optimised model; dotted line, initial solution] (M-K. Vai, S. Prasad, N. C. Li and F. Kai, *IEEE Transactions on Electron Devices*, Vol. 36, No. 4, pp. 761–762, April 1989 © IEEE).

3.4 Neural networks applied to modelling

The SA algorithm has been applied to device modelling in the previous section. SA avoids the local minimum entrapment problem and has been shown to be preferable to other optimisation methods since it is relatively insensitive to initial conditions. However, it inherits the time-consuming feature of iterative improvement methods. As was shown, the probabilistic hill-climbing operation increases the time taken for the completion of the optimisation. Furthermore, the solution has to slowly "cool down" according to the annealing procedure. Consequently, a large number of intermediate solutions have to be generated and evaluated at each pseudo-temperature which has to be decreased slowly from a large initial value. This disadvantage of the SA algorithm for optimisation can be overcome by using an artificial neural network (ANN).

ANNs are based on the human nervous system which consists of a distribution of neurons to carry messages back and forth to the brain. The ANN has been used successfully in many modelling and optimisation applications in engineering that are particularly useful when several tasks are to be performed in parallel and computation rates are required to be high [31, 37]. Neural networks are particularly attractive because of their speed and accuracy. Hence, neural networks have been developed into an alternative computer-aided approach to model and design devices and circuits. Neural networks represent a robust modelling approach to predict the behaviour of high-speed devices and circuits [29]. In comparison with various statistical methods and curve-fitting approaches for predicting system behaviour, the neural network approach features a learning process which fine tunes neural network parameters to interrelate the variables being modelled. A neural network may be developed to guide the solution generation of an SA optimisation process. This approach utilises the associative capability of a neural network to globally and concurrently evaluate the effect of varying all the parameters. When used in place of a physics-oriented device model, a neural network avoids the need to repeatedly solve the equations that describe the device physics.

Two classes of neural networks are used in modelling:

(i) Multi-layer perceptron neural networks
(ii) Recurrent Hopfield neural networks.

Regardless of the neural network architecture selected for an application, it consists of many processing elements called *neurons*, each connected to many others. Every connection entering a neuron has a weight assigned to it. This weight is used to amplify, attenuate or change the sign of the signal in the incoming connection. The input to the neural network is a vector of the data to be modelled. Each neuron operates on the outputs of other neurons according to its transfer function and delivers an output to other neurons. Often, the transfer function sums the incoming signals to determine the value of the neuron's next output signal. The result is an output vector representing some characteristics associated with the input.

In order to use neural network algorithms, it is necessary to determine an interconnection pattern, the weights and the transfer functions. The creation and training of an appropriate neural network for the problem on hand is difficult and time-consuming. However, an appropriately trained neural network provides fast and efficient solutions that have shown excellent results for different applications such as the modelling of transistor behaviour and microwave circuits as well as microwave impedance matching [28]. A neural network consists of a set of simultaneous non-linear equations that are capable of modelling any continuous function when the appropriate weights are determined. These networks are pictorially represented as neurons (circles) with interconnecting nerves (lines). Neural networks are very flexible tools for device modelling because they can be adapted to model different devices without change of equations. Adapting a physics-based model to a different physical device may involve a radical change of equations. This flexibility gave the impetus to the effort of finding much better ways of applying neural networks to device modelling. However, the problem of finding the appropriate weights for the network is a highly non-linear one, suggesting the necessity to use stochastic optimising algorithms such as SA and genetic algorithms.

3.4.1 Massively distributed computing networks

A general description of the distributed computing methodology is given by Vai and Prasad [31]. Massively distributed computing networks are a specific form of a non-linear system that maps an input to an output. A distributed computing network can be considered to be an asynchronous array processor with very simple processing elements (i.e. neurons). Figure 3.5 shows a typical processing element, henceforth referred to as a *neuron*, with n inputs ($i_1, \ldots, i_n$) and one output (Q). An input can be excitatory (indicated by a solid circle) or inhibitory (indicated by a hollow circle) and is assigned a weighting factor W_j. A threshold value T is associated with the neuron.

The function of a neuron can be described by the following equation which combines inputs $i_1, \ldots, i_n$ to form an overall input value I:

$$Q = \sum_{j=1}^{n} W_j \times i_j, \tag{3.7}$$

where W_j is positive for an excitatory input and negative for an inhibitory input. If the overall input value I is above the threshold value T associated with the neuron, the neuron fires and an output of $Q = 1$ is produced. Otherwise, the output remains at $Q = 0$. A neuron is also associated with a time constant (τ) that determines its output response time.

While the operation of a conventional computer is controlled by a series of instructions, called a *programme*, a massively distributed computing network is programmed by wiring up a set of neurons and setting the weights of these interconnections. The function of a distributed computing network can only be determined by considering the network as an integrated entity. No meaningful information can be extracted by examining a neuron isolated from its neighbours.

A distributed computing network is typically implemented by a hardware analogue circuit. Figure 3.6 shows the use of an operational amplifier configured as an integrating adder to carry out the function of a neuron. As shown in Figure 3.6, the input

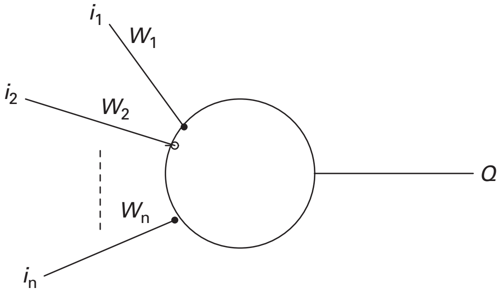


Fig. 3.5 Neuron structure.

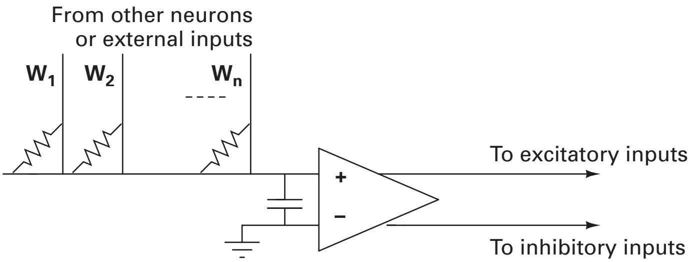


Fig. 3.6 Physical implementation of a neuron (M. Vai and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 43, No. 5, pp. 1087–1094, May 1995. ©1995 IEEE).

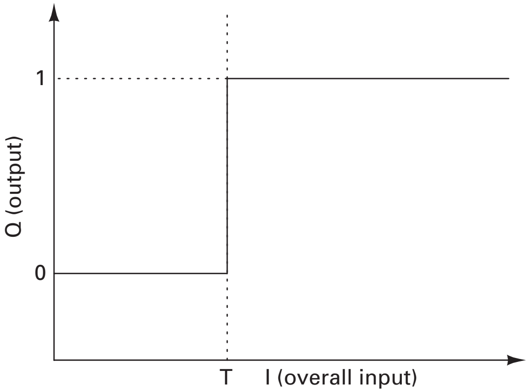


Fig. 3.7 Neuron Transfer Function (M. Vai and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 43, No. 5, pp. 1087–1094, May 1995. ©1995 IEEE).

weightings of such a neuron can be controlled by choosing appropriate resistance values connecting its inputs to the outputs of other neurons. The time constant (τ) of this neuron is determined by the capacitance connected at the operational amplifier input. The neuron transfer function shown in Figure 3.7 shows the input–output transfer function of a neuron.

3.4.2 Multi-layer perceptron neural networks

The training algorithm called *backpropagation* [16] is used in the application of multilayer perceptron (MLP) neural networks to device modelling. A multi-layer neural network with four layers (one input layer, two hidden layers and one output layer) is shown in Figure 3.8. Referring to the notations in Figure 3.8, $X = (x_1, \ldots, x_i, \ldots, x_m)$ is the input vector; $G = (g_1, \ldots, g_j, \ldots, g_n)$, $H = (h_1, \ldots, h_k, \ldots, h_p)$ and $Y = (y_1, \ldots, y_l, \ldots, y_q)$ are the outputs of the first hidden layer, the second hidden layer and the output layer, respectively; u_{ij} is the weight between the ith neuron and the jth neuron in the first hidden layer; v_{jk} is the weight between the jth neuron in the first hidden layer and the kth neuron in the second hidden layer; and w_{kl} is the weight between the kth neuron in the second hidden layer and the lth neuron in the output layer. Bias

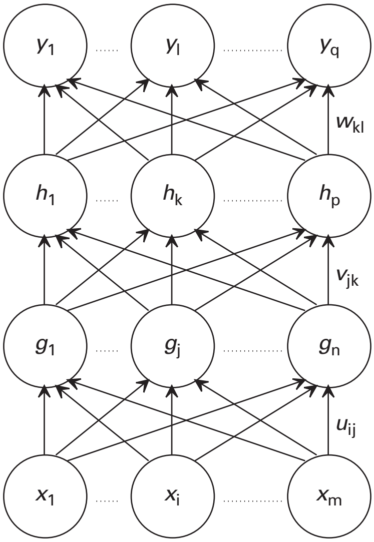


Fig. 3.8 Multi-layer neural network (M. Vai and S Prasad, *Int'l Journal of RF and Microwave CAE*, Vol. 9, No. 3, pp. 187–197, March 1999. © 1999 John Wiley & Sons). Reprinted with permission of John Wiley & Sons, Inc.

terms acting like weights on connections from units whose output is always 1 can also be provided to the neuron. They are not shown in Figure 3.8.

The output of the neural network is computed as

$$y_\ell = \frac{1}{1+e^{-\gamma_\ell}}, \tag{3.8}$$

where γ_ℓ is the weighted total input to the output neuron ℓ, which is defined as

$$\gamma_\ell = \sum_{k=1}^{p} h_k w_{k\ell}, \tag{3.9}$$

and p is the number of neurons in the second hidden layer. Similarly, the output of the second hidden layer H can be expressed as a function of the output of the first hidden layer G which can in turn be expressed as a function of the input vector X. The backpropagation training algorithm aims to adjust the weights of a MLP neural network in order to minimise the sum-squared error of the network, which is defined as

$$E(n) = \sum_{m=1}^{S} \left\{ \frac{1}{2} \sum_{l=1}^{q} [d_{ml} - y_{ml}(n)]^2 \right\}, \tag{3.10}$$

where n is the epoch number in the training process, S is the number of training data, q is the number of output variables and $d_m = (d_{m1} d_{m2}, \ldots, d_{mq})$ and $y_m = (y_{m1} y_{m2}, \ldots, y_{mq})$ are the mth desired and calculated output vectors, respectively. This is done by continually changing the values of the weights in the direction of steepest descent with respect to the error function E. The iteration process continues until the error function is minimised. The learning is performed by the many presentations of a prescribed set of training examples to the network. One complete presentation of the entire training set during the learning process is called an *epoch*. The learning process continues on an epoch-by-epoch basis until the weights of the network stabilise and the error function converges to a minimum value.

There are certain problems related to the architecture of an MLP neural network such as the determination of the number of hidden layers and the number of neurons in a hidden layer as well as under-fitting or over-fitting. The backpropagation learning algorithm and its derivatives are sensitive to the number of neurons in hidden layers. In general, a network with too few neurons will fail to model the data (i.e. under-fitting). While the more the number of neurons in hidden layers, the better the network can fit the data; if far too many neurons are used, over-fitting can occur. In the absence of a deterministic approach that can find the number of hidden layers and the number of neurons, a trial and error approach is taken. The hidden layers are adjusted to strike a balance between memorisation and generalisation.

A neural network trained with the relations between device parameters and behaviour can be used in place of conventional device models to speed up the simulation. Once a neural network model is trained, it provides a very fast prediction of results. Figure 3.9 shows a simplified flow chart of the circuit design process. Beginning with an initial solution, a series of solutions is generated. The circuit property of each solution is

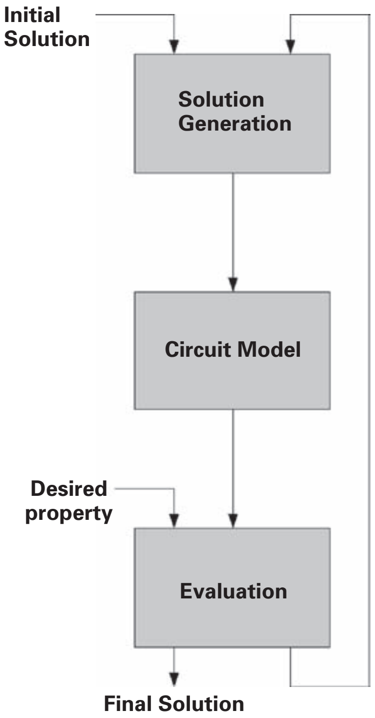


Fig. 3.9 Flow chart for a circuit design process (M. Vai and S. Prasad, *Int'l Journal of RF and Microwave CAE*, Vol. 9, No. 3, pp. 187–197, March 1999. ©1999 John Wiley & Sons). Reprinted with permission of John Wiley & Sons, Inc.

predicted by a circuit model and compared to the desired circuit property. If the solution on hand produces a circuit property close enough to the desired one, the design process is successful and terminated. Otherwise, another solution is generated and the above steps are repeated. The circuit model in Figure 3.9 often includes semiconductor devices which are commonly represented by the equations describing the physics of the particular device or equivalent circuit.

3.4.3 Hopfield recurrent neural networks

Although neural networks are known for their capability of learning the solutions to the problems that they are designed to solve, they also provide a framework for constructing special computing architectures to solve specific problems. The recurrent neural networks described here were proposed by Hopfield and are thus often referred to as *Hopfield networks* [11]. Consider a recurrent neural network of N neurons. If the activation of a neuron is updated according to the equation:

$$V_{\mathrm{i}}(t+1) = \mathrm{sgn}\left(\sum_{j=1}^{N} T_{\mathrm{ij}} V_{\mathrm{j}}(t) + I_{\mathrm{i}}\right), \tag{3.11}$$

where $V_{\mathrm{i}}(t) \in (0, 1)$ is the state of neuron i at moment t, T_{ij} is the weight associated with the link between neurons i and j, I_{i} is the internal threshold parameter of neuron i and $\mathrm{sgn}(x)$ is defined as

$$\mathrm{sgn}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{3.12}$$

It can be shown that an energy function defined as

$$\mathrm{Energy} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} T_{\mathrm{ij}} V_{\mathrm{i}} V_{\mathrm{j}} - \sum_{i=1}^{N} I_{\mathrm{i}} V_{\mathrm{i}} + K, \tag{3.13}$$

where K is a constant, is minimised.

The significance of a recurrent neural network is its ability to perform associative inference. There is no specific distinction between input and output vectors and a recurrent network perturbed by changing one or more neuron states will evolve into one of its consistent states which are the minima of its energy (Equation 3.13). The recurrent neural network is now applied to implement qualitative models, the function of which is to explore many competing hypotheses in a solution space with constraints. The Neural Network toolbox in MATLAB is available to allow fast and efficient device modelling.

Example: Modelling of HBTs using neural networks

Small-signal models of microwave devices are very useful in circuit design. Neural networks can be trained to learn the non-linear relationship between the small-signal transistor behaviour and the device input bias conditions. Figure 3.10 shows a representation of a neural network small-signal model of an HBT. The inputs of this model are the bias conditions of the device, i.e. the base current I_{b}, the collector–emitter voltage

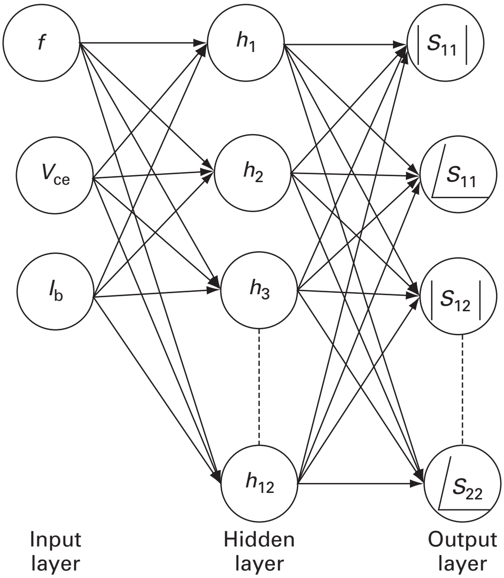


Fig. 3.10 Small-signal neural network model of the HBT.

V_{ce} and the frequency of operation f. The outputs of the model are the S parameters of the device.

As shown in Figure 3.10, the small-signal neural network model is a three-layer model – there are 3 layers in the input layer, 12 neurons in the hidden layer and 8 neurons in the output layer. The 3–12–8 structure of the neural network model is arrived at after various trials as being the best compromise between model accuracy and the time required to train the model. The model is trained with a sample set of measured S-parameter data. Different input bias conditions are then applied to the model and S-parameters derived from the model are compared to measured data. The data used to train the neural network model consist of measured S-parameters at 12 different bias conditions (V_{be} and $I_{c)}$, and a range of frequencies from 10 GHz to 40 GHz. The neural network modelling is tested for AlGaAs HBTs as well as SiGe HBTs. The models are accurate for both the material systems. The method should be applicable for all types of transistors regardless of the material system. The S-parameters for the devices are shown in Figures 3.11 and 3.12 for AlGaAs with bias $V_{ce} = 1.5\,\text{V}$, $I_b = 56\,\mu\text{A}$ and in Figures 3.13 and 3.14 for SiGe with bias $V_{ce} = 1.9\,\text{V}$, $I_b = 53\,\mu\text{A}$.

Figure 3.15 shows a representation of a neural network large-signal model of the device. The inputs of this model are the voltage bias conditions of the device: the collector–emitter voltage V_{ce} and the base–emitter voltage V_{be}. The outputs of the model are the two output currents of the device: the collector current I_c and the base current I_b.

As seen in Figure 3.15, the large-signal neural network model consists of three layers: an input layer consisting of two neurons, a hidden layer consisting of three neurons and an output layer of two neurons. The 2–3–2 structure of the neural network model was arrived at after various trials. It is the best compromise between model accuracy and the time required to train the model. The model is trained with a sample set of measured DC I–V characteristics and Gummel data. Different input voltage bias conditions are

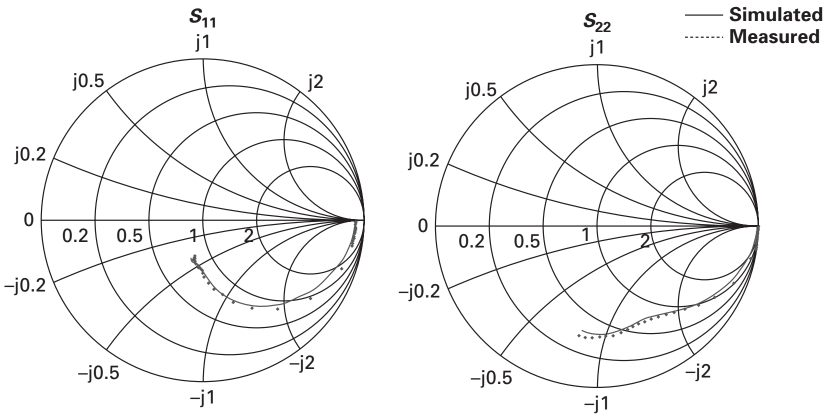


Fig. 3.11 The simulated and measured S_{11} and S_{22} for AlGaAs HBT.

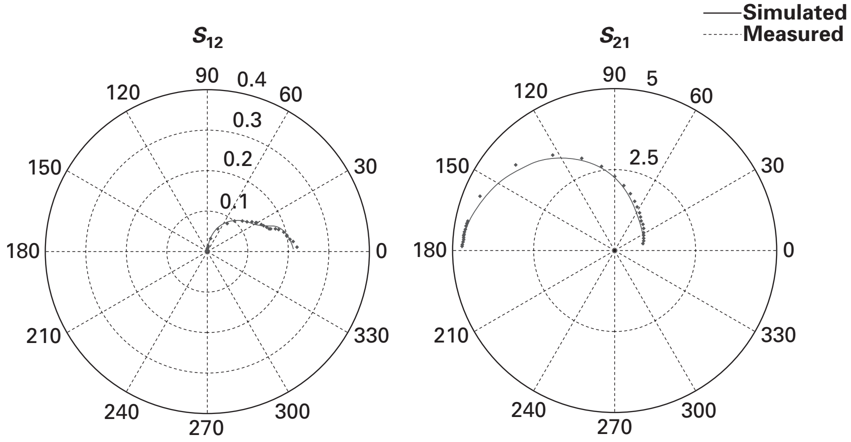


Fig. 3.12 The simulated and measured S_{12} and S_{21} for AlGaAs HBT.

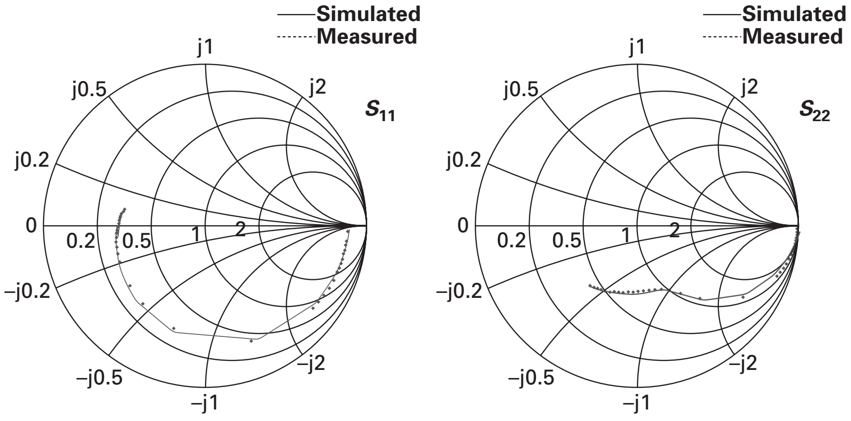


Fig. 3.13 The simulated and measured S_{11} and S_{22} for SiGe HBT.

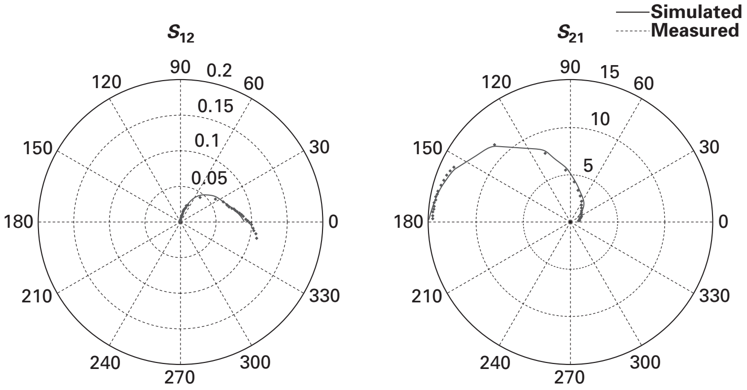


Fig. 3.14 The simulated and measured S_{12} and S_{21} for SiGe HBT.

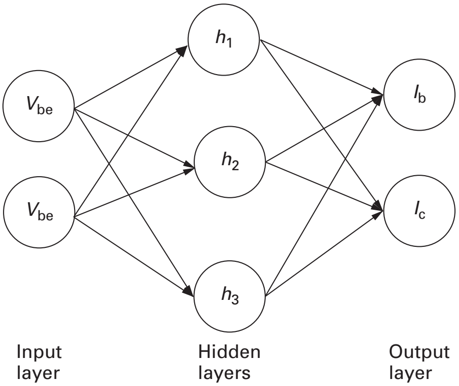


Fig. 3.15 Large-signal neural network model of the HBT.

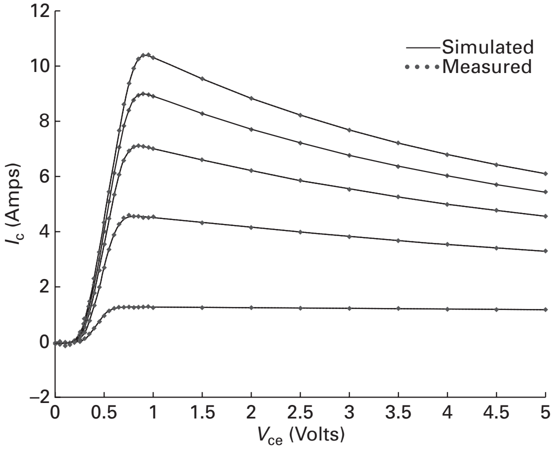


Fig. 3.16 I–V Characteristics of the AlGaAs HBT.

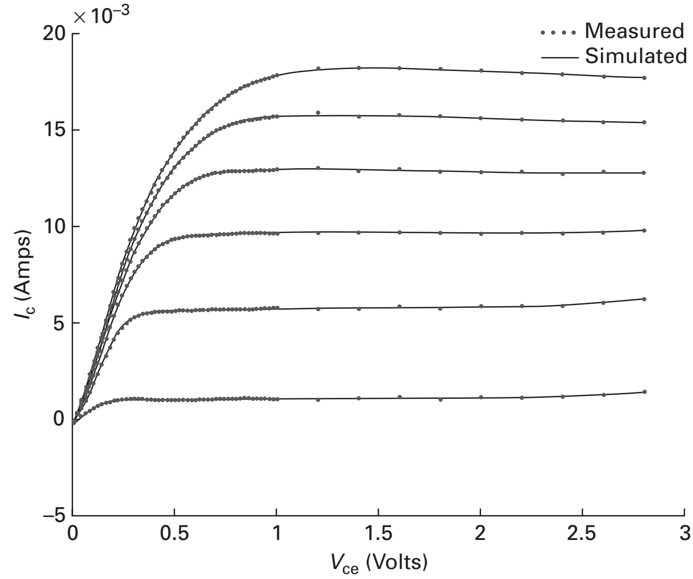


Fig. 3.17 I–V Characteristics of the SiGe HBT.

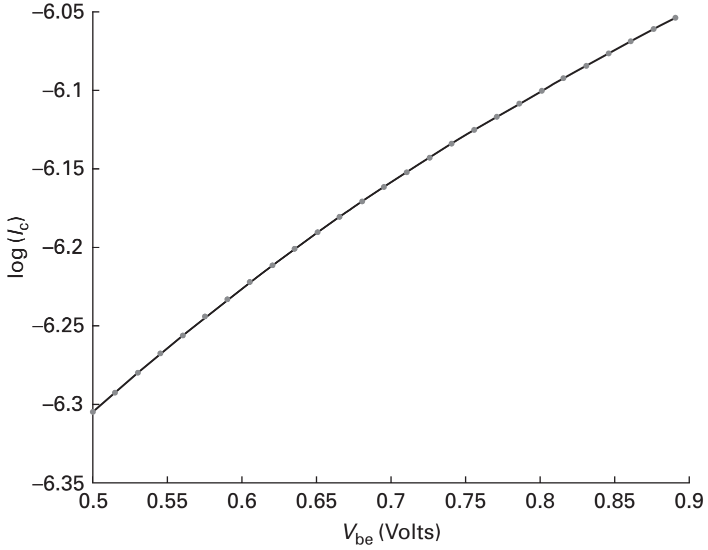


Fig. 3.18 Forward Gummel plot I_c versus V_{be} for AlGaAs HBT.

then applied to the model and output currents derived from the model are compared to measured data. The data used to train the neural network model consist of measured output currents (I_c and I_b) at different voltage bias conditions (V_{be} and V_{ce}). The DC I–V characteristics (measured and simulated) of the devices are shown in Figures 3.16 and 3.17. These figures show the collector current I_c versus the collector–emitter voltage V_{ce} at different base currents I_b.

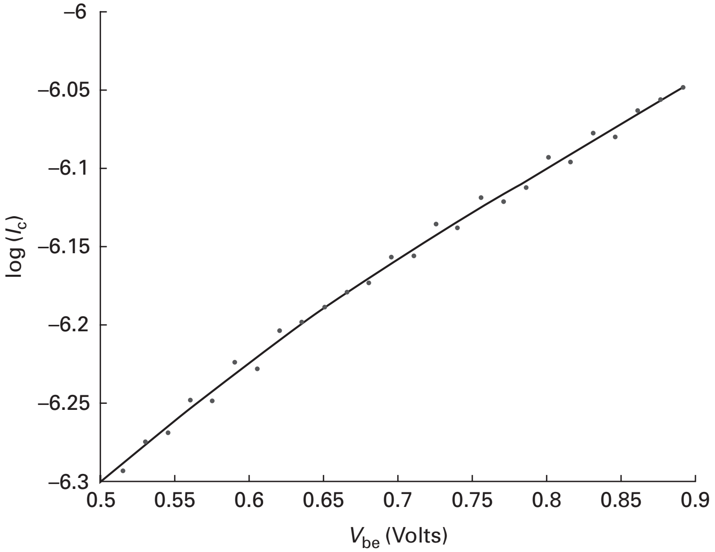


Fig. 3.19 Forward Gummel plot I_b versus V_{be} for AlGaAs HBT.

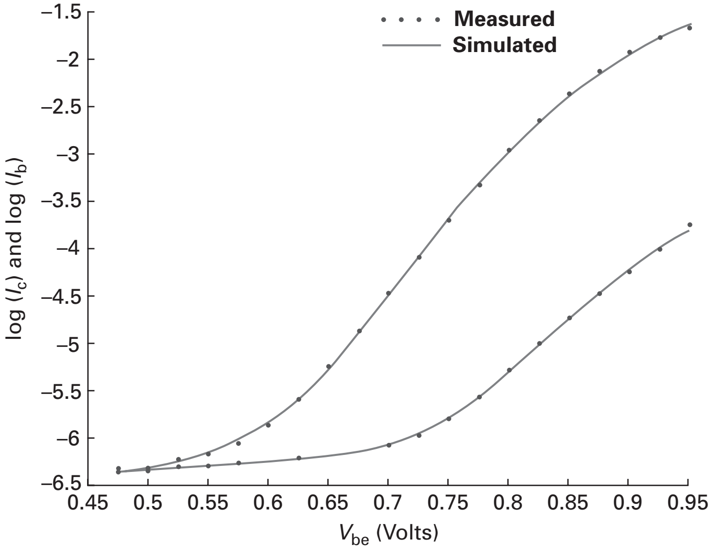


Fig. 3.20 Forward Gummel plot I_c (top curve) and I_b (bottom curve) versus V_{be} for SiGe HBT.

Figures 3.18, 3.19 and 3.20 show the forward Gummel plot of I_c and I_b versus V_{be}. The success of the application of neural network modelling in the example is evident from the very good correlation between the measured and simulated data.

3.5 Optimisation of neural networks by the genetic algorithm

Like neural networks, genetic algorithms are an optimisation strategy inspired by nature. Based on the Darwinian theory of evolution, these algorithms use the "survival of the fittest" paradigm to find the best solution to a problem [8, 10, 32]. They iteratively evaluate several possible solutions choosing the ones that are the closest fit to the desired solution. The possible solutions are called *chromosomes* and are usually represented as strings of binary numbers called *genes*.

The algorithm begins by randomly generating a number of chromosomes to form a population. Each chromosome is given a rank called a *fitness index* based on its closeness to the desired solutions. The highest ranked chromosomes have a greater chance of being selected for the next stage of the algorithm: the reproduction stage. In this stage, pairs of chromosomes are separated at selected points and their genes are exchanged in a process called *crossover* to form a new generation of chromosomes. Because the most fit chromosomes are likely to be selected for reproduction in every generation, each new population is likely to consist of better solutions. Coupled with the fact that the reproduction process naturally eliminates less fit chromosomes from the population, the population gets pushed towards the desired solution in every new generation. When trying to reach the optimal solution for the problem, a genetic algorithm has to avoid local extrema or pseudo-optimal solutions. To prevent convergence of the algorithm to these local extrema, the genetic algorithm uses a technique called *mutation* which inverts the value of a randomly chosen gene with a given probability. A flow chart for the genetic algorithm is given in Figure 3.21.

The genetic algorithm is a robust optimisation method well-suited to the difficult task of finding the optimum weights for a neural network. However, before it can be used to evolve the weights of a neural network, the following prerequisites must be worked out:

(i) *A suitable problem representation*
A chromosome of a neural network corresponds to the weight matrix for a layer. Substrings of the chromosome are made by concatenating the values of the real-valued weights between each neuron and the neurons in the previous layer. The substrings are then joined together into the chromosome for the layer. In some implementations, the real-valued weights are converted to binary values for enhanced gene exchange during crossover. However, it is sufficient to use strings of real-valued weights in most cases for genetic evolution.

(ii) *A fitness index*
A fitness index measures the closeness of the current output of the neural network to the desired output. It is the least squares norm of the difference.

(iii) *The reproduction strategy*
Reproduction involves two processes: crossover and mutation. During crossover, a pair of chromosomes are separated at a randomly chosen point along their length. The resulting substrings of the pair are switched and then joined to form child chromosomes. This is repeated for every pair of chromosomes in the population

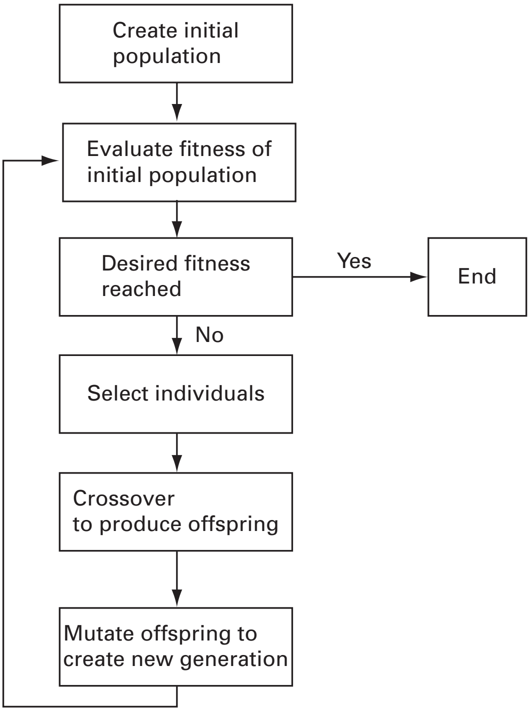

Create initial population
Evaluate fitness of initial population
Desired fitness reached
Yes
End
No
Select individuals
Crossover to produce offspring
Mutate offspring to create new generation

Fig. 3.21 Flow chart for the application of the genetic algorithm.

although some implementations will allow a small percentage of the fittest individuals to pass to the next generation without reproduction. Mutation is applied to a small percentage of the population. In this operation, the value of a gene is changed by adding a small value to it.

(iv) *The termination criterion*

Although, it would seem desirable to train the neural network till the error is as small as possible, this may be neither practical nor ideal. It may take up to several days to train a neural network. In addition, an excessively trained network can fail to produce correct outputs when presented with inputs not included in the training set. This is called *over-fitting*. The appropriate termination criterion for a given application is best determined from experiments. A common way is to terminate the algorithm after a fixed number of iterations.

3.6 Structured genetic algorithm

Neural networks are usually designed by determining the optimum values of weights for a fixed number of neurons. However, it is impossible to know in advance how many neurons are optimal for an application. This difficulty can be avoided by using the structured genetic algorithm (SGA), an algorithm that enables the determination of the optimum number of neurons and weight values simultaneously [4]. SGA uses a hierarchical representation for the genetic structure in which neurons are a high level layer of binary-valued genes controlling a lower level layer of weight genes. When a high level gene is on (value = 1), the lower level weights genes it controls are activated and used in the computation of the neural network's output. When a high level gene is off, the weights it controls are deactivated. Therefore, SGA is able to evaluate a variable number of neurons in the neural network. Neurons are turned on or off by the processes of mutation or crossover. The equations for SGA applied to a neural network with one hidden layer can be written as follows. Let X be a set of inputs to the networks and U the set of inner layer weights, then the output of a neuron in the middle layer is given by

$$y = \delta f\left(\sum_{i=1}^{m} u_{\mathrm{i}} x_{\mathrm{i}} + \tau\right), \tag{3.14}$$

where $f(x)$ is the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-\mathrm{x}}} \tag{3.15}$$

and δ is one when the neuron is active and zero otherwise.

If the set of the output weights is taken to be V, then the network outputs are given [36] by

$$z_{\mathrm{i}} = \sum_{i=1}^{m} v_{\mathrm{i}} y_{\mathrm{i}}. \tag{3.16}$$

Genetic algorithms are relatively slow algorithms and are best used to determine the weights for just the hidden layers of a neural network. The determined inner weights are then used to calculate the activations of the hidden layer neurons and this reduces the problem to a linear equation, where linear regression can be used to calculate the output layer weights.

The SGA for neural networks is given in Figure 3.22.

Example: Application of SGA to the modelling of a HEMT amplifier

The use of the SGA is illustrated by creating a neural network model of a HEMT class-F power amplifier as shown in Figure 3.23. IMN and OMN denote the input matching network and output matching network. A class-F amplifier is a highly efficient switching amplifier used in mobile commercial and military systems. The design and properties of amplifiers are described in Chapter 5.

Table 3.2 shows the inputs and outputs for one hidden layer neural network model. The upper limit for the number of neurons in the hidden layer was arbitrarily set to 40. Before training, the data were scaled with a linear transform to lie in the range of the sigmoid activation function (0.1–0.9) used for the neurons. At the start of training, the network's weights were initialised to small random values. The input to the SGA is shown in Table 3.3. Using the crossover and mutation operations, the initial weights were optimised until the termination condition was satisfied. The transfer function of the trained neural network and that of the HEMT are plotted in Figure 3.24.

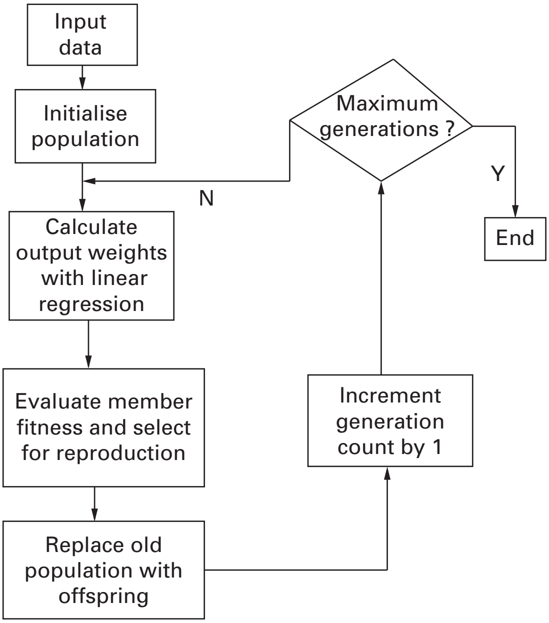

Input data
Initialise population
Calculate output weights with linear regression
Evaluate member fitness and select for reproduction
Replace old population with offspring
Increment generation count by 1
Maximum generations ?
N
Y
End

Fig. 3.22 Structured genetic algorithm for neural networks.

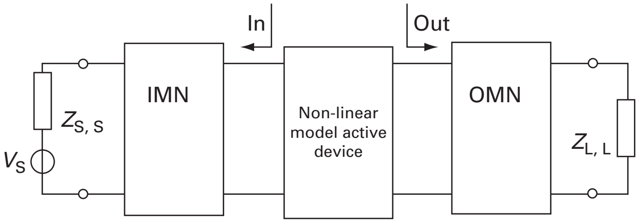


Fig. 3.23 Simplified power amplifier schematic.

Table 3.2 Neural network model for HEMT power amplifier

INPUTS	OUTPUTS
Input power	Output power
Gate voltage	DC gate current
DC drain current	Drain current

Table 3.3 SGA parameters for HEMT power amplifier model

Parameter	Value
Initial population size	200
Number of generations	10000
Number of input neurons	3
Number of output neurons	1
Maximum hidden neurons	40
Number of samples	140
Crossover probability	0.8%
Neuron layer mutation rate	0.0001%
Weights layer mutation rate	0.0001%

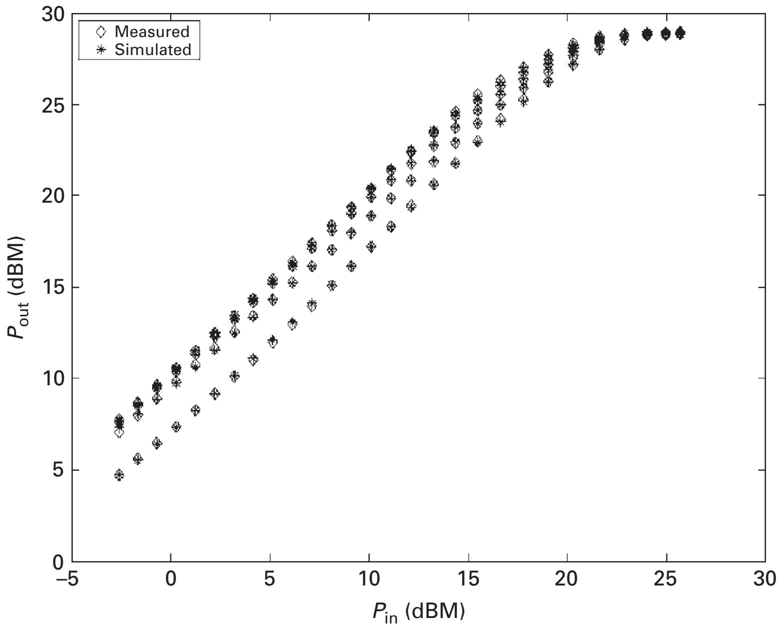


Fig. 3.24 Comparison of measured transfer function to values predicted by the neural network for the HEMT amplifier.

3.7 Semi-analytical device parameter extraction

Accurate, physically meaningful device models are necessary for circuit design, process technology design and optimum device design. The extraction of equivalent circuit parameters has been investigated by researchers for more than a decade [1–3, 5–7, 20, 25–27, 33–35].

A semi-analytical parameter extraction procedure for the HBT equivalent circuit developed by Li and Prasad [15] is presented here as one illustration of parameter extraction. It combines analytical and optimisation approaches. The significance of this procedure is that it is completely general and can be applied to any semiconductor device using the appropriate equivalent circuit.

3.7.1 Theoretical analysis

An AlGaAs common–emitter HBT is used to illustrate this procedure. The small-signal T-model equivalent circuit is shown in Figure 3.25. The box with dashed lines in Figure 3.25 encloses the inner shell without the pad parasitics. All the impedance parameters given below are for the inner shell. There are 16 elements in the equivalent circuit. Only C_{be}, r_e, α, R_{bc} and C_{bc} are considered to be bias-dependent and all the other elements are assumed to be bias-independent. The expressions for the two-port Z-parameters of the inner shell can be simplified in terms of the frequency ranges. The frequency ranges are characterised by $\omega C_{bc} R_{bc} \ll 1$ (low-frequency range), $\omega C_{bc} R_{bi}$, $\omega C_f R_{bi} \ll 1$ and $\omega C_{bc} R_{bc} \gg 1$ (intermediate frequency range) and $\omega C_{bc} R_{bi} \gg 1$ (high-frequency range). In the measurements on high-speed devices,

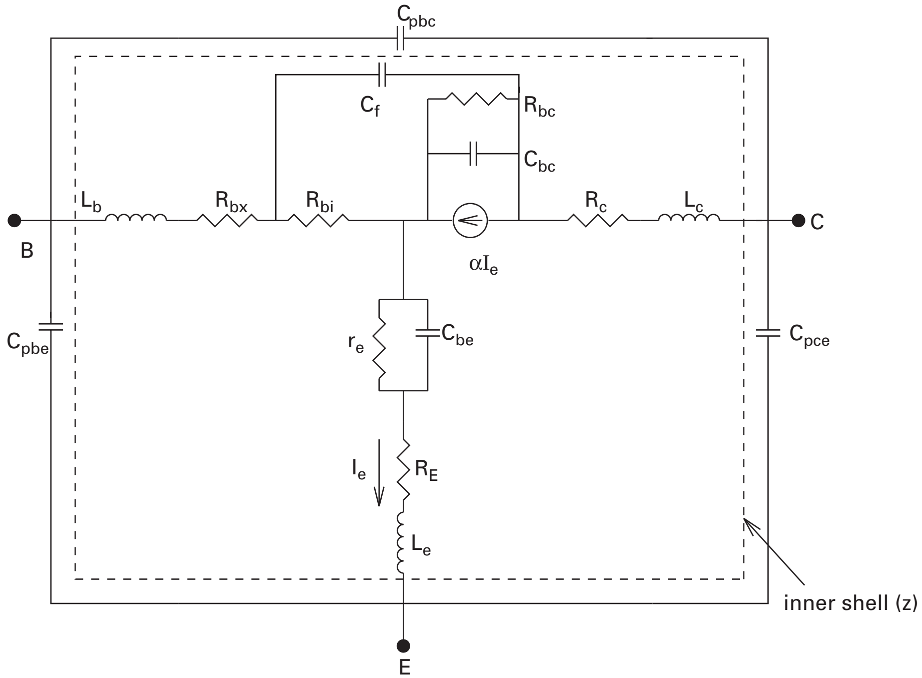


Fig. 3.25 The small-signal equivalent circuit of the AlGaAs/GaAs HBT (B. Li, S. Prasad, L.W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

most of the frequency data are located in the intermediate frequency range. In the low-frequency range,

$$Z_{11} - Z_{12} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} + R_{\mathrm{bi}} - j\omega R_{\mathrm{bc}} R_{\mathrm{bi}} C_{\mathrm{s}} \tag{3.17}$$

$$\begin{aligned} Z_{12} = {} & R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} \\ & + (1-\alpha) \times j\omega R_{\mathrm{bi}} C_{\mathrm{f}} (R_{\mathrm{bc}} - j\omega R_{\mathrm{bc}}^2 C_{\mathrm{s}}) \end{aligned} \tag{3.18}$$

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + R_{\mathrm{bc}} - j\omega R_{\mathrm{bc}}^2 C_{\mathrm{s}}, \tag{3.19}$$

Where $C_{\mathrm{s}} = C_{\mathrm{bc}} + C_{\mathrm{f}}$.

In the intermediate frequency range, the Z-parameters can be approximated as

$$Z_{11} - Z_{12} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} + R_{\mathrm{bi}} \frac{C_{\mathrm{bc}}}{C_{\mathrm{s}}} \tag{3.20}$$

$$\begin{aligned} Z_{12} = {} & R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} + (1-\alpha) \\ & \times R_{\mathrm{bi}} \times \frac{C_{\mathrm{f}}}{C_{\mathrm{s}}} \end{aligned} \tag{3.21}$$

$$\begin{aligned} Z_{22} - Z_{21} = {} & R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + \frac{1}{j\omega C_{\mathrm{s}}} + \frac{1}{\omega^2 R_{\mathrm{bc}} C_{\mathrm{s}}} \\ & - \frac{R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}}{C_{\mathrm{s}}^2} - j \frac{R_{\mathrm{bi}} C_{\mathrm{f}} C_{\mathrm{bc}}}{\omega C_{\mathrm{s}}^3 R_{\mathrm{bc}}}. \end{aligned} \tag{3.22}$$

In the high-frequency range, the simplified relation is

$$\begin{aligned} Z_{22} - Z_{21} = {} & R_{\mathrm{C}} + j\omega L_{\mathrm{c}} - \frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} + \frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} \\ & \times \left(\frac{1}{j\omega C_{\mathrm{f}} R_{\mathrm{bi}}} + \frac{1}{j\omega C_{\mathrm{bc}} R_{\mathrm{bi}}} \right) \end{aligned} \tag{3.23}$$

$$\alpha = \alpha_0 \left[\frac{1}{1 + \frac{j\omega}{\omega_0}} \right] e^{-\mathrm{j}\omega\tau}. \tag{3.24}$$

Example: Extraction of equivalent circuit elements

For the HBT used in this example, the intermediate frequency range is taken to be approximately from 0.5 GHz to 20 GHz, and the high-frequency range should go up to 40 GHz. The condition for the high frequency is relaxed and frequencies over 25 GHz are considered to be in the high-frequency range.

A. Extraction of the parasitic elements

If no test structure is available for extracting the parasitics, it is still possible to extract or estimate the pad capacitances from the HBT under cut-off operation [7, 24]. Under cut-off operation, we have zero V_{BE}, zero I_{c}, and variable V_{CB}. The HBT equivalent circuit of Figure 3.25 is reduced to the simplified circuit shown in Figure 3.26, provided the influence of the inductances and resistances remains negligible and the conditions

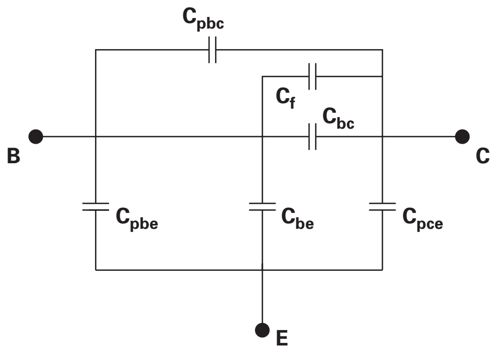


Fig. 3.26 The simplified HBT equivalent circuit under cutoff operations in which both junctions are reverse-biased and the influence of the inductances and resistances remains negligible (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

$R_{\mathrm{bc}} \gg (1/\omega C_{\mathrm{bc}})$ and $R_{\mathrm{bc}} \gg (1/\omega C_{\mathrm{be}})$ are satisfied. The capacitances in the equivalent circuit can be calculated directly:

$$C_{\mathrm{pbe}} + C_{\mathrm{be}} = \frac{\mathrm{Im}(Y_{11}) + \mathrm{Im}(Y_{12})}{\omega} \tag{3.25}$$

$$C_{\mathrm{pce}} = \frac{\mathrm{Im}(Y_{22}) + \mathrm{Im}(Y_{12})}{\omega} \tag{3.26}$$

$$C_{\mathrm{pbc}} + C_{\mathrm{f}} + C_{\mathrm{bc}} = -\frac{\mathrm{Im}(Y_{12})}{\omega}. \tag{3.27}$$

In the above equations, the $C_{\mathrm{pbe}}, C_{\mathrm{pbc}}, C_{\mathrm{pce}}$ and C_{f} are considered to be bias-independent and C_{be} and C_{bc} are bias-dependent elements. The value of C_{pce} can be calculated from Equation (3.26). Figure 3.27 clearly shows that C_{pce} is bias-independent.

C_{be}, the base–emitter junction capacitance, can be described by the equation

$$C_{\mathrm{be}} = \frac{C_{\mathrm{jbe0}}}{\left[1 + \left(V_{\mathrm{EB}}/V_{\mathrm{jb}}\right)\right]^{M_{\mathrm{jbe}}}}. \tag{3.28}$$

The extraction of C_{pbe} can be carried out by fitting the sum $C_{\mathrm{pbe}} + C_{\mathrm{be}}$ to Equation (3.28) at different reverse base–emitter voltages or by using the iteration method in which different values of V_{jbe}, M_{jbe} and C_{jbe0} are tried until the plot of $C_{\mathrm{pbe}} + C_{\mathrm{be}}$ versus $[1 + (V_{\mathrm{EB}}/V_{\mathrm{jbe}})]^{-M_{\mathrm{jbe}}}$ is a straight line shown in Figure 3.28.

Similarly, $(C_{\mathrm{f}} + C_{\mathrm{pbc}})$ can be extracted by fitting the sum $(C_{\mathrm{f}} + C_{\mathrm{pbc}} + C_{\mathrm{bc}})$ to the expression for junction capacitance at different base–collector voltages. However, it must be noted that it is difficult to distinguish between the base–collector coupling capacitance and extrinsic base–collector capacitance [24]. This is due to the fact that the distance between the base probe tip and the collector probe tip in probe stations used for most high-speed measurements is usually longer, and thus the coupling effect between base and collector contacts must be very small; furthermore, the influence of C_{pbc} can be absorbed by the extrinsic base–collector capacitance C_{f}. Thus, C_{pbc} can

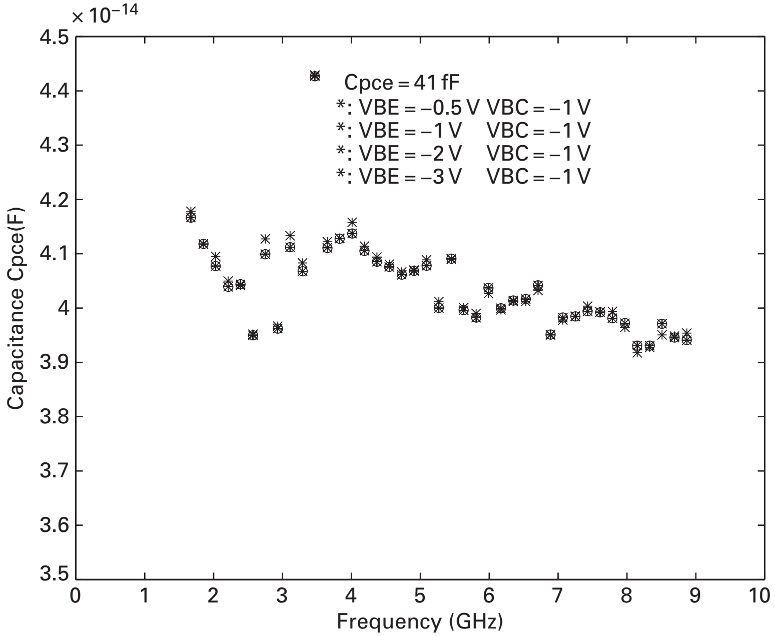


Fig. 3.27 The calculated C_{pce} versus frequency (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

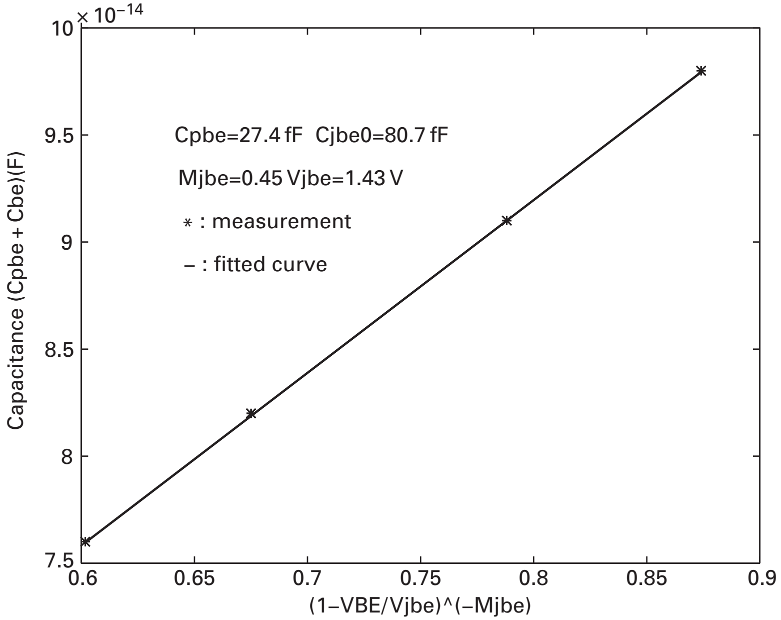


Fig. 3.28 The fitting of the sum $C_{\mathrm{pbe}} + C_{\mathrm{be}}$ to the expression for junction capacitance (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

be chosen to be zero. Such an assumption is also confirmed by the empirical optimisation procedures. The S-parameters measured over the frequency range of interest (here the range chosen was 50 MHz–36 GHz) are first converted to Y-parameters. After de-embedding the effect of the pad capacitances, the Y-parameters of the inner shell are converted to Z-parameters. Most of the elements are extracted from an analysis of the behaviour of the Z-parameters. Certain constraints are obtained to help in conditioning the optimisation procedure and to reduce the uncertainty.

B. The base–collector capacitance C_s

As indicated in Equation (3.66), the following approximation is valid in the intermediate frequency range:

$$\mathrm{Im}(Z_{22} - Z_{21}) = j\omega L_c + \frac{1}{j\omega C_s} - j\frac{R_{bi}C_f C_{bc}}{\omega C_s^3 R_{bc}}. \tag{3.29}$$

At the low end of the intermediate frequency range, the second term is much greater than the other terms on the right side of the equation. C_s can be extracted by the equation:

$$C_s = \frac{1}{\omega \mathrm{Im}(Z_{22} - Z_{21})}. \tag{3.30}$$

The extracted C_s from Equation (3.30) at the bias values for $V_{CE} = \{0.5\,\mathrm{V}, 1.0\,\mathrm{V}, 2.0\,\mathrm{V}, 4.0\,\mathrm{V}, 7.0\,\mathrm{V}\}$ and $I_B = 200\,\mu\mathrm{A}$ is shown in Figure 3.29.

At $V_{CE} = 0.5$ V, $1/\omega C_s \simeq 1224\,\Omega$, which is much larger than the other terms in the intermediate frequency range if the values extracted below are used. The base–collector

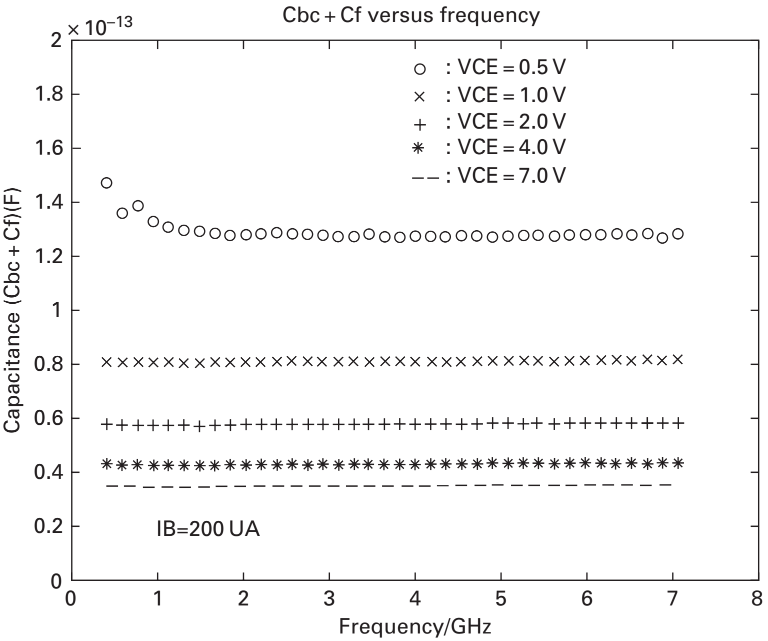


Fig. 3.29 The extracted value C_{bc} at different base collector voltages (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

capacitance is observed to decrease as V_{CE} increases. This results from the increased width of the base–collector depletion region due to the increased V_{CE}. The deviation of C_{s} is less than 5% except when $V_{\mathrm{CE}} = 0.5\,\mathrm{V}$. The base–collector junction is forward-biased at this value of V_{CE} when the intermediate frequency range moves up.

The extrinsic base–collector capacitance is generally a weak function of the base–collector junction voltage. In extreme cases, it can be considered to be independent of the bias variation, or the ratio of the extrinsic capacitance to the total base–collector capacitance is considered to be a constant. Practically, the extrinsic capacitance C_{f} is the in-between case. For simplicity, the extrinsic capacitance is considered to be fixed and extracted from the values of C_{s} at the different base–collector voltages. A method similar to that used for the extraction of the value of C_{pbe} and C_{pce} is applied. The value of C_{s} thus obtained can be compared with what is obtained from the cutoff measurement. The parameters for the base–collector junction capacitance are also extracted from this approach.

C. The collector contact lead inductor L_{c}

The collector lead conductor L_{c} can be calculated from Equation (3.29):

$$L_{\mathrm{c}} = \frac{1}{\omega}\left[\mathrm{Im}(Z_{22} - Z_{21}) + \frac{1}{\omega C_{\mathrm{s}}}\right]. \tag{3.31}$$

The third term in Equation (3.29) is assumed to be small enough and is neglected. This straightforward method is not as accurate as expected. The deviation of the extracted value of L_{c} is large and an accurate value of L_{c} is difficult to obtain. The reason for this is that the small error resulting from extracting C_{s} could lead to large errors in L_{c}. The differentiation of Equation (3.31) yields

$$\begin{aligned}\Delta L_{\mathrm{c}} &= \frac{1}{\omega^2}\left(-\frac{1}{C_{\mathrm{s}}^2}\right)\Delta C_{\mathrm{s}}\\ &= -\frac{1}{\omega^2 C_{\mathrm{s}}} \times \frac{\Delta C_{\mathrm{s}}}{C_{\mathrm{s}}}.\end{aligned} \tag{3.32}$$

L_{c} is very sensitive to even a 5% error in extracting C_{s}. The error in estimating L_{c} resulting from the error in the estimation of C_{s} is plotted in Figure 3.30. L_{c} is very sensitive to the error in extracting C_{s}. However, L_{c} is less sensitive to the error in estimating C_{s} if the magnitude of C_{s} becomes larger. Therefore, a good bias point to extract L_{c} would be zero bias at which the third term in the equation is negligible and the value of C_{s} is larger. The L_{c} extracted at zero bias is shown in Figure 3.31.

D. The base–collector resistance R_{bc}

The real part of $Z_{22} - Z_{21}$ in the middle frequency range is given by

$$\mathrm{Re}(Z_{22} - Z_{21}) = R_{\mathrm{c}} + \frac{1}{\omega^2 R_{\mathrm{bc}} C_{\mathrm{s}}} - \frac{R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}}{C_{\mathrm{s}}^2}. \tag{3.33}$$

If the term $1/\omega^2 R_{\mathrm{bc}} C_{\mathrm{s}}$ is much larger than the other two terms, R_{bc} can be approximately extracted from the real part of $Z_{22} - Z_{21}$ in the lower middle-frequency range:

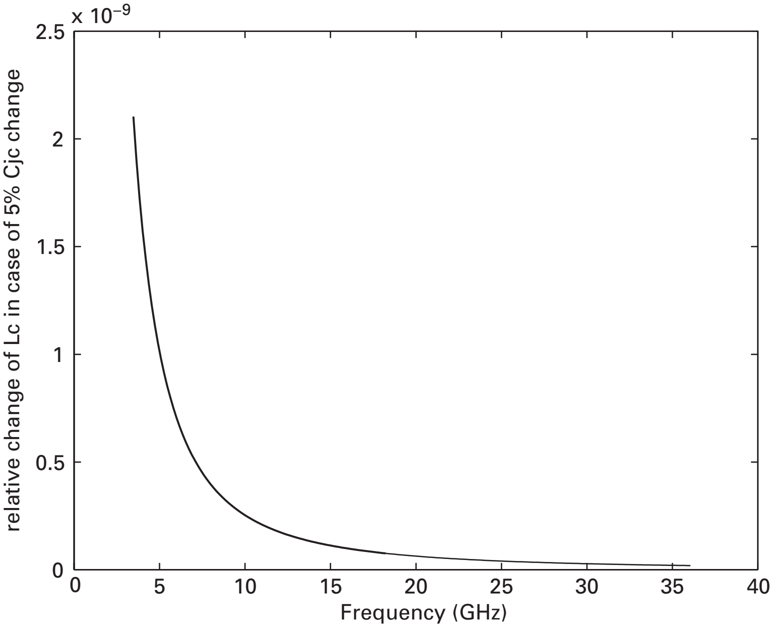


Fig. 3.30 The variation of L_c with a 5% error in the estimation of C_s (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

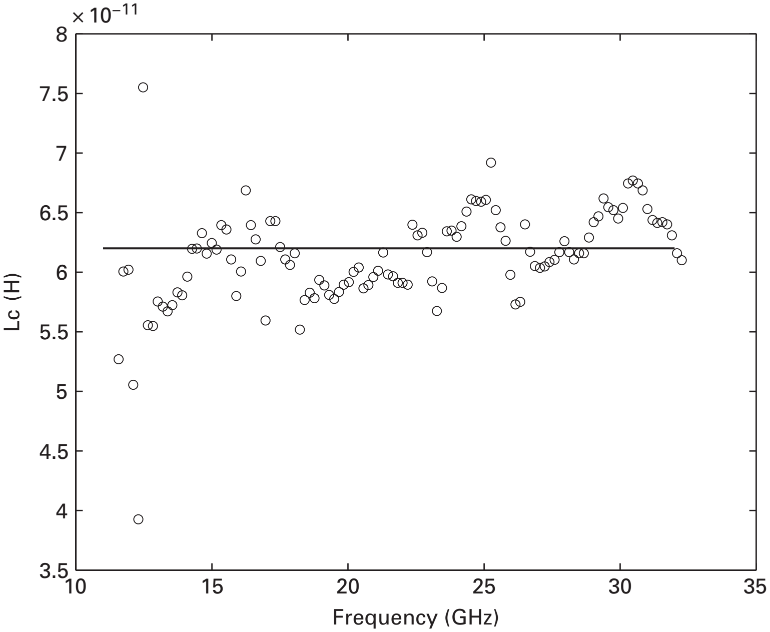


Fig. 3.31 The extracted L_c versus frequency (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

$$R_{\mathrm{bc}} = \frac{1}{\omega^2 C_{\mathrm{s}} \mathrm{Re}(Z_{22} - Z_{21})} \tag{3.34}$$

The second term in Equation (3.33) is inversely proportional to ω^2; therefore, the magnitude of the second term decreases rapidly. The other two terms cannot be neglected as the frequency increases to a certain point. However, the extracted value of R_{bc} is not significant since the value of R_{bc} is very large and does not affect the frequency response much as long as we are only concerned with forward operation. Figure 3.32 shows the extracted R_{bc} at the bias $I_{\mathrm{B}} = 200\,\mu\mathrm{A}$ and $V_{\mathrm{CE}} = \{0.5\,\mathrm{V}, 1.0\,\mathrm{V}, 2.0\,\mathrm{V}, 4.0\,\mathrm{V}, 7.0\,\mathrm{V}\}$. The magnitude of R_{bc} increases as V_{CE} increases.

Figure 3.33 shows the extracted R_{bc} without de-embedding the pad capacitances at the bias $I_{\mathrm{B}} = 200\,\mu\mathrm{A}$, $V_{\mathrm{CE}} = 1.0\,\mathrm{V}$. The magnitude of the calculated R_{bc} is negative beyond 1 GHz. This shows that physically meaningless values may be obtained if no de-embedding procedure is carried out.

E. The collector extrinsic resistance R_{c}

The R_{c} could be extracted by plotting $\mathrm{Re}(Z_{22} - Z_{21})$ versus $1/\omega^2$ in the high frequency range. The y-axis intercept is the value of R_{c}. The requirement for the high-frequency range is difficult to be achieved and the conditions for the requirement are relaxed. C_{s} is bias-dependent and the larger value of C_{s} could be obtained from S-parameters at zero bias. R_{c} should be extracted from zero bias by this method since the R_{c} is more significant in Equation (3.23) at zero bias.

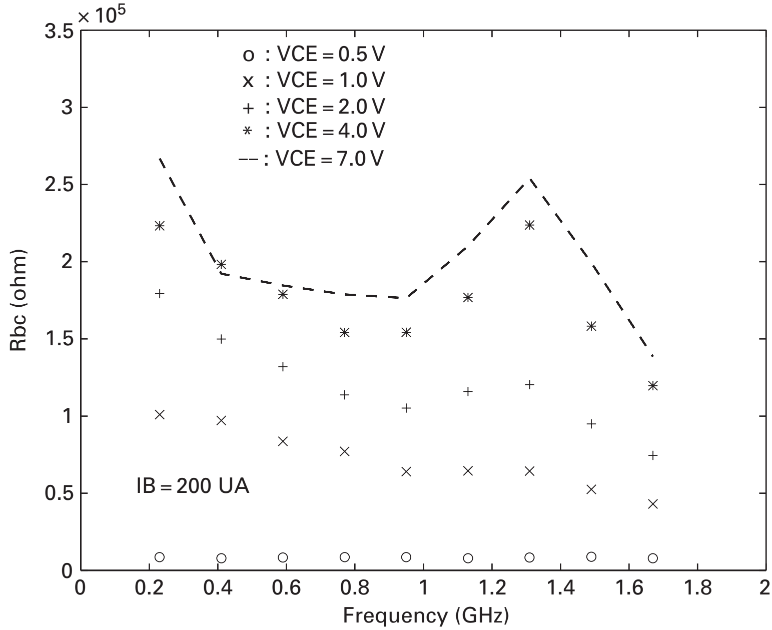


Fig. 3.32 The extracted R_{bc} at different base–collector voltages (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

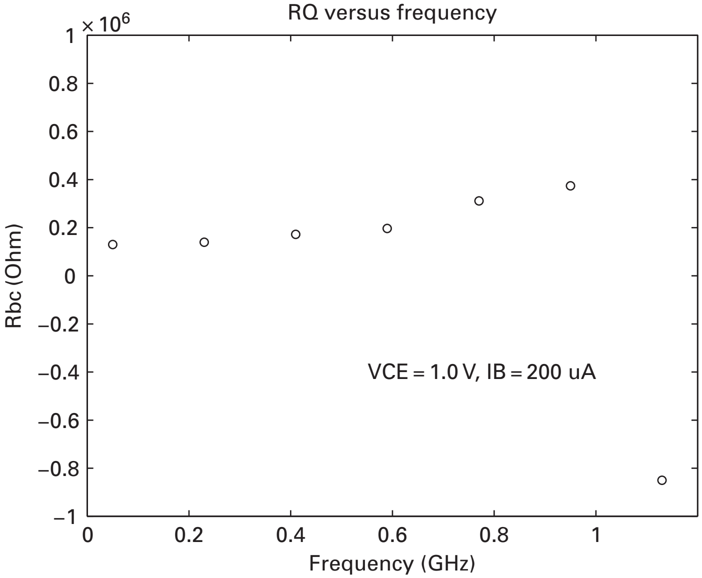


Fig. 3.33 The extracted R_{bc} at the bias $I_{\mathrm{B}} = 200\,\mu\mathrm{A}$, $V_{\mathrm{CE}} = 1.0\,\mathrm{V}$ without de-embedding the pad capacitances (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

F. The base contact lead inductor L_{b}

Applying the first-order approximation, L_{b} can be easily extracted from the imaginary part of $Z_{11} - Z_{12}$ in the middle-frequency range. That is

$$L_{\mathrm{b}} = \mathrm{Im}(Z_{11} - Z_{12})/\omega. \tag{3.35}$$

The extracted L_{b} at different biases is shown in Figure 3.34 without the de-embedding procedure.

The dependence of the value of L_{b} on the bias V_{CE} is attributable to the pad capacitance. After the de-embedding procedure is carried out, the extracted L_{b} is shown in Figure 3.35.

The magnitude variation of L_{b} at the different biases is very small and almost negligible, so L_{b} can be considered to be independent of bias.

G. The intrinsic and extrinsic base resistances

In principle, the sum of the intrinsic and extrinsic base resistances, $R_{\mathrm{bx}} + R_{\mathrm{bi}}$, can be extracted from the low-frequency data, and the extrinsic base resistance, R_{bx}, can be extracted from the high-frequency data if the equivalent circuit shown in Figure 3.25 describes the frequency response of the HBT accurately. However, most of the frequency data are located in the intermediate frequency range. The requirement for the high frequency condition is difficult to be satisfied, and the data at extremely low frequencies are not available. The constraints on the base resistances thus can be obtained from the real part of $Z_{11} - Z_{12}$ in the intermediate frequency range:

$$\mathrm{Im}(Z_{11} - Z_{12}) = R_{\mathrm{bx}} + R_{\mathrm{bi}}\frac{C_{\mathrm{bc}}}{C_{\mathrm{s}}}. \tag{3.36}$$

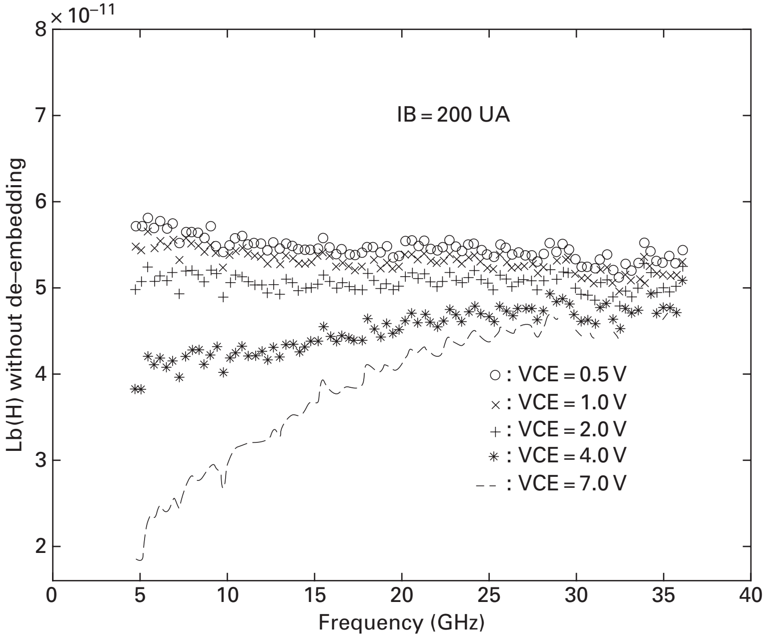


Fig. 3.34 The L_b versus frequency, in which the pad capacitances have not been de-embedded (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

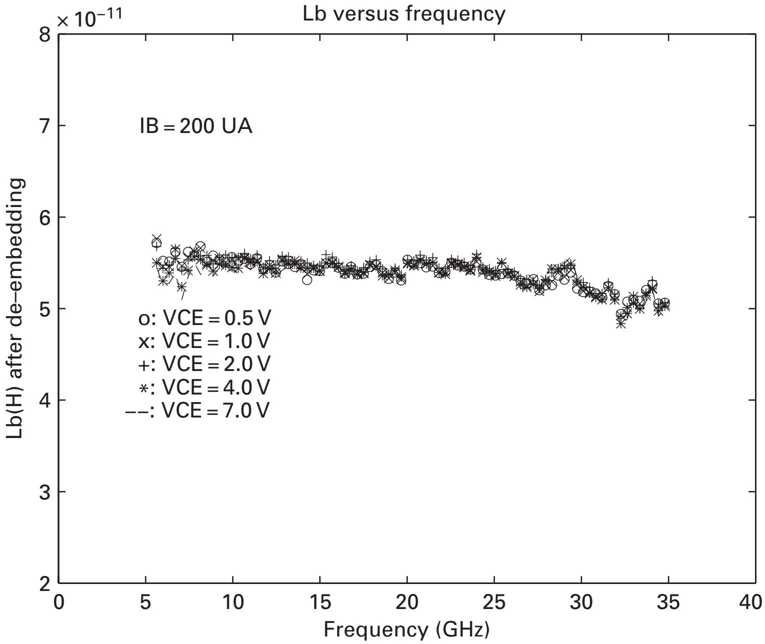


Fig. 3.35 The L_b versus frequency, in which the pad capacitance effect has been de-embedded (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

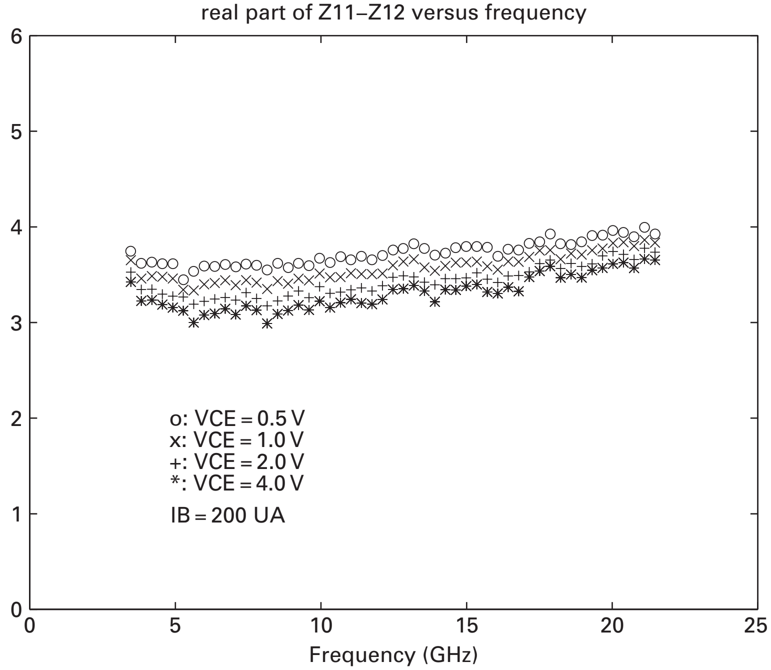


Fig. 3.36 The values of $R_{\text{bx}} + R_{\text{bi}}(C_{\text{bc}}/C_{\text{s}})$ after de-embedding the pad capacitances (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

The pad capacitances have a significant effect on the value obtained for $R_{\text{bx}} + R_{\text{bi}}\,(C_{\text{bc}}/C_{\text{s}})$. The maximum variations of $R_{\text{bx}} + R_{\text{bi}}(C_{\text{bc}}/C_{\text{s}})$ before and after de-embedding the pad capacitances are 5 and 0.5 respectively. The result after removing the pad parasitics is shown in Figure 3.36.

This value is used to constrain the optimisation procedure in order to obtain accurate values of R_{bx} and R_{bi}. The variation of $R_{\text{bx}} + R_{\text{bi}}(C_{\text{bc}}/C_{\text{s}})$ is due to the change of C_{bc} with the base–collector voltage V_{CE}. C_{bc} decreases as V_{CE} increases. This causes the ratio $C_{\text{bc}}/(C_{\text{bc}} + C_{\text{f}})$ to decrease and hence the magnitude of $R_{\text{bx}} + R_{\text{bi}}(C_{\text{bc}}/C_{\text{s}})$ decreases.

H. The emitter resistance R_{E} and base–emitter resistance r_{e}

$R_{\text{E}} + r_{\text{e}}$ can be obtained from the real part of Z_{12} in the intermediate frequency range. With the high collector current where the neutral base recombination is the dominant recombination, $R_{\text{E}} + r_{\text{e}}$ can be expressed as

$$R_{\text{E}} + r_{\text{e}} = R_{\text{E}} + \frac{n_{\text{f}}\,kT}{q\,I_{\text{E}}}. \tag{3.37}$$

The real part of Z_{12} in the intermediate frequency range is the sum of $r_{\text{e}} + R_{\text{E}}$. The plot of $r_{\text{e}} + R_{\text{E}}$ versus $1/I_{\text{E}}$ would give the values of R_{E}, r_{e} and ideality factor n_{f}.

I. The emitter lead inductor and base–emitter capacitance

$L_{\text{e}} - C_{\text{be}}r_{\text{e}}^2$ can be obtained from the imaginary part of Z_{12} in the low middle frequency range. In the case of high collector currents, the fraction of the depletion capacitance in the base–emitter capacitance C_{be} is small and C_{be} can be approximated to be proportional to I_{E}, and we also have $r_{\mathrm{e}} \propto 1/I_{\mathrm{E}}$. Therefore, the y intercept of the plot of $L_{\mathrm{e}} - r_{\mathrm{e}}^2 C_{\mathrm{be}}$ versus $1/I_{\mathrm{E}}$ gives the value of L_{e}. The value of $L_{\mathrm{e}} - C_{\mathrm{be}} r_{\mathrm{e}}^2$ at $f = 5\,\mathrm{GHz}$ is used for this purpose. $L_{\mathrm{e}} - C_{\mathrm{be}} r_{\mathrm{e}}^2$ is plotted versus $1/I_{\mathrm{e}}$ and shown in Figure 3.37.

Based on the values of L_{e} and r_{e} obtained previously, the value of C_{be} can be easily calculated. The value of C_{be} obtained in this way only serves to give the initial value of C_{be}. An accurate value of C_{be} is obtained from the optimisation procedure. It is noted that the magnitude of C_{be} is not sensitive to the optimisation procedure. This was also reported in [24] where the value of C_{be} is calculated from f_α (where f_α is the transport factor α cutoff frequency and $f_\alpha = 1/r_{\mathrm{e}} C_{\mathrm{be}}$). An accurate value of C_{be} is extremely difficult to obtain since changing the value of C_{be} does not change the error of optimisation much over the bias ranges for this example.

J. The transport factor α

The transport factor α can be calculated directly by

$$\alpha = \frac{Z_{21} - Z_{12}}{Z_{22} - Z_{21} - R_{\mathrm{c}} - j\omega L_{\mathrm{c}}}. \tag{3.38}$$

Assuming a single-pole approximation, one can write

$$\alpha = \frac{\alpha_0}{1 + \dfrac{j\omega}{\omega_\alpha}} e^{-\mathrm{j}\omega\tau}, \tag{3.39}$$

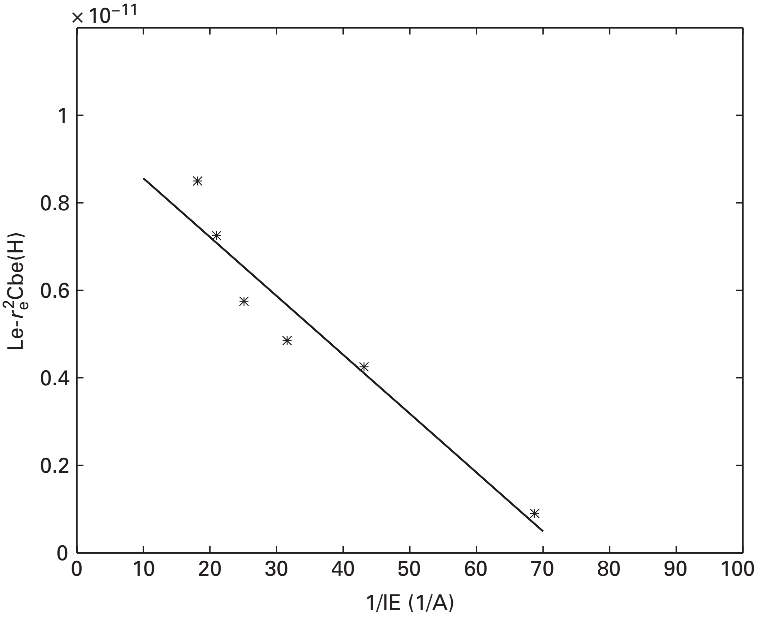


Fig. 3.37 $L_{\mathrm{e}} - C_{\mathrm{be}} r_{\mathrm{e}}^2$ versus the inverse emitter current. $L_{\mathrm{e}} = 9.99\,\mathrm{pH}$ (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

where ω_α can be expressed as

$$\omega_\alpha = \frac{1}{C_{\rm be} r_{\rm e}} = \frac{1}{\tau_{\rm b}} \tag{3.40}$$

where $\tau_{\rm b}$ is the base transit time and is related to physical parameters by $W_{\rm BC}^2/2D_{\rm n}$ for npn HBTs. The magnitude of $\alpha(\omega)$ at $I_{\rm B} = 200\,\mu$A with different collector–emitter voltages is shown in Figure 3.38 together with the fitted curve of the magnitude of α.

α_0 is obtained by taking the value of $|\alpha|$ at low frequency and ω_α (and therefore the base transit time $\tau_{\rm b} = 1/\omega_\alpha$) can be calculated directly at each frequency using

$$\tau_{\rm b} = \frac{\sqrt{\alpha_0^2 - |\alpha(\omega)|^2}}{\omega|\alpha(\omega)|}. \tag{3.41}$$

The calculated $\tau_{\rm b}$ at $I_{\rm B} = 200\,\mu$A with different collector–emitter voltages is shown in Figure 3.39. Since the base is heavily doped, the base width modulation effect in the HBT is negligible and therefore $\tau_{\rm b}$ should be a weak function of $I_{\rm B}$ and $V_{\rm CE}$. The dependence of $\tau_{\rm b}$ on the base current and collector–emitter voltage is not completely clear. One possible reason for the α dependence on $V_{\rm CE}$ is the self-heating effect in the HBT. The diffusion coefficient $D_{\rm n} = (kT/q)\mu_{\rm n}$ is a function of the temperature in which $\mu_{\rm n} \propto T^{-{\rm s}}$. An often quoted value of s is 2.3 (for intrinsic GaAs). It is noted that the $D_{\rm n}$ decreases when the dissipated power in the HBTs increases. Therefore, the $\tau_{\rm b}$ increases with larger $V_{\rm CE}$ values. The emitter–collector phase delay time can be calculated by

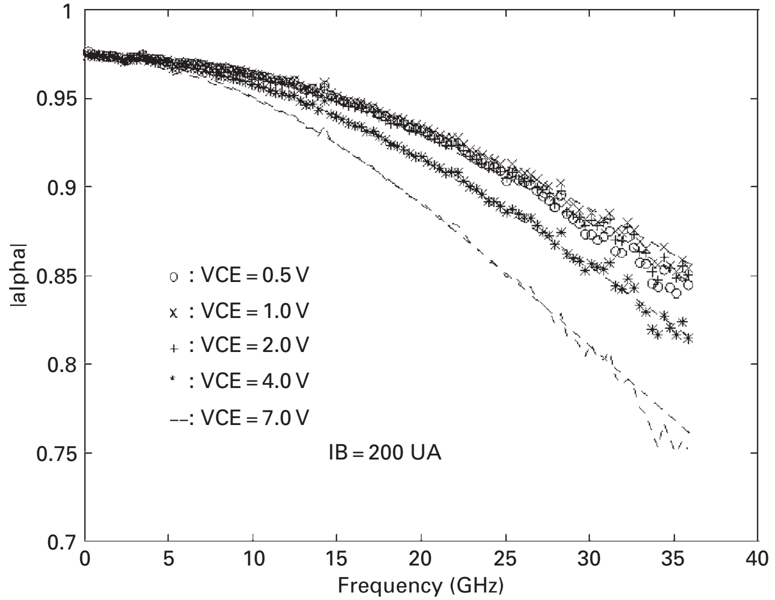


Fig. 3.38 $|\alpha|$ versus frequency at different collector emitter voltages (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

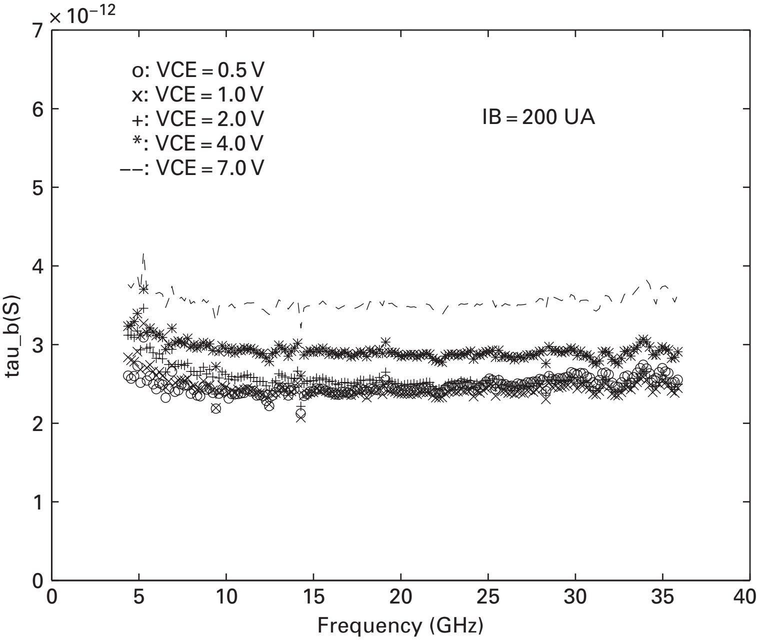


Fig. 3.39 τ_b versus frequency at different collector–emitter voltages (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

$$\tau = \frac{1}{\omega}\left[-\angle\alpha(\omega) - \tan^{-1}\left(\frac{\omega}{\omega_\alpha}\right)\right], \tag{3.42}$$

where $\tau = (m/1.2)\tau_b + \tau_c$ and $m \simeq 0.22$. The calculated τ versus frequency at $I_B = 200\,\mu A$ with different values of V_{CE} is shown in Figure 3.40.

When the collector–emitter voltage increases, the collector transit time $\tau_c = W_{BC}/2v_{sat}$ increases due to the larger base–collector space region. Therefore, the emitter–collector delay time increases as expected with the larger collector–emitter voltage. This might be explained by the self-heating effect in the HBT. The emitter–collector delay time is a monotonously increasing function of τ_b and τ_c. As the power dissipated in the HBT increases, the temperature of the intrinsic part of the HBT increases. That causes the diffusion coefficient D_n to decrease and results in the larger base transit time τ_b, and hence the larger emitter–collector delay time.

3.7.2 Results of the parameter extraction

The values of the bias-independent elements are given in Table 3.4.

All of the bias-independent elements are extracted from the procedure described above except for L_e, R_{bi} and R_{bx}. Accurate values of L_e, R_{bx} and R_{bi} are obtained from the empirical optimisation procedure. Let $a = R_{bx} + R_{bi}(C_{bc}/C_s)$ and $\gamma = C_{bc}/C_s$. The initial values of R_{bi} and R_{bx} are estimated from the variation of a. We have $R_{bx} = \Delta a/\Delta\gamma$. The calculated values of R_{bx} and R_{bi} are listed in Table 3.4. Instead of defining just the absolute error and just the relative error, the mixed relative and absolute

Table 3.4 The bias-independent parameters

Parameters	Values (analytical)	Values (optimised)
C_{pbc}(fF)	0	0
C_{pbe}(fF)	27.4	27.4
C_{pce}(fF)	41	42
C_{f}(fF)	16.5	16.5
L_{b}(pH)	55	55
L_{e}(pH)	9.9	5.46
L_{c}(pH)	61	61
$R_{\text{bx}}(\Omega)$	1.38	1.42
$R_{\text{bi}}(\Omega)$	2.3	4.049
$R_{\text{E}}(\Omega)$	1.832	1.832
$R_{\text{c}}(\Omega)$	4.99	4.99
Error	2.2 %	0.43 %

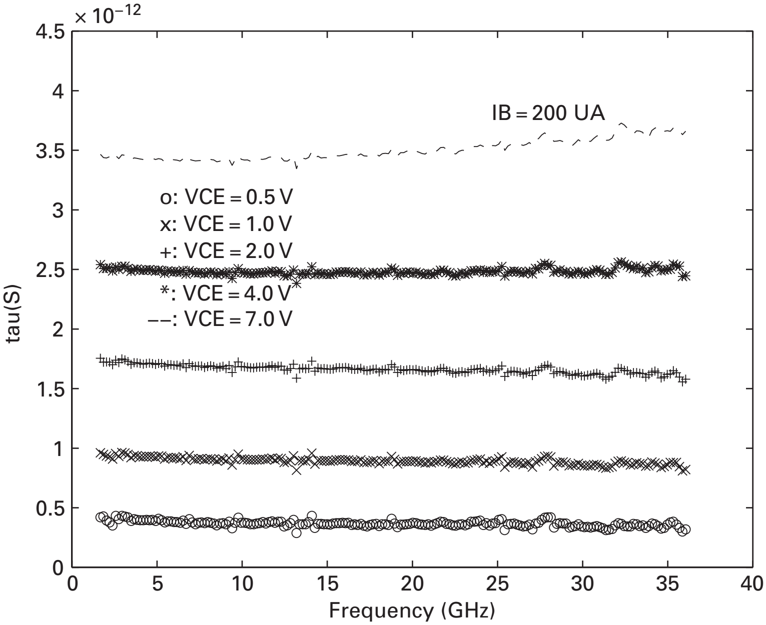


Fig. 3.40 τ versus frequency at different collector–emitter voltage (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

errors are used to obtain the best fit between the measurement and the simulation. The optimisation error is defined by

$$E = 1/4N \sum_{i=1}^{N} \left[\left(\frac{S_{11}^{\text{mea}} - S_{11}^{\text{sim}}}{S_{11}^{\text{mea}}} \right)^2 + \left(\frac{S_{12}^{\text{mea}} - S_{12}^{\text{sim}}}{S_{12}^{\text{mea}}} \right)^2 + \left(\frac{S_{21}^{\text{mea}} - S_{21}^{\text{sim}}}{S_{21}^{\text{mea}}} \right)^2 + \left(\frac{S_{22}^{\text{mea}} - S_{22}^{\text{sim}}}{S_{22}^{\text{mea}}} \right)^2 \right], \tag{3.43}$$

Table 3.5 The bias-dependent parameters

V_{CE}(V)	0.50000	1.0000	2.0000	4.0000	7.0000
I_{C}(mA)	6.0	6.2	6.1	5.9	5.6
C_{bc} (fF)	109.1	64.75	41.50	26.8	18.70
$r_{\mathrm{e}}(\Omega)$	4.168	4.0432	4.218	4.518	5.018
$R_{\mathrm{bc}}(\mathrm{k}\Omega)$	8.6	70	121	180	200
α_0	0.9741	0.9751	0.9751	0.9740	0.9740
$C_{\mathrm{be}}^{\mathrm{A}}$(fF)	344	344	344	344	344
f_{α}^{A}(GHz)	63	66	66	55	45
τ^{A} (ps)	0.42	0.95	1.7	2.52	3.42
$C_{\mathrm{be}}^{\mathrm{O}}$ (fF)	289	238	239	268	293
f_{α}^{O}(GHz)	58.2	58.9	52.3	47.10	39.9
τ^{O}(ps)	0.158	0.69	1.09	2.08	3.14
Error$^{\mathrm{A}}$	0.98%	0.84%	0.86%	0.62%	0.9%
Error$^{\mathrm{O}}$	0.50%	0.69%	0.43%	0.41%	0.56%

where N is the number of the frequency points. The errors between the measured and simulated S-parameters are also listed in Table 3.4.

The optimisation is carried out at the bias $I_{\mathrm{B}} = 200\,\mu\mathrm{A}$ and $V_{\mathrm{CE}} = 2.0\mathrm{V}$. The error between the measurement and the simulation at the bias $I_{\mathrm{B}} = 200\,\mu\mathrm{A}$ and $V_{\mathrm{CE}} = 2.0\,\mathrm{V}$ before the optimisation is already 2.2%. The bias-dependent parameters $C_{\mathrm{be}}, r_{\mathrm{e}}, C_{\mathrm{bc}}, R_{\mathrm{bc}}, \alpha_0, f_{\alpha}$ and τ at constant base current $I_{\mathrm{B}} = 200\,\mu\mathrm{A}$ and $V_{\mathrm{CE}} = \{0.5\,\mathrm{V}, 1.0\,\mathrm{V}, 2.0\,\mathrm{V}, 4.0\,\mathrm{V}, 7.0\,\mathrm{V}\}$ are given in Table 3.5. Superscript A represents the results from optimisation and superscript O represents the results from the direct analysis. Once the values of the bias-independent elements are known, all the bias-dependent values can be easily calculated and no further optimisation is needed. It is seen in Table 3.5 that, by using the directly calculated values of the bias-dependent elements, the error between simulation and measurement is very small. All the errors are less than 1%. Optimisations are also used. Only the three elements C_{be}, f_{α} and τ are optimised. The errors after optimisation are given in the Table 3.5. The variation of r_{e} is dependent on the collector current and the self-heating effect. As explained previously, accurate values of C_{be} are very difficult to obtain. The variation of C_{be} may result from the numerical techniques. The bias-dependent parameters $C_{\mathrm{be}}, r_{\mathrm{e}}, C_{\mathrm{bc}}, R_{\mathrm{bc}}, \alpha_0, f_{\alpha}$ and τ at constant collector–emitter voltage $V_{\mathrm{CE}} = 2.0\,\mathrm{V}$ and $I_{\mathrm{B}} = \{200\,\mu\mathrm{A}, 400\,\mu\mathrm{A}, 600\,\mu\mathrm{A}, 800\,\mu\mathrm{A}, 1000\,\mu\mathrm{A}\}$ are given in Table 3.6. The errors in using the analytical approach and in using the optimisation procedure based on the initial values obtained from the analytical approach are both given. It is seen that the errors, using the analytical approach, become higher if the collector currents increase. This is because the self-heating effect becomes more significant when the collector currents increase. However, the bias-independent elements are forced to be fixed in all the extraction procedures and they are practically functions of the device temperature. The thermal effect is absorbed by the bias-dependent elements after optimisation. Thus the errors become smaller. As expected, C_{be} increases with increased collector

Table 3.6 The bias-dependent parameters

$I_B(\mu A)$	200.000	400.000	600.000	800.000	1000.000
$I_C(mA)$	6.1	14.1	22.6	30.19	38.8
$C_{bc}(fF)$	41.50	37.4	32.7	28.75	27.15
$r_e(\Omega)$	4.218	1.798	1.133	0.8482	0.6912
$R_{bc}(k\Omega)$	200	200	200	200	200
α_0	0.9751	0.9785	0.9796	0.9790	0.9790
$C_{be}^{A}(fF)$	344	863	1275	1704	2190
$f_{\alpha}^{A}(GHz)$	66	72	94	114	102
$\tau^{A}(ps)$	1.7	1.6	1.52	1.73	1.2
$C_{be}^{O}(fF)$	0.239	0.682	1.070	1.537	1.984
$f_{\alpha}^{O}(GHz)$	52.3	64.4	79.1	88.9	81.2
$\tau^{O}(ps)$	1.09	1.02	0.98	0.68	0.45
Error^{A}	0.86%	2.6%	3.2%	5.2%	7.6%
Error^{O}	0.43%	0.62%	1.1%	2.1%	2.9%

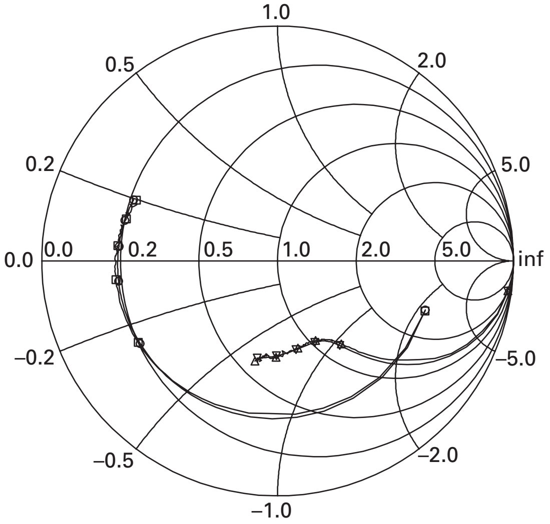


Fig. 3.41 The simulated and measured S_{11} and S_{22}. ○: measured S_{11}; □: simulated S_{11}; ∇: measured S_{22}; △: simulated S_{22} (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

currents. It is observed from Table 3.6 that C_{bc} decreases with the increased collector currents. One possible explanation is the self-heating effect and the modification of the base–collector space charge region by the injected carriers [24]. The effect of self-heating is not discussed here. The simulated and measured S-parameters at the bias $I_B = 200\,\mu A$ and $V_{CE} = 2.0\,V$ are shown in Figures 3.41–3.43. The excellent fit between the measured and modelled data shows that this procedure may be used successfully in device parameter extraction.

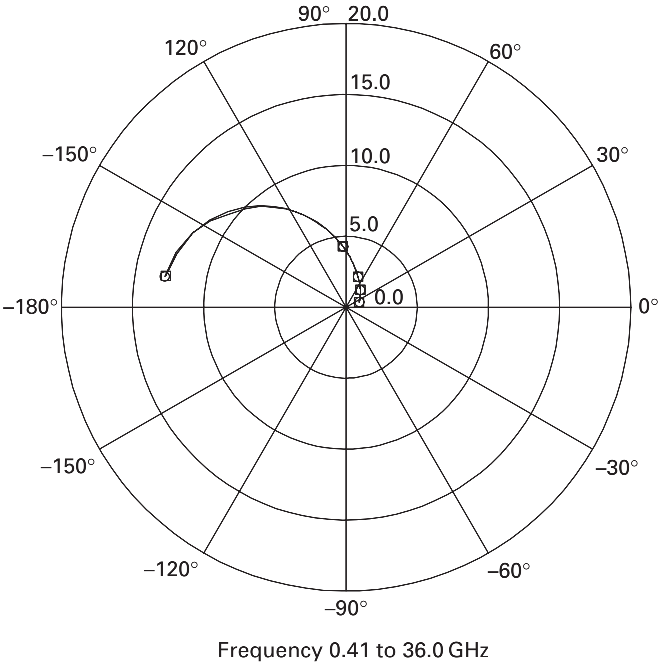


Fig. 3.42 The simulated and measured S_{21}. ○: measured S_{21}; □: simulated S_{21} (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

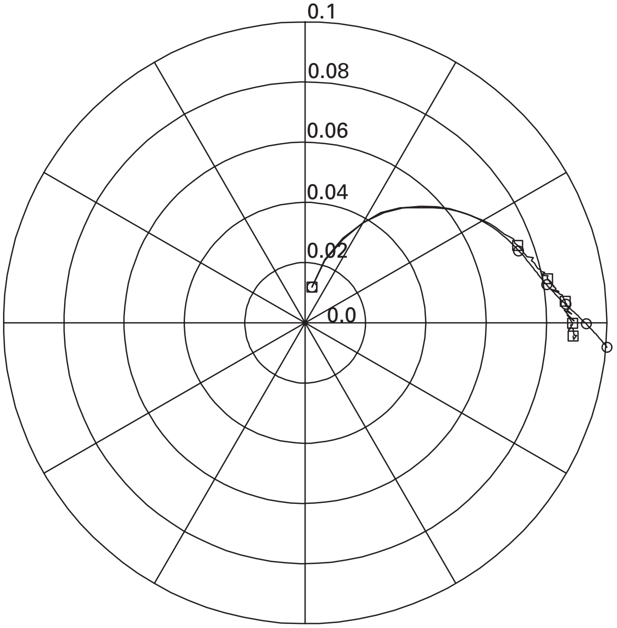


Fig. 3.43 The simulated and measured S_{12}. ○: measured S_{12}; □: simulated S_{12} (B. Li, S. Prasad, L.-W. Yang and S. C. Wang, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 10, pp. 1427–1435, October 1998. ©1998 IEEE).

In summary, the pad capacitances are extracted from the HBTs under cutoff operation. Most of the elements are obtained from the analysis of the behaviour of the Z-parameters. The values of uncertain elements are obtained from the optimisation at a specific bias. The initial values of these uncertain elements are also obtained from the analytical approach.

3.8 Basic expressions for small-signal parameter extraction

Small-signal equivalent circuit parameter extraction has been much addressed by several researchers [3, 6, 19, 20, 25, 26, 33, 34]. The basic expressions and approximations for the Z-parameters in different frequency ranges and under different bias conditions are treated in this section [14].

3.8.1 Theoretical approximations of Z-parameters for the HBT

The simplified T equivalent circuit of the HBT after de-embedding pad capacitances is shown in Figure 3.44. The T equivalent circuit is used here since it is more physically meaningful than the π equivalent circuit. The latter circuit is the mathematical representation of transistor operation. It is also easier to infer the large-signal model from the

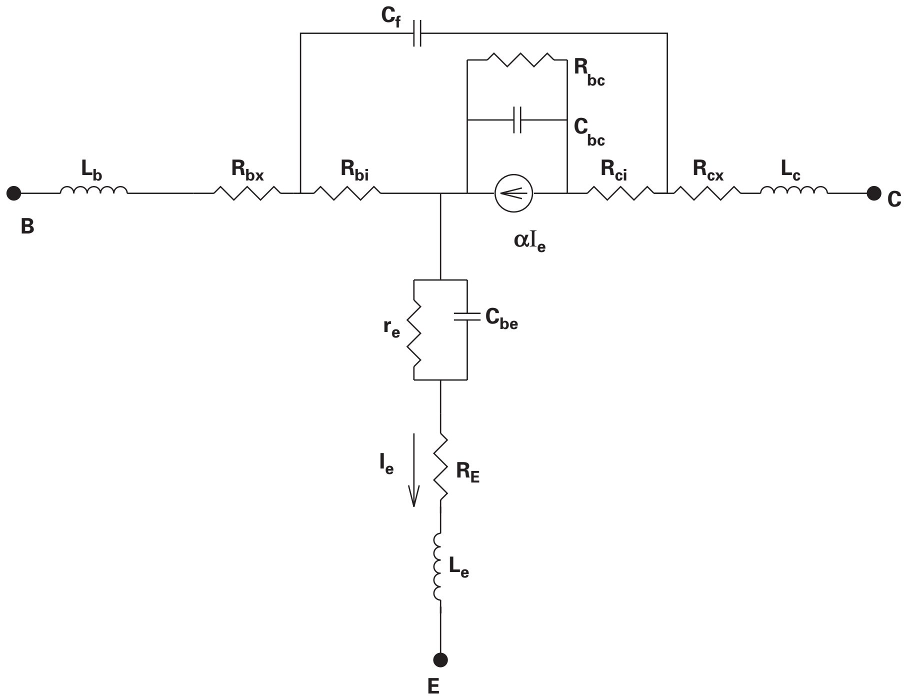


Fig. 3.44 Simplified T-type equivalent circuit, in which pad parasitics have been de-embedded (B. Li and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 5, pp. 534–539, May 1999. ©1999 IEEE).

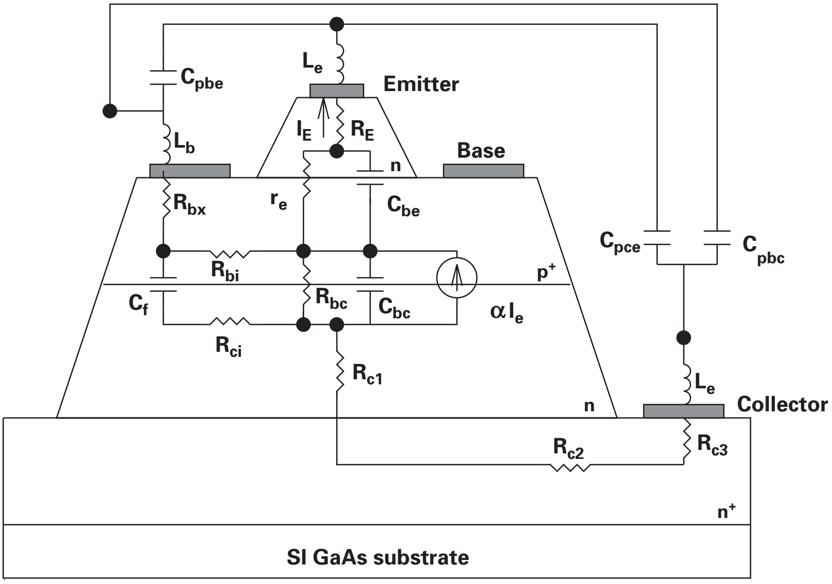


Fig. 3.45 Physical significance of the elements in the small-signal equivalent circuit (B. Li and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 5, pp. 534–539, May 1999. ©1999 IEEE).

bias-dependent element parameters by the T equivalent circuit since there are straightforward physical equations for the bias-dependent intrinsic elements. Figure 3.45 depicts the physical significance of each of the circuit elements in Figure 3.44 and also includes the extrinsic parasitic capacitances C_{pbe}, C_{pbc} and C_{pce}. The active portion of the HBT was modelled using C_{be}, C_{bc}, r_{e}, αI_{e}, R_{bc} and C_{f}. R_{E} is the extrinsic emitter resistance which consists of the contact resistance and emitter region resistance. The extrinsic collector resistance was divided into three parts: R_{c1}, R_{c2} and R_{c3}, which are, respectively, the resistance due to the n-collector, the n^+ access region, and the collector contact. The intrinsic collector resistance is represented by R_{ci} which characterises the distribution effect of the base–collector junction at the collector side. R_{c1}, R_{c2} and R_{c3} are lumped together as R_{cx} in Figure 3.44. Similarly, the extrinsic base resistance consists of a contact resistance R_{b1} and an access resistance R_{b2}. R_{b1} and R_{b2} are lumped together as R_{bx} in Figure 3.44. R_{bi} is the intrinsic base resistance. Finally, the distribution effect of the base–collector junction is modelled by the elements R_{bi}, R_{bx}, C_{f}, C_{bc} and R_{bc}. C_{pbe}, C_{pbc} and C_{pce} modelled the coupling between the base–emitter, the base–collector and the collector–emitter interconnection layers. L_{e}, L_{b} and L_{c} are the contact leads of the emitter, the base and the collector respectively:

$$Z_{\text{BC}} = \frac{R_{\text{bc}}}{1 + j\omega R_{\text{bc}} C_{\text{bc}}} \tag{3.44}$$

$$Z_{\text{E}} = R_{\text{E}} + j\omega L_{\text{e}} + \frac{r_{\text{e}}}{1 + j\omega r_{\text{e}} C_{\text{be}}} \tag{3.45}$$

$$Z_{\mathrm{F}} = \frac{1}{j\omega C_{\mathrm{f}}} \tag{3.46}$$

$$Z_{\mathrm{C}} = R_{\mathrm{c}} + j\omega L_{\mathrm{c}} \tag{3.47}$$

$$Z_{\mathrm{B}} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} \tag{3.48}$$

$$\alpha = \alpha_0 \frac{1}{1 + \dfrac{j\omega}{\omega_\alpha}} e^{-\mathrm{j}\omega\tau}. \tag{3.49}$$

The two-port network Z-parameters are given as follows (the derivation for two-port Z-parameters of this equivalent circuit is straightforward using basic circuit theory and the details are not given here):

$$\begin{aligned}
Z_{11} &= Z_{\mathrm{B}} + Z_{\mathrm{E}} + \frac{R_{\mathrm{bi}}[Z_{\mathrm{F}} + R_{\mathrm{ci}} + (1-\alpha)Z_{\mathrm{BC}}]}{\triangle} \\
Z_{12} &= Z_{\mathrm{E}} + R_{\mathrm{bi}} \times \frac{R_{\mathrm{ci}} + (1-\alpha)Z_{\mathrm{BC}}}{\triangle} \\
Z_{21} &= Z_{\mathrm{E}} + \frac{R_{\mathrm{ci}}R_{\mathrm{bi}} + (1-\alpha)Z_{\mathrm{BC}}R_{\mathrm{bi}} - \alpha Z_{\mathrm{F}}Z_{\mathrm{BC}}}{\triangle} \\
Z_{22} &= Z_{\mathrm{C}} + Z_{\mathrm{E}} + \frac{[(1-\alpha)Z_{\mathrm{BC}} + R_{\mathrm{ci}}](Z_{\mathrm{F}} + R_{\mathrm{bi}})}{\triangle},
\end{aligned} \tag{3.50}$$

where $\triangle = R_{\mathrm{bi}} + R_{\mathrm{ci}} + Z_{\mathrm{F}} + Z_{\mathrm{BC}}$.

Let $R_{\mathrm{ci}} = 0$ (in most cases, it is hard to distinguish between R_{ci} and R_{cx}, and R_{ci} and R_{cx} are lumped together as R_{c}), the following equations are obtained after some simple calculations:

$$Z_{11} - Z_{12} = Z_{\mathrm{B}} + \frac{Z_{\mathrm{F}}R_{\mathrm{bi}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}} \tag{3.51}$$

$$Z_{12} = Z_{\mathrm{E}} + \frac{(1-\alpha)Z_{\mathrm{BC}}R_{\mathrm{bi}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}} \tag{3.52}$$

$$Z_{12} - Z_{21} = \frac{\alpha Z_{\mathrm{F}}Z_{\mathrm{BC}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}} \tag{3.53}$$

$$Z_{22} - Z_{21} = Z_{\mathrm{C}} + \frac{Z_{\mathrm{F}}Z_{\mathrm{BC}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}}. \tag{3.54}$$

The most important term is

$$\Re = \frac{Z_{\mathrm{F}}Z_{\mathrm{BC}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}}. \tag{3.55}$$

This term $\Re$ is a complicated function of frequency and all the right sides of the Equations (3.51)–(3.54) can be expressed as the sum of a simple function of frequency and this term (except for a scaling constant).

Within the extreme low frequency range in which $\omega C_{\mathrm{bc}} R_{\mathrm{bc}} \ll 1$,

$$\begin{aligned}\Re &\simeq \frac{Z_{\mathrm{F}} Z_{\mathrm{BC}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}}} \\ &\simeq \frac{R_{\mathrm{bc}}}{j\omega (C_{\mathrm{f}} + C_{\mathrm{bc}}) R_{\mathrm{bc}} + 1} \\ &\simeq R_{\mathrm{bc}} - j\omega R_{\mathrm{bc}}^2 C_{\mathrm{s}},\end{aligned} \tag{3.56}$$

where we have used the assumption $R_{\mathrm{bc}} \gg R_{\mathrm{bi}}$ and let $C_{\mathrm{s}} = C_{\mathrm{f}} + C_{\mathrm{bc}}$, since this is almost always the case in the forward or saturation applications. A typical extreme low frequency can be calculated by using the typical element values. In modern process technology in which the base resistance and base–collector capacitance have been greatly reduced, the typical values $C_{\mathrm{bc}} \simeq 5 \times 10^{-14}$ F and $R_{\mathrm{bc}} \simeq 5 \times 10^4\ \Omega$, therefore the frequency $f \simeq 0.1/(2\pi R_{\mathrm{bc}} C_{\mathrm{bc}}) = 6.8$ MHz. For microwave applications, this typical frequency is very low. Substituting the expression in Equation (3.56) into Equations (3.51), (3.52) and (3.54), the equations reduce to

$$Z_{11} - Z_{12} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} + R_{\mathrm{bi}} - j\omega R_{\mathrm{bc}} R_{\mathrm{bi}} C_{\mathrm{s}} \tag{3.57}$$

$$\begin{aligned}Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} \\ + (1 - \alpha) \times j\omega R_{\mathrm{bi}} C_{\mathrm{f}} (R_{\mathrm{bc}} - j\omega R_{\mathrm{bc}}^2 C_{\mathrm{s}})\end{aligned} \tag{3.58}$$

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + R_{\mathrm{bc}} - j\omega R_{\mathrm{bc}}^2 C_{\mathrm{s}}. \tag{3.59}$$

The base–emitter resistance is small in forward bias and therefore, Z_{E} is approximated as $R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}}$ in this analysis. Z_{12} can be simplified by discarding the higher-order terms:

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} + j(1 - \alpha)\omega R_{\mathrm{bi}} R_{\mathrm{bc}} C_{\mathrm{s}}. \tag{3.60}$$

At low frequency, $\alpha \to \alpha_{\mathrm{DC}} \to 1$. Hence, Z_{12} is further simplified as

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}}. \tag{3.61}$$

The intermediate frequency range is characterised by $\omega C_{\mathrm{bc}} R_{\mathrm{bi}}, \omega C_{\mathrm{f}} R_{\mathrm{bi}} \ll 1$, but $\omega C_{\mathrm{bc}} R_{\mathrm{bc}} \gg 1$. The approximation for $\Re$ is given as

$$\begin{aligned}\Re &= \frac{Z_{\mathrm{F}} Z_{\mathrm{BC}}}{Z_{\mathrm{F}} + Z_{\mathrm{BC}} + R_{\mathrm{bi}}} \\ &\simeq \frac{Z_{\mathrm{F}} Z_{\mathrm{BC}}}{Z_{\mathrm{F}} + Z_{\mathrm{BC}}} \times \left(1 - \frac{R_{\mathrm{bi}}}{Z_{\mathrm{F}} + Z_{\mathrm{BC}}}\right)\end{aligned} \tag{3.62}$$

$$\begin{aligned}&\simeq \left(\frac{1}{j\omega C_{\mathrm{s}}} + \frac{1}{\omega^2 C_{\mathrm{s}}^2 R_{\mathrm{bc}}}\right)\left(1 - \frac{j\omega C_{\mathrm{f}} C_{\mathrm{bc}} R_{\mathrm{bi}}}{C_{\mathrm{s}}}\right) \\ &\simeq \frac{1}{j\omega C_{\mathrm{s}}} + \frac{1}{\omega^2 C_{\mathrm{s}}^2 R_{\mathrm{bc}}} - \frac{R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}}{C_{\mathrm{s}}^2} - \frac{j R_{\mathrm{bi}} C_{\mathrm{f}} C_{\mathrm{bc}}}{\omega C_{\mathrm{s}}^3 R_{\mathrm{bc}}} \\ &\simeq \frac{1}{j\omega C_{\mathrm{s}}} + \frac{1}{\omega^2 C_{\mathrm{s}}^2 R_{\mathrm{bc}}} - \frac{R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}}{C_{\mathrm{s}}^2},\end{aligned} \tag{3.63}$$

where higher-order terms have been discarded. A typical intermediate frequency is calculated by using the typical element values. Assume $R_{\mathrm{bi}} = 10\,\Omega$, the typical value of frequency is $0.1/(2\pi R_{\mathrm{bi}} C_{\mathrm{bc}}) \simeq 0.64\,\mathrm{GHz}$. The value $\omega C_{\mathrm{bc}} R_{\mathrm{bc}} \simeq 10$ appropriately justifies the assumption. Substituting the approximation in Equation (3.63) into Equations (3.51), (3.52) and (3.54), we obtain

$$Z_{11} - Z_{12} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} + R_{\mathrm{bi}} \frac{C_{\mathrm{bc}}}{C_{\mathrm{s}}} - \frac{j R_{\mathrm{bi}} C_{\mathrm{f}}}{\omega R_{\mathrm{bc}} C_{\mathrm{s}}^2} \tag{3.64}$$

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} + (1-\alpha) R_{\mathrm{bi}} \left(\frac{C_{\mathrm{f}}}{C_{\mathrm{s}}} \right) \tag{3.65}$$

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + \frac{1}{j\omega C_{\mathrm{s}}} + \frac{1}{\omega^2 R_{\mathrm{bc}} C_{\mathrm{s}}} - \frac{R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}}{C_{\mathrm{s}}^2}. \tag{3.66}$$

As indicated in the above expressions, initially $1/\omega^2 (C_{\mathrm{f}} + C_{\mathrm{bc}})^2 R_{\mathrm{bc}}$ is much larger than $R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}} / C_{\mathrm{s}}^2$ and R_{c}. $R_{\mathrm{e}}(Z_{22} - Z_{21})$, where R_{e} denotes the real part can hence be used to extract R_{bc}. With increase of frequency, the second term decreases rapidly and $\mathrm{Re}(Z_{22} - Z_{21})$ is possible to be negative as described in [20], where the intrinsic base resistance R_{bi} is much higher in InP-based HBTs than in GaAs-based HBTs. After the higher-order terms are discarded, Equations (3.64) and (3.65) are reduced to

$$Z_{11} - Z_{12} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} + R_{\mathrm{bi}} \frac{C_{\mathrm{bc}}}{C_{\mathrm{s}}} \tag{3.67}$$

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} \tag{3.68}$$

The fifth term in Equation (3.65) consists of only a very small fraction of the real part of Z_{12}; therefore, the approximation is generally reasonable. However, for InP-based HBTs it may account for up to 18% of the real part of Z_{12} due to the large base resistance [20]. Self-consistent iterations can be used to correct the approximation errors [20].

The very high frequency range is defined by $\omega C_{\mathrm{bc}} R_{\mathrm{bi}}, \omega C_{\mathrm{f}} R_{\mathrm{bi}} \gg 1$

$$\begin{aligned} \Re &\simeq \frac{Z_{\mathrm{F}} Z_{\mathrm{BC}}}{R_{\mathrm{bi}}} \times \frac{1}{1 + \dfrac{Z_{\mathrm{F}} + Z_{\mathrm{BC}}}{R_{\mathrm{bi}}}} \\ &\simeq -\frac{1}{\omega^2 C_{\mathrm{bc}} C_{\mathrm{f}} R_{\mathrm{bi}}} \times \left(1 - \frac{Z_{\mathrm{F}} + Z_{\mathrm{BC}}}{R_{\mathrm{bi}}} \right) \\ &\simeq -\frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} + \frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} \left(\frac{1}{j\omega C_{\mathrm{f}} R_{\mathrm{bi}}} + \frac{1}{j\omega C_{\mathrm{bc}} R_{\mathrm{bi}}} \right). \end{aligned} \tag{3.69}$$

The typical very high frequency is calculated by the values given above: $f = 10/(2\pi R_{\mathrm{bi}} C_{\mathrm{bi}}) = 63.8\,\mathrm{GHz}$. Since the very high frequency is very hard to reach, we normally relax the high frequency requirement to above 20 GHz. Substituting Equation (3.69) into Equations (3.51), (3.52) and (3.54), we obtain

$$Z_{11} - Z_{12} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} + \frac{1}{j\omega C_{\mathrm{f}}} + \frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} \left(\frac{C_{\mathrm{bc}}}{C_{\mathrm{f}} + 1} \right) \tag{3.70}$$

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} + \frac{(1-\alpha)}{j\omega C_{\mathrm{bc}}} + \frac{(1-\alpha)}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} \left(\frac{C_{\mathrm{f}}}{C_{\mathrm{bc}}} + 1 \right) \tag{3.71}$$

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} - \frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} + \frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}} C_{\mathrm{f}}} \left(\frac{1}{j\omega C_{\mathrm{f}} R_{\mathrm{bi}}} + \frac{1}{j\omega C_{\mathrm{bc}} R_{\mathrm{bi}}} \right). \tag{3.72}$$

We divide the frequency range into three subdivisions. The boundaries of these three ranges are not strict and highly dependent on the applied biases, process technology and device design. The magnitudes of R_{bc}, R_{bi}, C_{F}, C_{bc} and r_{e} affect the definitions of the frequency range. The network analyser is used to measure S-parameters and then the S-parameters are converted to Z-parameters. The lowest frequency of the network analyser is about several tenths of MHz. Therefore, S-parameter data at the low frequency (satisfied by the condition $\omega R_{\mathrm{bc}} C_{\mathrm{s}} \ll 1$) are not of much use for analysis except in the reverse or deep saturation region.

For the analytical approach to parameter extraction, we clarify that the intermediate frequency range is below 5 GHz and the high frequency range is greater than 10 GHz. The method of clarifying the frequency range before extracting the element parameters is arbitrary; however, the justification needs to be made during the procedure of parameter extraction.

3.8.2 Z-Parameters at zero bias or 'cold' biases

At zero bias ($V_{\mathrm{BE}} = 0\,\mathrm{V}$ and $V_{\mathrm{CE}} = 0\,\mathrm{V}$) or 'cold' bias ($V_{\mathrm{BE}} = 0\,\mathrm{V}$ and $I_{\mathrm{B}} = 0\,\mathrm{A}$; 'cold' is adopted from the MESFET modelling techniques), the current gain becomes zero and the device behaves like a passive component. The base–emitter resistance r_{e} is also much larger. The magnitude of this resistance is normally greater than $10^4\ \Omega$, which is within the same order of magnitude as the collector–base resistance. Therefore, the equivalent circuit becomes much simpler and so are the Z-parameters. With $\alpha = 0$, we have

$$Z_{11} - Z_{12} = Z_{\mathrm{B}} + \frac{Z_{\mathrm{F}} R_{\mathrm{bi}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}} \tag{3.73}$$

$$Z_{12} = Z_{21} = Z_{\mathrm{E}} + \frac{Z_{\mathrm{BC}} R_{\mathrm{bi}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}} \tag{3.74}$$

$$Z_{22} - Z_{21} = Z_{\mathrm{C}} + \frac{Z_{\mathrm{F}} Z_{\mathrm{BC}}}{Z_{\mathrm{BC}} + Z_{\mathrm{F}} + R_{\mathrm{bi}}}. \tag{3.75}$$

The approximation of the Z-parameters is almost the same as before except for Z_{12} and Z_{21}. The approximation of Z_{E} in this case will be rather different. The Z_{E} at the low frequency range in which $\omega r_{\mathrm{e}} C_{\mathrm{be0}} \ll 1$ is satisfied is given by

$$Z_{\mathrm{E}} \simeq R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be0}}. \tag{3.76}$$

Substituting into Equation (3.74), Z_{12} can be written as

$$Z_{12} \simeq R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be0}} + j\omega R_{\mathrm{bi}} C_{\mathrm{f}} (R_{\mathrm{bc}} - j\omega R_{\mathrm{bc}}^2 C_{\mathrm{s}}). \tag{3.77}$$

For simplicity, higher-order terms can be discarded in the above equation without affecting the accuracy.

When the frequency increases to the point at which $\omega r_{\mathrm{e}} C_{\mathrm{be0}} \gg 1$, Z_{E} is approximated as

$$Z_{\mathrm{E}} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + \frac{1}{j\omega C_{\mathrm{be0}}} + \frac{1}{\omega^2 r_{\mathrm{e}} C_{\mathrm{be}}^2}. \tag{3.78}$$

Actually, the assumption of $\omega r_{\mathrm{e}} C_{\mathrm{be0}}$, $\omega R_{\mathrm{bc}} C_{\mathrm{bc0}} \gg 1$ and $\omega C_{\mathrm{be0}} R_{\mathrm{bi}}$, $\omega C_{\mathrm{bc0}} R_{\mathrm{bi}} \ll 1$ is easily satisfied in the frequency range of interest; therefore,

$$Z_{11} - Z_{12} = R_{\mathrm{bx}} + j\omega L_{\mathrm{b}} + \frac{C_{\mathrm{bc}}}{C_{\mathrm{s}}} R_{\mathrm{bi}} \tag{3.79}$$

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + \frac{1}{j\omega C_{\mathrm{be0}}} + \frac{1}{\omega^2 r_{\mathrm{e}} C_{\mathrm{be}}^2} + \frac{C_{\mathrm{f}}}{C_{\mathrm{s}}} R_{\mathrm{bi}} \tag{3.80}$$

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + \frac{1}{j\omega C_{\mathrm{s}}} + \frac{1}{\omega^2 R_{\mathrm{bc}} C_{\mathrm{s}}^2}. \tag{3.81}$$

In the above equations, the higher-order terms have been discarded for simplicity. If the frequency increases further and $1/(\omega C_{\mathrm{be0}})$, $1/(\omega C_{\mathrm{bc0}}) \ll R_{\mathrm{bi}}$, then Z_{12} is given by

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + \frac{1}{j\omega C_{\mathrm{be0}}} + \frac{1}{\omega^2 r_{\mathrm{e}} C_{\mathrm{be}}^2} + \frac{1}{j\omega C_{\mathrm{bc}}} + \frac{1}{\omega^2 R_{\mathrm{bi}} C_{\mathrm{bc}}} \left(\frac{1}{C_{\mathrm{bc}}} + \frac{1}{C_{\mathrm{f}}} \right). \tag{3.82}$$

In the case of both junctions being reverse-biased, the approximation of the Z-parameters might be rather similar to this case at zero bias.

3.8.3 Parameter extraction

The model parameter extractions are based on the assumption that all the extrinsic and parasitic elements are bias-independent and only elements in the intrinsic part of the device vary with bias. This also means that the extrinsic and parasitic elements can be extracted by using multiple bias information. Some element values are sensitive to the bias point at which the extraction procedure is carried out.

Since the conditions at the intermediate frequency are easily satisfied in the frequency measurement range, we should make the most use of the Z-parameters in the intermediate frequency range. The extraction procedure is shown in Figures 3.46–3.49. The information for extracting the element parameters is overwhelming; hence different schemes can be developed. The extraction procedure shown here is just one of them. The superscripts used are explained as follows: R represents the real part of the corresponding Z-parameters; I the imaginary part; F the forward bias; 0 the zero bias; L the low frequency range; M the intermediate frequency range; and H the high frequency range.

From $Z_{11} - Z_{12}$, we extract L_{b} by plotting $\mathrm{Im}(Z_{11} - Z_{12})/\omega$ versus frequency in high frequency range. $R_{\mathrm{bx}} + R_{\mathrm{bi}}$ can be obtained from $\mathrm{Re}(Z_{11} - Z_{12})$ at the extreme low frequency. R_{bx} might be obtained from $\mathrm{Re}(Z_{11} - Z_{12})$ at the high frequency. The difference will be R_{bi}.

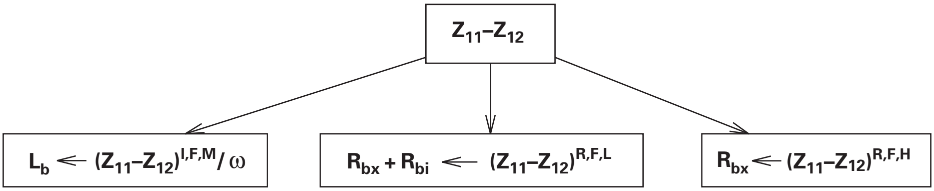


Fig. 3.46 The element values extracted from $Z_{11} - Z_{12}$ (B. Li and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 5, pp. 534–539, May 1999. ©1999 IEEE).

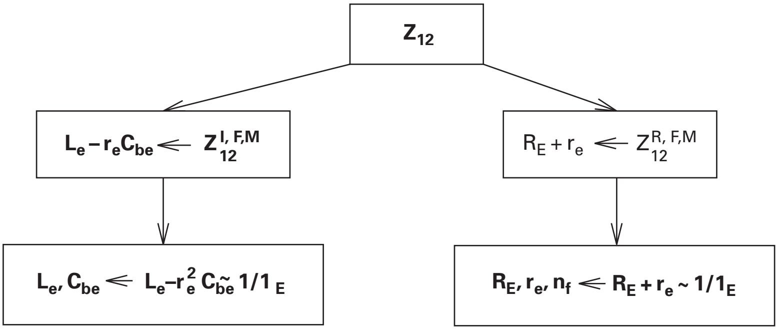


Fig. 3.47 The element values extracted from Z_{12} (B. Li and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 5, pp. 534–539, May 1999. ©1999 IEEE).

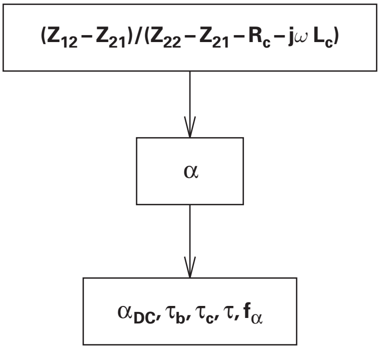


Fig. 3.48 The element values extracted from $(Z_{12} - Z_{21})/(Z_{22} - Z_{21} - R_c - j\omega L_c)$ (B. Li and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 5, pp. 534–539, May 1999. ©1999 IEEE).

Figure 3.47 shows how to extract R_E, r_e, L_e and C_{be}. $R_E + r_e$ is easily obtained from the real part of Z_{12} in the low or intermediate frequency range and we have

$$R_E + r_e = R_E + \frac{n_f kT}{qI_E}. \tag{3.83}$$

Therefore, we plot the curve $R_E + r_e$ versus $1/I_E$. R_E, r_e and n_f should be easily extracted. In the case of high collector currents, the fraction of the depletion capacitance

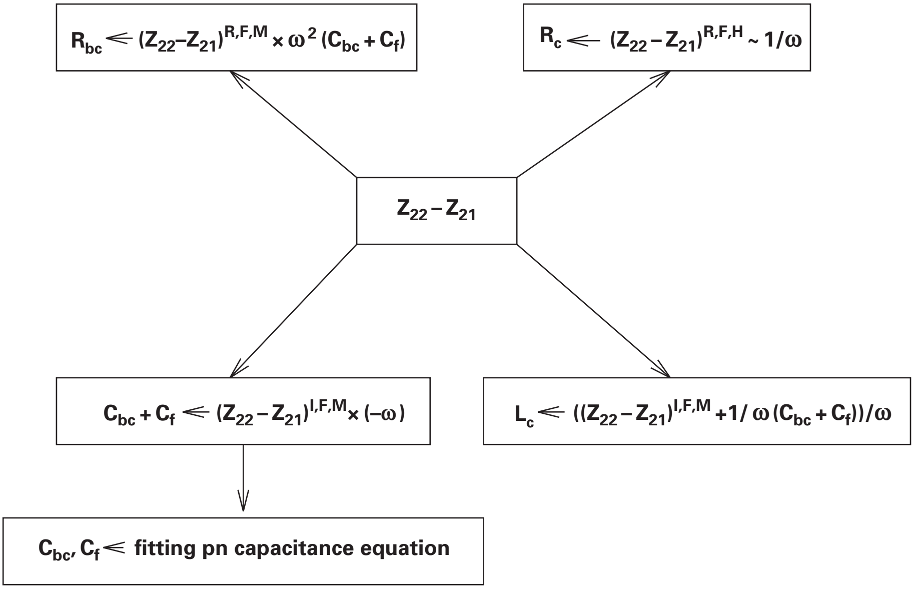

$R_{bc} \Leftarrow (Z_{22}–Z_{21})^{R,F,M} \times \omega^2 (C_{bc} + C_f)$
$R_c \Leftarrow (Z_{22} – Z_{21})^{R,F,H} \sim 1/\omega$
$Z_{22} – Z_{21}$
$C_{bc} + C_f \Leftarrow (Z_{22} – Z_{21})^{I,F,M} \times (–\omega)$
$L_c \Leftarrow ((Z_{22} – Z_{21})^{I,F,M} +1/\omega (C_{bc} + C_f))/\omega$
$C_{bc}, C_f \Leftarrow$ fitting pn capacitance equation

Fig. 3.49 The element values extracted from $Z_{22} - Z_{21}$ (B. Li and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 5, pp. 534–539, May 1999. ©1999 IEEE).

in the base–emitter capacitance C_{be} is small and C_{be} can be approximated to be proportional to I_{E}, and we also have $r_{\text{e}} \propto 1/I_{\text{E}}$. Therefore, the y intercept of plot of $L_{\text{e}} - r_{\text{e}}^2 C_{\text{be}}$ versus $1/I_{\text{E}}$ gives the value of L_{e}. Once L_{e} is known, C_{be} can be easily calculated.

The extraction from $Z_{22} - Z_{21}$ is somewhat more complex. R_{bc} is extracted from the expression $\text{Re}(Z_{22} - Z_{21})\omega^2 C_{\text{s}}$ at the intermediate frequency or from the real part of $Z_{22} - Z_{21}$ at the low frequency. R_{c} is the vertical axis intercept of the plot of $\text{Re}(Z_{22} - Z_{21})$ versus $1/\omega^2$ in the high frequency range. $C_{\text{f}} + C_{\text{bc}}$ is extracted from $(Z_{22} - Z_{21}) \times (-\omega)$ at the intermediate frequency. There are two different assumptions about C_{f}. One assumes that C_{f} is bias-independent. Therefore, we can obtain C_{f} and C_{bc} by fitting $C_{\text{f}} + C_{\text{bc}}$ into the following bias-dependent equation:

$$C_{\text{f}} + C_{\text{bc}} = C_{\text{f}} + \frac{C_{\text{jbc0}}}{\left(1 - \dfrac{V_{\text{BC}}}{V_{\text{jbc}}}\right)^{M_{\text{jbc}}}}. \tag{3.84}$$

On the other hand, C_{f} might be considered to vary with the bias rather than be fixed as in the Gummel–Poon model. The fraction of C_{f} in the base–collector capacitance is constant. For this case, given R_{bx}, R_{bi} and $C_{\text{bc}}/C_{\text{s}}$ can be obtained from the real part of $Z_{11} - Z_{12}$ at the intermediate frequency or calculated from a knowledge of the geometry. We can calculate C_{bc} and C_{f} once the fraction is given.

L_{c} is calculated by the expression $\text{Im}[(Z_{22} - Z_{21}) + 1/(\omega(C_{\text{bc}} + C_{\text{f}}))]/\omega$ in the high frequency range.

α can be extracted easily once we know R_{c} and L_{c}. This is shown in Figure 3.47. The extraction of parameters related to α has been reported by Pehlke and Pavlidis [19].

3.8.4 The approximation at $R_{\mathrm{bi}} = 0$

This is the case in which the distributed effect of the base–collector region is not necessarily taken into account. The equivalent circuit is reduced and so are the Z-parameters. Let $R_{\mathrm{bi}} = 0$ in Equations (3.51)–(3.54), we have

$$Z_{11} - Z_{12} = R_{\mathrm{b}} + j\omega L_{\mathrm{b}} \tag{3.85}$$

$$Z_{12} = R_{\mathrm{E}} + j\omega L_{\mathrm{e}} + r_{\mathrm{e}} - j\omega r_{\mathrm{e}}^2 C_{\mathrm{be}} \tag{3.86}$$

$$Z_{12} - Z_{21} = \alpha \times Z_{\mathrm{BC}} \tag{3.87}$$

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + Z_{\mathrm{BC}}. \tag{3.88}$$

The element parameter extraction becomes trivial in this particular situation.

In the extreme low frequency range,

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + R_{\mathrm{bc}} - jR_{\mathrm{bc}}^2 C_{\mathrm{bc}}. \tag{3.89}$$

In the intermediate frequency range,

$$Z_{22} - Z_{21} = R_{\mathrm{C}} + j\omega L_{\mathrm{c}} + \frac{1}{j\omega C_{\mathrm{bc}}} + \frac{1}{\omega^2 R_{\mathrm{bc}} C_{\mathrm{bc}}^2}. \tag{3.90}$$

R_{b} and L_{b} are given by Equation (3.85). Equation (3.86) gives $R_{\mathrm{E}}, r_{\mathrm{e}}, L_{\mathrm{e}}$ and C_{be}. Equation (3.88) gives $R_{\mathrm{c}}, R_{\mathrm{bc}}, C_{\mathrm{bc}}$ and L_{c}. The extraction of α is the same as described above.

The T small-signal equivalent circuit for HBTs after de-embedding pad capacitances has been used to derive the Z-parameter expressions. The simplification of the Z-parameters based on the range of measured frequencies was developed in order to provide guidelines for directly extracting element parameters. The information from the measured S-parameters is normally overwhelming and different schemes of extraction procedure could be developed from these approximations. One fully analytical extraction procedure has been provided here. Z-parameters at multiple biases are utilised to extract the small-signal parameters. These results can also be used for the parameter extractions of BJTs.

3.9 Small-signal model of the collector-up (inverted) HBT

The device used in this example is a $5 \times 10\,\mu\mathrm{m}^2$ InGaAs/InAlAs/InP inverted HBT with $f_{\mathrm{T}} = 23\,\mathrm{GHz}$ and $f_{\max} = 20\,\mathrm{GHz}$. The small-signal equivalent circuit of the device [17, 18] is shown in Figure 3.50.

$L_{\mathrm{b}}, L_{\mathrm{c}}$ and L_{e} and $C_{\mathrm{pbe}}, C_{\mathrm{pbc}}$ and C_{pce} are parasitic inductances and capacitances respectively, and $R_{\mathrm{b}}, R_{\mathrm{c}}$ and R_{e} are extrinsic resistances. The active portion of the HBT is modelled by intrinsic elements $C_{\mathrm{e}}, r_{\mathrm{e}}, C_{\mathrm{jc}}, \alpha$ and R_{jc}, where $\alpha = \alpha_{\mathrm{F}}/[1 + j(f/f_\alpha)]e^{-\mathrm{j}\omega\tau}$. α_{F} is the dc value of the transport factor, τ is the transit time of collector current and f_α is the α 3 dB frequency. RF-measurements indicate that the HBT under zero bias ($I_{\mathrm{b}} = 0\,\mathrm{A}$, $V_{\mathrm{CE}} = 0\,\mathrm{V}$) can be represented by a passive network. Therefore, in

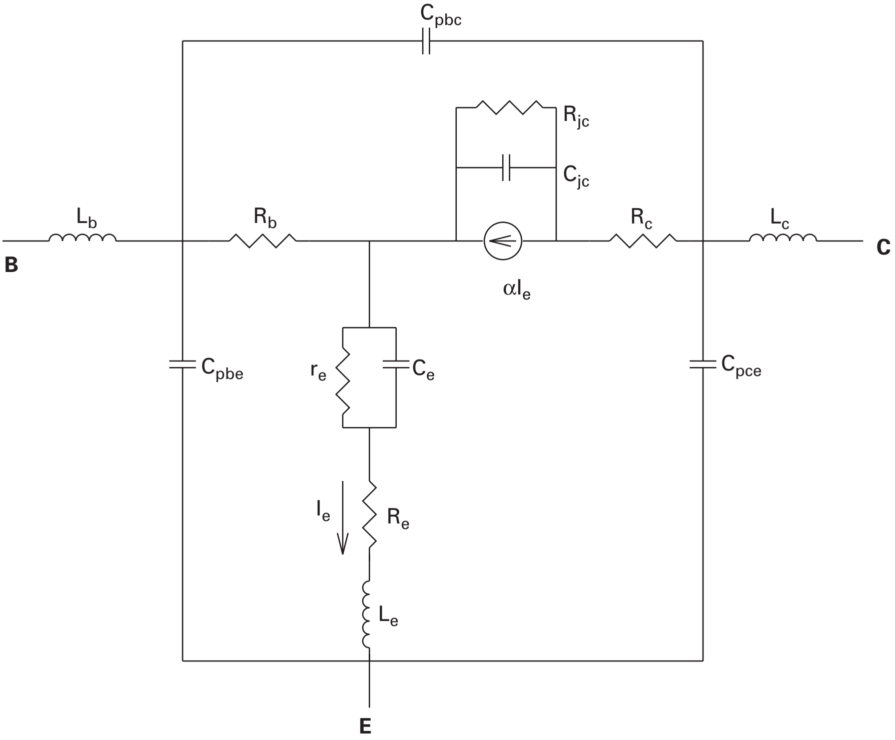


Fig. 3.50 Small-signal equivalent circuit of the inverted HBT (B. Li and S. Prasad, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 45, No. 7, pp. 1135–1137, July 1997. ©1997 IEEE).

this case, the transport factor α is negligible. Moreover, the dynamic resistance of the base–collector p–n junction and the base–emitter p–n junction is assumed to be very large. The uncertainty of numerical optimisation in zero bias can be reduced. The initial values that are assumed in optimisation are calculated from the physical and geometrical parameters. The parasitic elements, C_{pbe}, C_{pbc}, C_{pce}, L_{b}, L_{e} and L_{c}, are obtained from zero bias numerical optimisation and assumed to be invariant with bias [18]. Their values are listed in Table 1.

The other elements under non-zero bias are extracted by the following analytical approach:

- Convert the S-parameters to Z-parameters and remove the parasitic series elements L_{b} and L_{c};
- Convert the Z-parameters to Y-parameters and remove the parasitic shunt elements C_{pbe}, C_{pce} and C_{pbc};
- Convert the Y-parameters to h-parameters.

The elements of the equivalent circuit, excluding the parasitic effects, are easily extracted using the procedure described in [19].

The elements R_{b}, L_{e}, R_{e} and R_{c} are basically constant over the entire frequency range of interest and do not show significant variation with bias. Therefore, these elements can be considered to be fixed. The bias-dependent elements are C_e, C_{jc}, R_{jc}, α_F, τ, f_α and r_e. The consideration of the bias variation of these elements is sufficient for accurate small-signal modelling [17, 18].

3.10 Problems

(1) The scattering parameters of the HEMT in the common source configuration are given. The Y-parameters are related to the equivalent circuit parameters as follows:

$$Y_{11} + Y_{12} = j\omega C_{GS} \tag{3.91}$$

$$-Y_{12} = j\omega C_{GD} \tag{3.92}$$

$$Y_{22} + Y_{12} = g_D + j\omega C_{DS} \tag{3.93}$$

$$Y_{21} - Y_{12} = \frac{g_m}{1 + j\left(\frac{\omega}{\omega_0}\right)\tau_1}, \tag{3.94}$$

where

$$\tau_1 = \omega_0 \frac{-C_{DS}}{g_D} \tag{3.95}$$

$$\omega_0 = \frac{\mu_n (V_{GS} - V_T)}{L^2}. \tag{3.96}$$

The scattering parameters were measured at $V_{GS} - V_T = 0.4\,\text{V}$. The mobility $\mu_n = 4400\,\text{cm}^2/\text{V-s}$ and $L = 1\,\mu\text{m}$. $V_{DS} = 0.5\,\text{V}$. $g_m = 0.049\,\text{S}$, $g_D = 0.014\,\text{S}$.

Use the SA algorithm to obtain optimised values of the equivalent circuit after converting the scattering parameters to admittance parameters [14] as shown in Table 3.7.

(2) What are the advantages of combinatorial optimisation methods over analytical methods of device optimisation? What are the disadvantages?

(3) A researcher wants to optimise the parameters of a small-signal equivalent circuit for the HBT using the genetic algorithm. The parameters are shown in Table 3.8. Assume that there is a procedure for measuring the S-parameters of the HBT for each variation of the parameters of the circuit shown in the table.
(a) Write down the steps of a genetic algorithm for the optimisation.
(b) Write a fitness function for the genetic algorithm.
(c) Modify this genetic algorithm for optimising the small-signal parameters of a FET. Compare the difficulty of doing this to the difficulty of modifying an analytic method for the same procedure.

(4) A student wants to create a neural network model for a power amplifier with the parameters shown in Table 3.9. He/she decides to use the SGA to simultaneously optimise both the weight values and the number of neurons in the neural network.

Table 3.7 HEMT scattering parameters

Frequency (GHz)	S_{11}	S_{12}	S_{21}	S_{22}
5	$0.55\angle-158$	$0.05\angle 1.80$	$0.53\angle 14.5$	$0.83\angle 176$
10	$0.75\angle-166$	$0.06\angle 180$	$0.63\angle 18.5$	$0.82\angle 173$
20	$0.8\angle 179$	$0.07\angle 26$	$0.68\angle 28.5$	$0.82\angle 167$
30	$0.826\angle 169$	$0.07\angle 35$	$0.079\angle 33.5$	$0.83\angle 161$
40	$0.79\angle 161$	$0.009\angle 33$	$0.09\angle 34.5$	$0.81\angle 155$

Table 3.8 Circuit parameters

Component	Definition
L_b(pH)	Base inductance
L_c(pH)	Collector inductance
L_e(pH)	Emitter inductance
C_c(fF)	Collector capacitance
C_e(pF)	Emitter capacitance
$R_\mathrm{b}(\Omega)$	Base resistance
$R_\mathrm{c}(\Omega)$	Collector resistance
τ(ps)	Transit time
β_o	Current gain

Table 3.9 Neural network inputs and outputs

Inputs	Outputs
Gate length	Output power
Gate voltage	DC gate current
DC drain current	Drain current

(a) Draw a diagram of a chromosome that can be used by the SGA.

(b) Is there any other way to simultaneously optimise both the weight values and the number of neurons in the network?

(5) The scattering parameters for an HBT are given in Table 3.10.

Use the semi-analytical parameter extraction method to determine the element values in the equivalent circuit and then obtain the scattering parameters to compare with the given measured parameters.

(6) Assume that the base–collector extrinsic capacitance could be lumped into the intrinsic C_bc and develop the strategy for the small-signal model parameter extraction procedure.

Table 3.10 HBT S-parameters

Frequency (GHz)	$\lvert S_{11}\rvert$	$\angle S_{11}$	$\lvert S_{21}\rvert$	$\angle S_{21}$	$\lvert S_{12}\rvert$	$\angle S_{12}$	$\lvert S_{22}\rvert$	$\angle S_{22}$
0.45	0.955	−4.3	4.651	174.83	0.008	85.1	0.997	−2.78
0.5	0.954	−4.83	4.648	174.25	0.009	84.2	0.996	−3.1
0.6	0.953	−5.78	4.638	173.19	0.01	84.11	0.996	−3.73
0.7	0.952	−6.71	4.631	172.13	0.012	83.44	0.995	−4.32
0.8	0.951	−7.66	4.622	171.05	0.013	82.7	0.994	−4.93
0.9	0.95	−8.68	4.614	170.05	0.015	82.79	0.993	−5.51
1	0.948	−9.52	4.601	168.94	0.016	82.61	0.992	−6.12
2	0.92	−18.95	4.457	158.29	0.034	75.73	0.974	−12.12
4	0.833	−35.63	4.005	139.02	0.062	66.77	0.921	−22.78
6	0.726	−49.36	3.482	122.93	0.086	55.57	0.853	−31.24
8	0.628	−60.18	3.001	109.46	0.102	48.19	0.797	−38.1
10	0.544	−68.72	2.593	98.35	0.115	42.08	0.75	−43.67
12	0.476	−74.48	2.264	89.05	0.123	37	0.716	−48.35
14	0.43	−79.5	2.009	81.24	0.132	34.11	0.694	−52.23
16	0.39	−84.14	1.805	73.81	0.141	30.94	0.68	−56.36
18	0.352	−87.56	1.633	66.91	0.149	27.43	0.669	−60.25
20	0.326	−88.9	1.489	61.32	0.153	25.01	0.663	−63.83
22	0.306	−89.49	1.382	55.44	0.158	22.24	0.661	−67.38
24	0.299	−90.98	1.293	50.96	0.163	20.41	0.662	−70.79
26	0.292	−94.44	1.227	46.22	0.167	19.91	0.672	−73.97
28	0.283	−95.78	1.173	41.1	0.176	18.33	0.68	−77.75
30	0.272	−98.22	1.127	35.42	0.184	16.84	0.69	−81.95
32	0.2569	−99.7	1.073	29.92	0.194	14.21	0.688	−85.81
34	0.246	−102.13	1.024	25.25	0.201	11.68	0.692	−89.2
36	0.238	−100.66	0.974	21.13	0.206	8.92	0.697	−92.91
38	0.232	−101.63	0.963	16.41	0.213	6.63	0.699	−96.4
40	0.223	−101.26	0.918	11.77	0.22	3.11	0.694	−99.63

(7) What are the advantages of the inverted (also known as the *emitter-down* or *collector-up*) HBT? If it is promising in some power application, why is it not popular yet? Compare the emitter-up HBT and the emitter-down HBT in terms of the device parameters?

(8) Use the small-signal equivalent circuit of the collector-up HBT (Figure 3.50) to determine the basic expressions for the circuit parameters.

(9) The base–collector capacitance C_f is much smaller in the inverted HBT. Does that make the model parameter extraction easy?

(10) In compact device modelling, layout pads for measurement are not the part of the device which needs to be removed (De-embedding procedure) from the measurement data. In industry, open/short structures are used.

(a) Design your own open/short structure. How do you think the pad size, ground layout, probe type or frequency affect your design?

(b) Draw the equivalent circuit for the open structure.

(c) Draw the equivalent circuit for the short structure.

(d) The S-parameters for the open/short structures are given. What information can you extract from them?

(11) An engineer is developing the diode model at a particular bias. Outline the procedure he/she is going to develop. Use the given S-parameter file. If there is inductance in the structure that he/she forgets to remove, how will that affect the final result? If the resistor and inductor are ignored, do you get the right result? Explain your answer.

(12) There is a set of S-parameter files which were taken for the diode. Use the files to extract the diode capacitance model parameters, assuming that parasitic elements have been de-embedded out.

(13) Given the HBT S-parameters measured at the cold condition, use what you learned to extract the model parameters.

(14) Given the HBT S-parameters measured at the constant V_{ce} condition, use what you learned to extract the model parameters.

(15) Given the HBT S-parameters measured at constant I_b condition, use what you learned to extract the model parameters.

(16) The intrinsic part of the small-signal equivalent circuit of the MESFET which is similar to the hybrid pi model of the bipolar transistor is given in Figure 3.51. Outline the extraction procedure to extract the model parameters assuming negligible g_{ds}. Now the full equivalent circuit of the MESFET is given in Figure 3.52. Assume $L_g = L_d = 25\,\text{pH}$, $L_s = 10\,\text{pH}$, $R_s = R_d = 10\,\Omega$, $R_g = 1\,\Omega$,

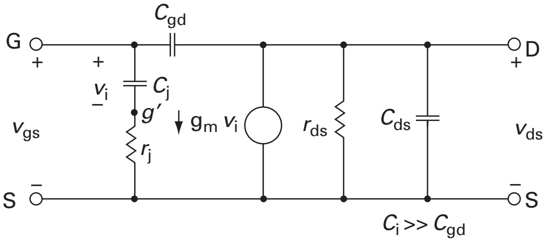


Fig. 3.51 Intrinsic MESFET circuit.

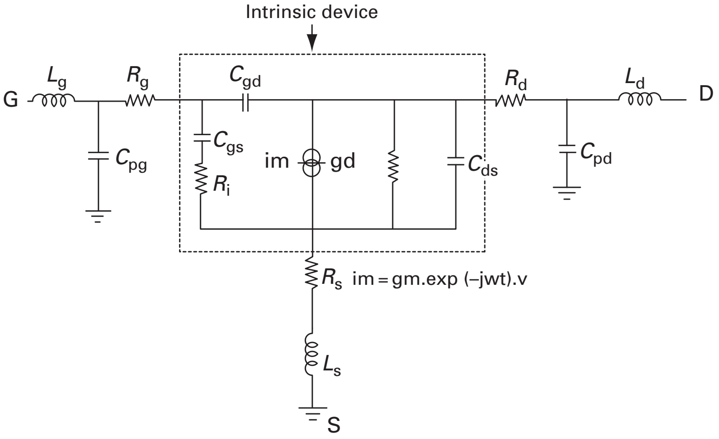


Fig. 3.52 MESFET equivalent circuit.

$C_{\mathrm{pg}} = C_{\mathrm{pd}} = 25\,\mathrm{fF}$ and $g_{\mathrm{ds}} = 0\,\mathrm{S}$. Extract the other model parameters based on the given set of S-parameters.

Note that the necessary data files are available on the Web labelled according to the problem numbers.

References

[1] Bousnina S., Falt C., Mandeville P., Kouki A. B., Ghannouchi F. M. (2002). An accurate on-wafer deembedding technique with application to HBT devices characterizaion. *IEEE Trans. Microw. Theory Tech. MTT-50*, 2 (February), 420–424.

[2] Bousnina S., Mandeville P., Kouki A. B., Surridge R., Ghannouchi F. M. (2002). Direct parameter-extraction method for HBT small signal model. *IEEE Trans. Microw. Theory Tech. MTT-50*, 2 (February), 529–536.

[3] Costa D., Liu W., Jr. J. S. H. (1991). Direct extraction of the AlGaAs/GaAs heterojunction bipolar transistor small-signal equivalent circuit. *IEEE Trans. Electron Devices ED-38*, 9 (September), 2018–2024.

[4] Dasgupta D., McGregor D. (1992). SGA: a structured genetic algorithm. Tech. Rep. IKBS-8-92, University of Strathclyde.

[5] Degachi L., Ghannouchi F. M. (2006). Systematic and rigorous extraction method of HBT small-signal model parameters. *IEEE Trans. Microw. Theory Tech. MTT-54*, 2 (February), 682–688.

[6] Dvorak M. W., Bolognesi C. (2003). On the accuracy of direct extraction of the heterojunction-bipolar-transistor equivalent-circuit model parameters C_π, C_{BC} and R_{E}. *IEEE Trans. Microw. Theory Tech. MTT-51*, 6 (June), 1640–1649.

[7] Gobert Y., Tasker P. J., Bachem K. H. (1997). A physical, yet simple, small-signal equivalent circuit for the heterojunction bipolar transistor. *IEEE Trans. Microw. Theory Tech. MTT-45*, 1 (January), 149–153.

[8] Goldberg D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

[14] Gonzalez G. (1997). *Microwave Transistor Amplifiers, Analysis and Design*. Prentice Hall.

[10] Holland J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

[11] Hopfield J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci. 81*, 10 (May), 3088–3092.

[12] Kirkpatrick S., Gelatt C. D. Jr, Vecchi M. P. (1983). Optimization by simulated annealing. *Science 220*, 4598 (May), 671–680.

[13] Li B., Prasad S. (1997). Harmonic and two-tone intermodulation distortion analyses of the inverted InGaAs/InAlAs/InP HBT. *IEEE Trans. Microw. Theory Tech. MTT-45*, 7 (July), 1135–1137.

[14] Li B., Prasad S. (1999). Basic expressions and approximations in small-signal parameter extraction for HBT's. *IEEE Trans. Microw. Theory Tech. MTT-47*, 5 (May), 534–539.

[15] Li B., Prasad S., Yang L.-W., Wang S. C. (1998). A semianalytical parameter-extraction procedure for HBT equivalent circuit. *IEEE Trans. Microw. Theory Tech. MTT-46*, 10 (October), 1427–1435.

[16] Lippmann R. (1987). An introduction to computing with neural nets. *IEEE ASSP Mag. 4*, 5 (April), 4–22.

[17] Meskoob B., Prasad S., Vai M., Fonstad C., Vlcek J. C., Sato H. (1992). A small-signal equivalent circuit for the collector-up InGaAs/InAlAs/InP heterojunction bipolar transistor. *IEEE Trans. Electron Devices ED-39*, 11 (November), 2629–2632.

[18] Meskoob B., Prasad S., Vai M., Vlcek J. C., Sato H., Fonstad C. G. (1992). Bias-dependence of the intrinsic element values of InGaAs/InAlAs/InP inverted heterojunction bipolar transistor. *IEEE Trans. Microw. Theory Tech. MTT-40*, 5 (May), 1012–1014.

[19] Pehlke D. R., Pavlidis D. (1992). Evaluation of the factors determining HBT high frequency performance by direct analysis of S-parameter data. *IEEE Trans. Microw. Theory Tech. MTT-40*, 12 (December), 2367–2373.

[20] Rios J. M. M., Lunardi L. M., Chandrasekhar S., Miyamoto Y. (1997). A self-consistent method for complete small-signal extraction of InP-based heterojunction bipolar transistors (HBTs). *IEEE Trans. Microw. Theory Tech. MTT-45*, 1 (January), 39–45.

[21] Roblin P., Kang S., Ketterson A., Morkoc H. (1987). Analysis of MODFET microwave characteristics. *IEEE Trans. Electron Devices ED-34*, 9 (September), 1919–1927.

[22] Romeo F., Sangiovanni-Vincentelli A. (1987). Probabilistic hill climbing algorithms: properties and applications. In *1985 Chapel Hill Conference on VLSI*, H. Fuchs, ed. Computer Science Press, 671–680.

[23] Rutenbar R. A. (1989). Simulated annealing algorithm: an overview. *IEEE Circ. Dev. Mag. 5*, 1 (January), 19–26.

[24] Samelis A., Pavlidis D. (1997). DC to high-frequency HBT-model parameter evaluation using impedance block conditioned optimization. *IEEE Trans. Microw. Theory Tech. MTT-45*, 6 (June), 886–897.

[25] Schaper U., Holzapfl B. (1995). Analytical parameter extraction of the HBT equivalent circuit with T-like topology from measured S-parameters. *IEEE Trans. Microw. Theory Tech. MTT-40*, 3 (March), 493–498.

[26] Sheinman B., Wasige E., Rudolph M., Sidorov V., Cohen S., Ritter D. (2002). A peeling algorithm for extraction of the HBT small-signal equivalent circuit. *IEEE Trans. Microw. Theory Tech. MTT-50*, 12 (December), 2804–2810.

[27] Spiegel S. J., Ritter D., Hamm R. A., Feygenson A., Smith P. R. (1995). Extraction of the InP/GaInAs heterojunction bipolar transistor small-signal equivalent circuit. *IEEE Trans. Electron Devices ED-42*, 6 (June), 1059–1064.

[28] Vai M., Prasad S. (1993). Automatic impedance matching with a neural network. *IEEE Microw. Guided Wave Lett. MGWL-3*, 10 (October), 353–354.

[29] Vai M., Prasad S. (1999). Neural networks in microwave circuit design – beyond black-box models. *Int. J. RF and Microwave CAE 9*, 3 (May), 187–197.

[30] Vai M., Prasad S., Li N., Kai F. (1989). Modeling of microwave semiconductor devices using simulated annealing optimization. *IEEE Trans. Electron Devices ED-36*, 4 (April), 761–762.

[31] Vai M., Prasad S. (1995). Microwave circuit analysis and design by a massively distributed computing network. *IEEE Trans. Microw. Theory Tech. MTT-43*, 5 (May), 1087–1094.

[32] vanRooij A., Jain L. C., Johnson R. P. (1996). *Neural Network Training Using Genetic Algorithms*. World Scientific.

[33] Wei C. J., Huang J. C. M. (1995). Direct extraction of equivalent circuit parameters for heterojunction bipolar transistors. *IEEE Trans. Microw. Theory Tech. MTT-43*, 9 (September), 2035–2039.

[34] Willén B., Rohner M., Schnyder I., Jäckel H. (2002). Improved automatic parameter extraction of InP-HBT small signal equivalent circuits. *IEEE Trans. Microw. Theory Tech. MTT-50*, 2 (February), 580–583.

[35] Yang T. -R., Tsai J. M. -L., Ho C. -L., Hu R. (2007). SiGe HBTs small-signal pi modeling. *IEEE Trans. Microw. Theory Tech. MTT-55*, 7 (July), 1417–1424.

[36] Zaabab A. H., Zhang Q., Nakhla M. (1995). A neural network modeling approach to circuit optimization and statistical design. *IEEE Trans. Microw. Theory Tech. MTT-43*, 6 (June), 1349–1358.

[37] Zhang Q. -J., Gupta K., Devabhaktuni V. K. (2003). Artificial neural networks for RF and microwave design–from theory to practice. *IEEE Trans. Microw. Theory Tech. MTT-51*, 4 (April), 1339–1350.

4 Optoelectronics

4.1 Executive summary

All land line communication systems currently use optical fibres as the channel due to their very low attenuation, and deployed systems operate up to 40 Gigabits per second (Gbps), with experimental systems at 160 Gbps and higher. These high speed systems use very stable sources, usually the distributed feedback lasers, with external modulators and very fast detectors. Direct modulation of lasers leads to chirp, caused by the laser frequency varying when modulated, and therefore is not used in these systems. However, interest remains in high speed lasers for several other applications. This chapter discusses high speed optoelectronic devices which include light-emitting diodes (LEDs), semiconductor lasers, photoconductors, p–i–n diode photodetectors, avalanche photodetectors, metal–semiconductor–metal photodetectors, travelling wave photodetectors and briefly the phototransistor. The physics of the devices are outlined and then the parameters that make these devices fast are discussed with specific examples from the literature.

4.2 Optical sources

Two types of sources are widely used in optical communication systems, and they are the LED and the semiconductor laser. LEDs are inherently slow; their response times are determined by the lifetime of the carriers in the active region, and in most LEDs this is between 2 ns and 10 ns. Thus, small-signal response of LEDs is of the order of 100 MHz, and the large-signal response is smaller than this. Lasers are complex devices, and the response time is determined by the so-called *relaxation oscillation frequency*, and this is due to the interaction between the photons which have a finite lifetime, which is a measure of the cavity quality factor, Q, the differential gain of the structure, the lifetime of the carriers and other parameters [1]. Most lasers, when directly modulated, suffer from spectral broadening, and therefore direct modulation is not used for long haul fibre-optic systems; but when this broadening is not important, direct modulation may be used.

In this section, the fundamental aspects of LEDs and lasers are discussed, followed by an outline of the different types of lasers. To calculate the response of lasers, the rate equations are introduced and these are solved analytically for the relaxation frequency f_r. A short discussion of the noise mechanisms of lasers follows. A further section is included on the very high f_r lasers.

4.2.1 Preliminaries

The generation of non-thermal light in a semiconductor requires the creation of hole–electron pairs which recombine radiatively to emit photons at the bandgap energy, or impurity level to valence band energy. The more recent source of light is the intra-band relaxation of electrons from a higher energy level to a lower energy level in the conduction band of a quantum well, to emit photons, and this has resulted in the quantum cascade lasers. In both these sources of light, a means of generation of the excess electron–hole pairs or excitation of the carriers to a higher energy state requires the expenditure of energy. If this energy is from an electrical source, then this is called *electrical pumping* leading to electroluminescent emission. Alternatively, optical pumping with photons of a higher energy than the emitted photons is also a possibility. In this chapter, only electroluminescent sources are considered, and quantum cascade lasers are not discussed.

4.2.2 Light-emitting diodes

The ubiquitous green and red LEDs are fabricated in GaP which is an indirect gap semiconductor. The impurities of zinc oxide and nitrogen form isoelectronic bands which give rise to the red and green emissions, respectively. Other LEDs include the blue GaN device, the near infrared GaAs 870 nm device and the InGaAsP 1300 nm and 1550 nm devices.

The mechanisms that create the LED are based on p–n junctions. A forward-biased p–n junction injects minority carriers to both sides of the junction, and these diffuse away from the junction to recombine radiatively and non-radiatively. The radiative recombination results in emission of photons, with energy $h\nu$ approximately equal to the bandgap energy $\mathrm{E_g}$. The recombination is spontaneous, which implies that emission has random phase, and the linewidth is of the order of a few $\mathrm{k_B T}$, where $\mathrm{k_B}$ is Boltzmann's constant, and T is the temperature in kelvin, and the emission is incoherent. The internal quantum efficiency is a measure of how efficiently the injected carriers produce light, since some of them recombine non-radiatively. Define the total recombination rate R_{total}, in recombination events per second, as the sum of the radiative recombination rate R_{rr} and the non-radiative recombination rate R_{nr}. Then, the internal quantum efficiency is given by the ratio of the radiative recombination rate to the total rate:

$$\eta_{\mathrm{i}} = \frac{R_{\mathrm{rr}}}{R_{\mathrm{total}}} = \frac{R_{\mathrm{rr}}}{R_{\mathrm{rr}} + R_{\mathrm{nr}}}. \tag{4.1}$$

However, lifetimes are more easily measured or inferred, and the quantum efficiency is recast in terms of the lifetimes. The radiative recombination rate is defined as the ratio of the non-equilibrium carrier density N to the radiative lifetime τ_{rr}, and similarly, the non-radiative recombination rate is the ratio of the carrier density N to the non-radiative lifetime τ_{nr}, and thus $R_{\mathrm{rr}} = N/\tau_{\mathrm{rr}}$ and $R_{\mathrm{nr}} = N/\tau_{\mathrm{nr}}$. It follows that the total lifetime τ_{total} is given by

$$\frac{1}{\tau_{\mathrm{total}}} = \frac{1}{\tau_{\mathrm{rr}}} + \frac{1}{\tau_{\mathrm{nr}}}. \tag{4.2}$$

Hence, substituting in the Equation (4.1), the internal quantum efficiency is given by

$$\eta_\mathrm{i} = \frac{\tau_\mathrm{nr}}{\tau_\mathrm{nr} + \tau_\mathrm{rr}}. \tag{4.3}$$

The total lifetime is obtained from the definition

$$R_\mathrm{nr} + R_\mathrm{rr} = R_\mathrm{total} = \frac{N}{\tau_\mathrm{total}}, \tag{4.4}$$

and the total recombination rate R_total is given by

$$R_\mathrm{total} = A_\mathrm{nr}N + B_\mathrm{r}N^2, \tag{4.5}$$

where A_nr is the non-radiative recombination coefficient and B_r is the radiative recombination coefficient. It follows that

$$\tau_\mathrm{total} = (A_\mathrm{nr} + B_\mathrm{r}N)^{-1}. \tag{4.6}$$

In direct gap semiconductors, the radiative lifetime is comparable to the non-radiative lifetime, whereas in indirect gap semiconductors, the non-radiative lifetime is much shorter than the radiative lifetime. Thus, the internal quantum efficiency is about 0.5 for direct gap semiconductors and this improves very considerably for lasers to almost unity when the radiative lifetime becomes very small compared to the non-radiative lifetime. In indirect gap semiconductors, silicon and germanium for example, the internal quantum efficiency is of the order of 10^{-5}. Surface recombination is also a problem as this is non-radiative and decreases the internal quantum efficiency. To reduce this, a heterojunction layer of a higher band gap may be deposited above the emission layer. The other problem is the absorption of the emitted photons in the generating layer, and there is little that can be done about this. The higher bandgap layer has very low absorption, and this heterolayer may help marginally.

Consider the structure of the LED which is usually a p–n or p^+–n or n^+–p junction. It is necessary to have the junction close to the surface so that the emitted light is able to escape from the material into the air. Initially, consider the carrier injection process in a p–n junction when it is forward-biased. A p–n junction when forward-biased injects holes into the n region and electrons in the p region. With forward bias of V_f, it can be shown that the current density in the p–n junction is given by

$$J = q\left(\frac{D_\mathrm{p}p_\mathrm{n0}}{L_\mathrm{p}} + \frac{D_\mathrm{n}n_\mathrm{p0}}{L_\mathrm{n}}\right)(e^{qV_\mathrm{f}/k_\mathrm{B}T} - 1), \tag{4.7}$$

where D_p and D_n are the minority hole and minority electron diffusion constants in the n-type and p-type material respectively adjacent to the junction; L_p and L_n are the diffusion lengths of the minority holes and minority electrons in the n- and p-type material respectively adjacent to the junction; p_n0 and n_p0 are the minority hole and electron densities in the n- and p-type material; q is the charge magnitude of an electron; k_B is Boltzmann's constant; T is the temperature in kelvin and $L_\mathrm{p,n} = \sqrt{\tau_\mathrm{p,n}D_\mathrm{p,n}}$. In some LEDs, the junction is n^+–p, in which case p_n0 in the n^+ region is very small, as it is equal to n_i^2/n^+, and in many III–V materials n_i, the intrinsic density, is very low, for example in GaAs it is about 10^6 cm^{-3}. The hole injection term which is the first term in

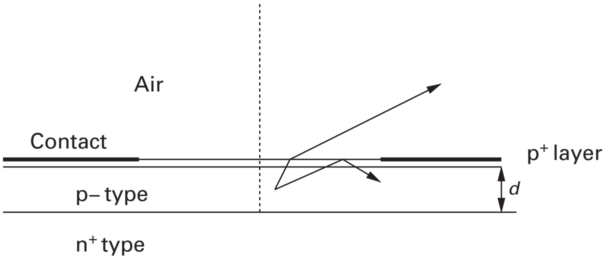


Fig. 4.1 Sketch of the n^+–p junction, with a very thin p^+ and the contacts of the LED. Also shown is the effect of total internal reflection in the escape of light from the p^+ layer.

Equation (4.7) is small, and the injection is largely from the n^+ region into the p region. Thus, the injection efficiency, which is the ratio of the injected current into the p region to the total injected current is almost unity due to the fact that the injected current into the n^+ region is much smaller than that injected into the p region. This is the case for the p^+–n junction as well, since the minority carrier density in the highly doped region is always much smaller than that in the lower doped region, and the injection efficiency is assumed to be unity. However, as indicated above, the internal quantum efficiency is of the order of 50% for direct gap semiconductors. As the junction temperature rises, the non-radiative recombination increases and so this figure is typical for room temperature devices.

Assuming that the injection is primarily from the n^+ side of the n^+–p junction into the p region. The minority electron distribution in the p layer is of the form $\Delta n e^{-z/L_\mathrm{n}}$, where the z direction is normal to the surface in Figure 4.1, and z is the distance from the junction towards the surface. This minority carrier distribution may be shown to be equivalent to $n_{\mathrm{p}0}(e^{qV/k_\mathrm{B}T} - 1)e^{-z/L_\mathrm{p}}$. Thus, the radiative recombination is over this distribution, and for all practical purposes it can be shown that this is equivalent to an uniform distribution over the distance L_n. Assume that the thickness of the active layer is d, which is assumed to be much larger than the diffusion length L_n. Then, integrating this distribution

$$\int_0^\infty \Delta n e^{-z/L_\mathrm{n}} dz = \Delta n L_\mathrm{n}, \tag{4.8}$$

which shows that Δn may be considered uniformly distributed over L_n. The choice of the upper limit of ∞ is because $d \gg L_\mathrm{n}$. The lifetimes in GaAs and InGaAsP vary from 2 ns to about 10 ns, and mobility of minority carriers of 1000 cm^2 (V.s)$^{-1}$ gives L_n values of less than 0.5 μm, and the active layer thickness may exceed this value. In the following sections, the tacit assumption is that the minority carrier density is uniform over L_n.

In Figure 4.1, the top p layer has the injected minority carriers recombining to create the photon source. The spontaneous emission is isotropic and the trajectories of the photons at the surface to the air region result in total internal reflection when the incidence angle exceeds the critical angle. Thus, extraction efficiency of the emission needs to take into account this total internal reflection, when the incidence angle exceeds the critical angle given by

$$\theta_c = \sin^{-1} \frac{n_0}{n_2} \tag{4.9}$$

where n_0 is the index of air equal to 1 and n_2 is the index of the top layer of the LED through which the emitted photons escape. For normal incidence of the photons, the reflection coefficient Γ is given by

$$\Gamma = \frac{n_0 - n_2}{n_0 + n_2}. \tag{4.10}$$

The fractional transmitted power or the transmissivity $(1 - \Gamma^2)$ for normal incidence, and is given by

$$T(0) = \frac{4n_0 n_2}{(n_0 + n_2)^2}. \tag{4.11}$$

As the photon angle of incidence varies from the normal to θ_c, the reflection coefficient changes, depending on the polarisation, which averages out between the two perpendicular and parallel cases for the spontaneous emission, and the transmissivity also changes [7]. The external quantum efficiency is obtained from

$$\eta_{ext} = \frac{1}{4\pi} \int_0^{\theta_c} T(\theta) 2\pi \sin\theta d\theta. \tag{4.12}$$

Since the expressions for the transmissivity vary with θ, and are difficult to integrate, $T(\theta)$ is replaced by $T(0)$. Substituting for $T(0)$ from Equation (4.11) and using Equation (4.9), the external quantum efficiency becomes

$$\eta_{ext} = \frac{1}{n_2(n_0 + n_2)^2}. \tag{4.13}$$

For a value of n_2 of 3.5 and n_0 of unity, η_{ext} is of the order of 1.4%, which suggests that most of the light generated is trapped inside the device. The presence of a heterojunction layer on the top surface, discussed below, complicates this expression as the index of this top layer is lower than that of the active layer. An anti-reflection coating on the surface helps to improve this factor considerably. The obvious method of extracting light from the p-layer is to place a hemispherical lens, index identical to the p-layer, on the surface. If there is a hetero-layer, then this needs to modified further. In the case of the p layer, the extraction efficiency becomes very high, but the problem is getting a suitable lens with the same index.

The optical power emitted by the LED is given by

$$P_{opt} = \eta_{int} \eta_{ext} (h\nu) \frac{I}{q}, \tag{4.14}$$

where I/q is the number of electrons or holes that are injected into the active region per second; the internal quantum efficiency η_{int} defines the fraction of them that recombine radiatively; and the external quantum efficiency η_{ext} the fraction of the generated photons that escape from the active layer. The total quantum efficiency of the LED is a measure of its performance and is the ratio of the output optical power to the input electrical power, which is given as $P_0 = V_0 I$, where V_0 is the voltage drop across the device and the current is given by I. Thus, substituting from Equation (4.14)

$$\eta_{\text{total}} = \eta_{\text{int}}\eta_{\text{ext}}\frac{h\nu}{qV_0}. \tag{4.15}$$

Neglecting contact resistance drops, $h\nu \approx E_{\text{g}} \approx qV_0$, where E_{g} is the band gap in eV, which then makes the total quantum efficiency, which is also the external power efficiency or the wall plug efficiency:

$$\eta_{\text{total}} \approx \eta_{\text{int}}\eta_{\text{ext}}. \tag{4.16}$$

This is usually less than a few percent, unless other techniques are used to extract the light more efficiently. Edge-emitting LEDs, discussed below, are therefore much more efficient. For visible LEDs the luminosity is also an issue [3], but this is not considered here.

The responsivity R of the LED is defined as the ratio of the emitted optical power to the current, and substituting from Equation (4.14), is given by

$$R = \eta_{\text{int}}\eta_{\text{ext}}\frac{h\nu}{q}. \tag{4.17}$$

This is of the order of 0.01 W A^{-1} for the above values of η unless this becomes much larger.

The spectral width of the emission $\Delta\nu$ of LEDs is approximately defined by $k_{\text{B}}T/qh$ in Hz and peaks at $(E_{\text{g}} + k_{\text{B}}T/2q)/h$ [1]. The full width half maximum (FWHM) is $\sim 1.8k_{\text{B}}T/qh$, and at room temperature ($T = 300\,\text{K}$) is about 11 THz. The spectral width in wavelength $\Delta\lambda$ varies as $\Delta\lambda = \Delta\nu(\lambda^2/c)$, and so varies from about 30 nm to 90 nm.

Modulation response

The LED is effectively a n$^+$–p junction, and therefore the usual diode current relationship holds, as discussed above:

$$I = I_0(e^{qV/k_{\text{B}}T} - 1) \tag{4.18}$$

To calculate the modulation response, the small-signal behaviour is obtained by initially biasing the LED at some bias point, where the current is given by I_{b}. The carrier lifetime in the active layer determines the modulation rate of these LEDs. If N is the density of carriers and I is the current that flows into the LED, the following rate equation determines the carrier dynamics:

$$\frac{dN}{dt} = \frac{I}{q\text{Vol}} - \frac{N}{\tau_{\text{total}}} \tag{4.19}$$

where Vol is the volume of the active region. Under steady-state conditions, the time dependence is zero and at a bias current of I_{b}, then

$$N_{\text{b}} = \frac{I_{\text{b}}\tau_{\text{total}}}{q\text{Vol}}, \tag{4.20}$$

where N_b is the carrier density at bias current I_b. To obtain the small-signal response, let the current have an ac component given by $I_m e^{j\omega_m t}$, and similarly the carrier density also has a small-signal term $N_m e^{j\omega_m t}$. Note that it is assumed that $I_m \ll I_b$ the bias current, and similarly $N_m \ll N_b$. Thus

$$I(t) = I_b + I_m e^{j\omega_m t} \tag{4.21}$$

$$N(t) = N_b + N_m e^{j\omega_m t}. \tag{4.22}$$

Substituting in Equation (4.19) two component equations arise; the steady-state equation is satisfied by the result in Equation (4.20). The time varying equation gives rise to the following solution

$$N_m(\omega_m) = \frac{\dfrac{I_m \tau_{total}}{(q\text{Vol})}}{1 + j\omega_m \tau_{total}}. \tag{4.23}$$

The expression for optical power output given in Equation (4.14) varies as I/q, which is proportional to the carrier density. Thus, the modulated optical power $P_m(\omega)$ varies as N_m. It follows that

$$P_m(\omega_m) = P_{opt} \frac{1}{\left[1 + (\omega_m^2 \tau_{total}^2)\right]^{1/2}}, \tag{4.24}$$

where P_{opt} is the zero frequency steady-state output power. The frequency at which half power is obtained by setting the denominator of the above equation to 2, which leads to

$$f_{3dB} = \sqrt{3} \frac{1}{2\pi \tau_{total}}. \tag{4.25}$$

This equation confirms that the bandwidth is inversely proportional to the carrier lifetime.

LED structures

The basic planar LED is shown in Figure 4.2. In general, it is necessary to have a thin highly doped contact layer to reduce the contact resistance, and the emerging light has to pass through it. The absorption creates additional loss. If the active layer is intrinsic or undoped, with p$^+$ and n$^+$ layers on either side, then the injection into this layer is from both junctions. In practice, the intrinsic layer is always unintentionally doped as either p$^-$ or n$^-$, where one of the junctions is a p–n junction and the other is a high–low junction; but injection takes place from both junctions.

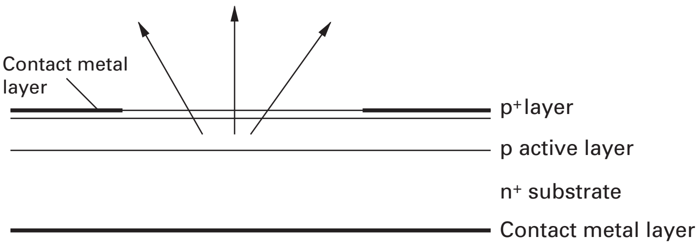


Fig. 4.2 Sketch of the n$^+$–p junction LED with a p$^+$ contact layer.

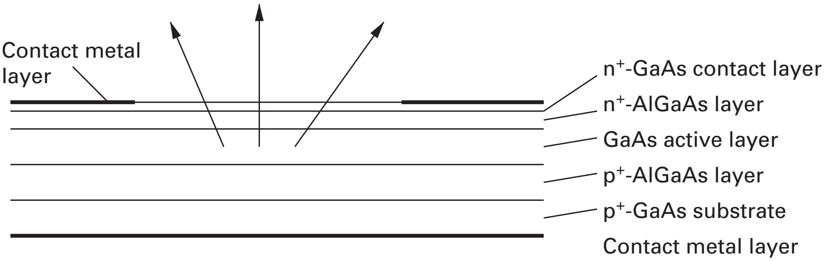


Fig. 4.3 Schematic diagram of the heterojunction LED.

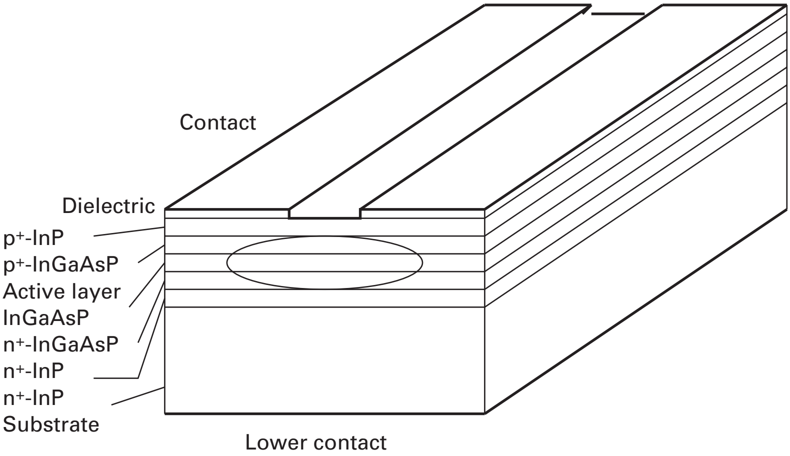


Fig. 4.4 Schematic diagram of the edge-emitting LED.

A variation of this structure is to have higher bandgap heterostructure layers both above and below the active layer, appropriately doped. The diffusion of the injected electrons and holes injected from the active layer is blocked by the heterojunctions that are now formed on both sides of the active layer, as shown in Figure 4.3. These heterolayers prevent the diffusion of the minority carriers to the surface, and therefore prevent surface recombination. Most LEDs designed at the current time use the heterolayers where available.

A further variation on this is the placement of dielectric mirrors below the active layer to reflect the light emitted towards the substrate. A second variant is the edge-emitting LED shown in Figure 4.4, in which the heterolayers above and below the active layers together with the active layer act as a waveguide. A high reflectivity mirror at one facet makes this a superluminescent diode; with mirrors on both ends and with adequate gain, the LED may operate as a laser. The external quantum efficiency of the edge-emitting LED is very much higher than the surface-emitting devices because the transmissivity at the edge facet is of the order of 0.7 for the waveguide index n_2 of 3.5. With a high reflectivity facet coating at one end, the external quantum efficiency is also about 0.7 for n_2 of 3.5. Anti-reflection coating on transmitting facet would increase the transmissivity to almost unity, and the external quantum efficiency also becomes nearly unity. Thus, the total quantum efficiency or the wall plug efficiency of edge-emitting LEDs is determined largely by η_{int}, which is of the order of 0.5 in direct gap semiconductors.

The normal emission of the surface-emitting LED is Lambertian, which implies that the emission intensity at angle θ from the normal is $\cos\theta$, which means that the beam width is 120°. For the edge-emitting LED, the emission is elliptic in form, and the beam width is about 30° in the horizontal plane but remains 120° in the vertical plane.

LEDs are used in optical fibre communication systems that operate with multimode fibres, as the broad emission spectrum prevents modal noise being a problem [17]. A diode proposed by Burrus [4] has the fibre bonded into the face of the surface-emitting LED.

4.2.3 Semiconductor lasers

While LEDs utilise spontaneous emission for the emitted light, lasers operate on stimulated emission-generated light. Lasers are optical oscillators in which the gain medium is in a cavity; the light acquires gain in the medium between reflections from the ends of the cavity until steady state is reached, when the gain becomes saturated. The simplest version is a Fabry Perot cavity with the gain medium between two mirrors.

In this section, semiconductor lasers are discussed, as they may be designed to be high-speed and high frequency lasers. While solid state, fibre and other types of lasers may produce extremely short pulses, known as ultrafast lasers, their repetition rates are typically 80 MHz, to a maximum of a few GHz, and are not considered 'high-speed' or 'high frequency' lasers, and therefore not discussed here. The semiconductor laser threshold condition is first considered, then waveguides used in these lasers are outlined, and different types of lasers are discussed. This is followed by the derivation of the rate equation for these lasers, and the solutions for various conditions. Discussion of noise in semiconductor lasers is also included. Subsequently, quantum well, quantum dot lasers and vertical cavity lasers are briefly discussed.

Basic concepts

When the gain medium is in a cavity formed by two mirrors, shown schematically in Figure 4.5, light in the form of an electromagnetic wave, electric field amplitude $\mathcal{E}_0$, travels from one end of the cavity to the other end, where it is reflected by the end

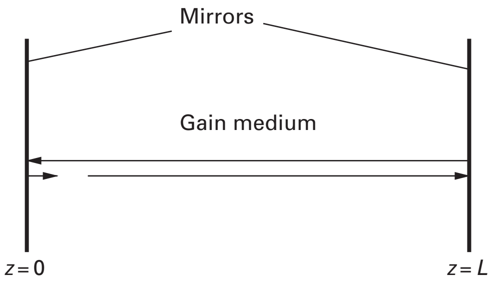


Fig. 4.5 Schematic diagram of gain medium in a cavity formed by two end mirrors.

mirror, and then propagates back, and is then reflected by the second mirror to return to its starting point. Suppose that the length of the cavity is L, this determines the distance travelled by the wave between reflections. Assume that the gain of the medium is given by $g/2$ Nepers per unit length, and hence the intensity gain is g per cm. Suppose the medium internal loss is $\alpha_{\text{int}}/2$ Nepers per unit length, or the intensity loss is α per cm, the medium index is n, and the reflection coefficients of the mirrors are r_1 and r_2. For the structure to commence oscillation, the loop gain should be unity or larger, which implies that the field at the starting point $\mathcal{E}_0$ has to undergo these reflections and is also subjected to gain to become after one round trip

$$\mathcal{E}_0 = \mathcal{E}_0(2L) = \mathcal{E}_0 e^{-j2k_0nL} e^{(g-\alpha_{\text{int}})L} |r_1||r_2|. \tag{4.26}$$

The real part of this equation gives

$$g = \alpha_{\text{int}} + \frac{1}{L} \ln \frac{1}{(|r_1||r_2|)}. \tag{4.27}$$

The mirror reflectivity is defined as $R_1 = |r_1|^2$, and similarly $R_2 = |r_2|^2$. Substituting in the above equation,

$$g = \alpha_{\text{int}} + \frac{1}{2L} \ln \frac{1}{(R_1 R_2)} = \alpha_{\text{int}} + \alpha_{\text{mir}}. \tag{4.28}$$

The first term on the right-hand side is the medium loss without pumping and the second term is the mirror loss term, usually denoted by α_{mir}. For the gain medium of GaAs, the index is about 3.45, and the mirror is assumed to be formed by cleaved facets which result in plane mirrors, parallel to each other. The reflection coefficient of each facet mirror is

$$r_{1,2} = \frac{3.45 - 1}{3.45 + 1} = 0.551. \tag{4.29}$$

The facet mirror reflectivity is 0.303, and hence the mirror loss term α_{mir} is $1.194\,\text{cm}^{-1}$. The internal loss term is usually between 10 and $20\,\text{cm}^{-1}$, and therefore the gain needs to be 11.194–$21.194\,\text{cm}^{-1}$ for the round trip gain to be unity, and generally the gain needs to be higher than this value.

In practice, only part of the wave obtains gain from the active region, and this is defined by the confinement factor Γ, which is discussed later. Thus, the oscillation condition for the laser in Equation (4.28) becomes

$$\Gamma g = \alpha_{\text{int}} + \alpha_{\text{mir}}. \tag{4.30}$$

Also the complex propagation constant in the gain medium now is given by

$$\beta = nk_0 - j\left(\frac{\alpha}{2}\right), \tag{4.31}$$

where n is the guide effective index and the loss term α is the net loss, and given by

$$\alpha = \alpha_{\text{int}} + \alpha_{\text{mir}} - \Gamma g. \tag{4.32}$$

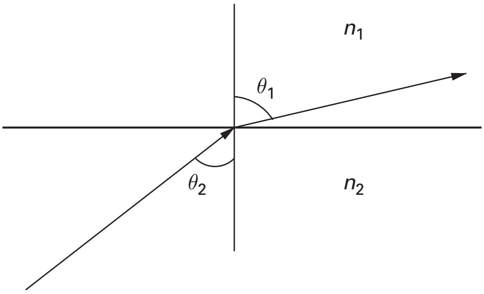


Fig. 4.6 Plane interface between medium 1, index n_1, and medium 2, index n_2, and the incident and transmitted light directions given by θ_1 and θ_2 respectively to the normal.

The imaginary part of Equation (4.26) determines the phase requirement:

$$2k_0 nL = 2m\pi \quad \text{or} \quad \nu_\mathrm{m} = \frac{mc}{(2nL)}, \tag{4.33}$$

where m is the longitudinal mode number, which may take values of $1, 2, 3, \ldots$, but cannot be zero, as the solution is then trivial. Note that the mode spacing in frequency is given by $c/(2nL)$. All longitudinal modes, with the values of m, satisfy this equation but only some of them are valid for a device. Since the medium gain is band-limited, the modes within this gain region are all excited at threshold when the gain is just larger than the losses. With increasing gain, generated by current pumping, the mode competition results in those close to the gain peak growing at the expense of the other modes.

Optical waveguides in semiconductors lasers

The light in the gain medium discussed above needs to be confined and guided, and this requires that the medium is in the form of an optical waveguide. An optical waveguide utilises the concept of total internal reflection that occurs when light emerges from a higher index medium to a lower index medium. In Figure 4.6, light in the form of an electromagnetic plane wave in a medium of index n_2 is incident on the plane interface between the media, at an angle θ_2 to the normal. The second medium index is given by n_1, with $n_2 > n_1$, and the light emerges into the second medium at an angle θ_2, obtained from Snell's law:

$$n_2 \sin\theta_2 = n_1 \sin\theta_1 \tag{4.34}$$

Since $n_2 > n_1$, it follows that $\theta_1 > \theta_2$. At the critical angle of $\theta_{2\mathrm{c}}$, the value of θ_1 becomes $\pi/2$, which implies that the emerging light travels along the interface. For incident angles greater than $\theta_{2\mathrm{c}}$, total internal reflection occurs, and with a second interface below the first, similar total internal reflection occurs so that the light remains confined to the high index region as shown in Figure 4.7. A similar confinement may occur in the plane normal to this to obtain two-dimensional guiding with appropriate layers to provide for an index guiding structure.

The solution of the wave equation for this three-layer guide used in a laser results in both even and odd modes. The even mode is sketched in Figure 4.8. Semiconductor lasers are classified as gain-guided lasers or index-guided lasers. In gain-guided

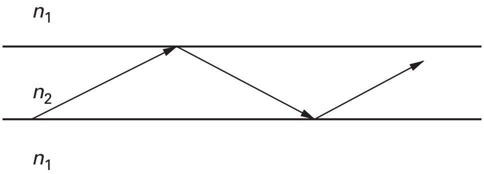


Fig. 4.7 Slab waveguide which provides confinement in the transverse plane.

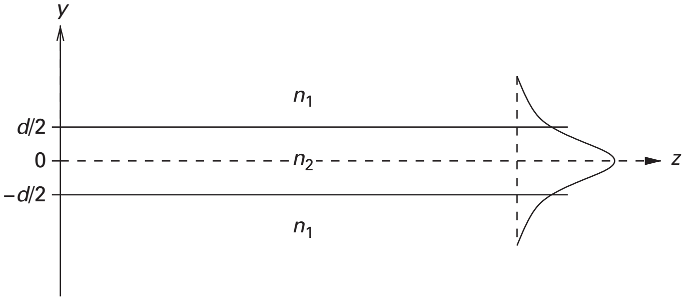


Fig. 4.8 Three-layer slab symmetric guide, which is typical for laser structures: the upper layer is p-type, the guide layer of higher index, is generally undoped, and the lower layer is n-type. The guide layer thickness is d. The even mode field distribution is sketched .

lasers, the mode confinement in the lateral direction, the horizontal direction, is not designed into the structure, and the mode is guided by the gain region of the structure. In index-guided lasers, the waveguide is well defined and the mode is confined in both the transverse direction, the vertical direction normal to the plane of the wafer and the lateral direction. A very popular structure among research scientists is the ridge laser, which is weakly index-guided. In all these lasers, the transverse confinement is obtained by the design of the heterolayers. Gain-guided, index-guided and weakly index-guided lasers are shown schematically in Figure 4.9.

Optical waveguides are analysed using Maxwell's equations, and the solution of the two-dimensional guide problem may be performed by a variety of methods. The results of this analysis enables the active region which acts as the guide to be designed. These guides are generally designed to be single mode in the $x - y$ plane. The width of the guide is usually designated by w, its thickness by d and the length of the laser is L, leading to an active volume of Lwd. The current is assumed to flow in the contact over the length of laser and the width of w, which results in a current flow area of Lw. The analysis of the guide is generally performed, assuming that there is no loss or gain, and these terms are added as perturbations. The analysis determines the propagation constant of the guide β, and also its effective relative permittivity ϵ_{reff} and effective index n_{eff}. The phase and group velocities are different in optical guides, and the corresponding effective indices are also obtained. An important parameter is the confinement factor Γ. This defines the fraction of the power contained in the active guide region to the total power in the particular mode of the guide.

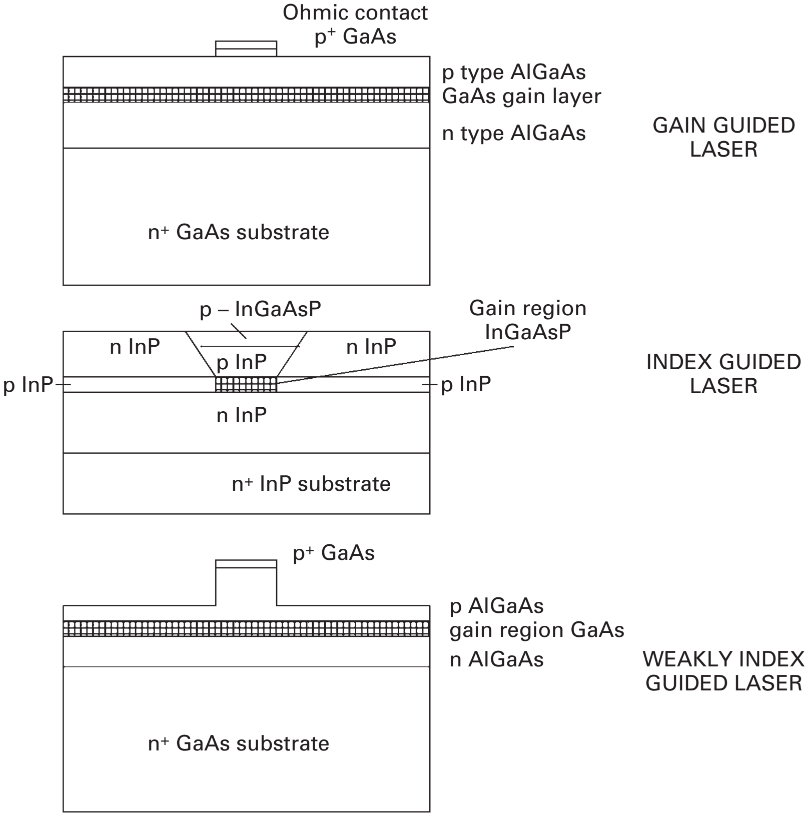


Fig. 4.9 Schematic diagram of the gain-guided laser, the index-guided laser and the ridge laser which is a weakly index-guided structure.

The effective index method is discussed next; it leads to approximate analytic expressions for the design of the guide layer. In this method, the guiding in the transverse, vertical, direction is analysed, and effective indices are calculated for each of the different transverse regions as they vary in the lateral direction. With the effective indices known for these transverse regions in the lateral, horizontal, direction, the one-dimensional guide in the lateral direction may be analysed, to obtain the solution of the entire guide.

Maxwell's equations lead to the Helmholtz's equation, assuming that the cross-section of the waveguide remains constant in the z direction and the propagation constant does not vary with z. Thus, the wave equation is of the form:

$$\nabla^2 \mathcal{E} + \epsilon_{\mathrm{r}}(x, y) k_0^2 \mathcal{E} = 0, \tag{4.35}$$

where $\epsilon_{\mathrm{r}}(x, y)$ is the structure permittivity which varies with both x and y directions but does not vary with the z direction. The dielectric constant also varies with pumping, typically the increase in carriers results in a small decrease in index, and vice versa. The dielectric constant is complex due to absorption or gain, and these effects are added as perturbations. Ignoring these effects enables the modes of the guiding structure to be obtained, and using the effective index approach allows both index-guided structures and gain-guided laser structures to be analysed.

In the effective index approach, the assumption is that the solution may be separated into a y-varying, transverse varying, component in the form of a slab waveguide, and a similar effective index varying guide in the x direction. Assume that the solution of the Equation (4.35) obtained by the separation of variables is of the form:

$$\boldsymbol{\mathcal{E}} = \mathbf{a}\xi(y; x)\psi(x)e^{-\mathrm{j}\beta z}, \tag{4.36}$$

where β is the propagation constant and **a** is unit vector in the direction of the $\mathcal{E}$-field, $\boldsymbol{\mathcal{E}}$, which defines the mode polarisation. Substituting in Helmholtz's Equation (4.35),

$$\frac{1}{\psi}\frac{d^2\psi}{dx^2} + \frac{1}{\xi}\frac{d^2\xi}{dy^2} + \left[k_0^2\epsilon_\mathrm{r}(x, y) - \beta^2\right] = 0, \tag{4.37}$$

where $k_0 = \omega\sqrt{(\epsilon_0\mu_0)}$, which defines the propagation constant in free space.

The next step is to solve the transverse, y directed, field distribution and with it the effective propagation constant $\beta_\mathrm{eff}(x)$ for a fixed value of x. Using the transverse part of the Equation (4.37)

$$\frac{d^2\xi}{dy^2} + \left[k_0^2\epsilon_\mathrm{r}(x, y) - \beta_\mathrm{eff}^2(x)\right]\xi = 0. \tag{4.38}$$

The lateral, x directed, field distribution and the propagation constant β are then obtained from the equation:

$$\frac{d^2\psi}{dx^2} + \left[\beta_\mathrm{eff}^2 - \beta^2\right]\psi = 0. \tag{4.39}$$

Consider the transverse modes of a typical laser structure, which may have as many as four or five layers in the slab guide form. However, the principle of the design is obtained from the three-layer guide shown in Figure 4.8. The solution of Equation (4.38) for the even TE mode ($\boldsymbol{\mathcal{E}}$ is directed along the x, lateral, direction) is of the form:

$$\xi = A_\mathrm{e}\cos(\kappa y) \qquad \text{for } |y| \leq d/2 \tag{4.40}$$

$$\quad = B_\mathrm{e}e^{-\gamma(|y|-d/2)} \qquad \text{for } |y| \geq d/2, \tag{4.41}$$

where

$$\kappa = k_0\left(n_2^2 - n_\mathrm{eff}^2\right)^{0.5} \tag{4.42}$$

$$\gamma = k_0\left(n_\mathrm{eff}^2 - n_1^2\right)^{0.5}. \tag{4.43}$$

Note that n_1 and n_2 are the refractive indices of the cladding and the guide layer respectively, and $n_2 > n_1$ for guiding.

In the TE mode case, the only components of field present are $\mathcal{E}_\mathrm{x}$, $\mathcal{H}_\mathrm{y}$ and $\mathcal{H}_\mathrm{z}$. The boundary conditions require the continuity of ξ and $d\xi/dy$ at $|y| = d/2$, and these correspond to the continuity of the $\mathcal{E}_\mathrm{x}$ component across the interfaces, and the continuity of the $\mathcal{H}_z$ component across the interfaces, respectively. This leads to the following two equations:

$$A_e \cos\left(\frac{\kappa d}{2}\right) = B_e \tag{4.44}$$

and

$$\kappa A_e \sin\left(\frac{\kappa d}{2}\right) = \gamma B_e. \tag{4.45}$$

Dividing the above two equations,

$$\kappa \tan\left(\frac{\kappa d}{2}\right) = \gamma. \tag{4.46}$$

The solution of this equation gives the values of β_{eff} from which the effective index is obtained through the relationship $\beta_{\text{eff}} = k_0 n_{\text{eff}}$.

For the odd TE modes, it may be shown that the dispersion relationship becomes

$$-\kappa \cot\left(\frac{\kappa d}{2}\right) = \gamma. \tag{4.47}$$

This is obtained by setting the initial solution for ξ in Equation (4.40) as $\sin(\kappa y)$ instead of $\cos(\kappa y)$.

For the TM modes, the $\boldsymbol{\mathcal{E}}$ vector is along the y direction, normal to the interfaces in Figure 4.8. In this case, the only components of fields present are $\mathcal{E}_y$, $\mathcal{H}_x$ and $\mathcal{E}_z$. From the Maxwell curl equations, $\mathcal{E}_z$ is proportional to $d\mathcal{E}_y/dy$. Thus, the boundary conditions are the continuity of the y component of electric flux across the interfaces and the continuity of $\mathcal{E}_z$ across the interfaces. These lead to the following equations for the even TM modes:

$$n_2^2 A_e \cos\left(\frac{\kappa d}{2}\right) = n_1^2 B_e \tag{4.48}$$

and

$$\kappa A_e \sin\left(\frac{\kappa d}{2}\right) = \gamma B_e. \tag{4.49}$$

Dividing the above two equations gives the dispersion relationship:

$$\kappa n_1^2 \tan\left(\frac{\kappa d}{2}\right) = n_2^2 \gamma. \tag{4.50}$$

For the odd TM mode, the dispersion relationship is given by

$$-\kappa n_1^2 \cot\left(\frac{\kappa d}{2}\right) = n_1^2 \gamma. \tag{4.51}$$

Consider the TE mode solutions, squaring Equations (4.42) and (4.43), and adding results in

$$\kappa^2 + \gamma^2 = k_0^2 \left(n_2^2 - n_1^2\right). \tag{4.52}$$

This is the equation of a circle in the $\kappa - \gamma$ plane, and the intersection of the circle with the curves defined by the Equations (4.46) and (4.47) provides the modal solutions for these equations. Note that multiple solutions are likely to occur as both tan and cot are periodic functions, and also depending on the parameters of the guide, defined by n_1, n_2, d and the wavelength which defines k_0.

At the cutoff of the guide which implies that the guide is no longer guiding, then $\gamma = 0$. Note that γ may not become negative, as it would imply exponential growth in the cladding region which is non-physical, and therefore the smallest value γ takes zero. When $\gamma = 0$, then from Equations (4.46) and (4.47)

$$\kappa d = p\pi, \tag{4.53}$$

where p is an integer whose even and odd values satisfy Equations (4.46) and (4.47) respectively, corresponding to the even and odd modes. Now let

$$D = p\pi = \kappa d. \tag{4.54}$$

Now for $\gamma = 0$, Equation (4.52) becomes

$$D = k_0 \left(n_2^2 - n_1^2\right)^{0.5} d, \tag{4.55}$$

where D is the normalized guide layer thickness. For a single transverse TE mode guide, this requires $D < \pi$. The layer thickness d for a single transverse mode guide is obtained as follows. Setting $k_0 = 2\pi/\lambda$ in the above Equations gives

$$d < \frac{\lambda}{2}\left(n_2^2 - n_1^2\right)^{-0.5}. \tag{4.56}$$

For GaAs/AlGaAs at 870 nm wavelength, with indices of 3.45 and 3.41, this suggests the guide thickness of less than 830 nm. The usual thickness is of the order of 0.2–0.5 μm. If the AlGaAs layer has a higher value of index, closer to that of GaAs, then the thickness of this layer may be larger. For InGaAsP lasers [2], the layer thickness is also of the order of 0.2 μm, and in this case the layer thickness needs to be less than 0.48 μm. According to [2], this relationship holds for lasers in the wavelength range 1.1–1.65 μm. The confinement factor in the transverse direction Γ_T is a measure of how much of the mode power lies in the guide active region, and is calculated from the equation:

$$\Gamma_T = \frac{\int_{-d/2}^{d/2} \xi^2(y)dy}{\int_{-\infty}^{\infty} \xi^2(y)dy}. \tag{4.57}$$

Performing the integration and simplifying, [2] derives this as

$$\Gamma_T \cong \frac{D^2}{(2 + D^2)}. \tag{4.58}$$

The effective index of the guide in the transverse direction is given as [2]:

$$n_{\text{eff}}^2 \cong n_1^2 + \Gamma_T \left(n_2^2 - n_1^2\right). \tag{4.59}$$

The lateral modes are next evaluated. The loss of laser structures may be as high as 5–10 Np cm^{-1}, and therefore in principle these losses need to be taken into account. However, when pumped, the gain which is typically of the order of 50 Np cm^{-1} and larger, allows the guide to become transparent, which implies that the loss is cancelled

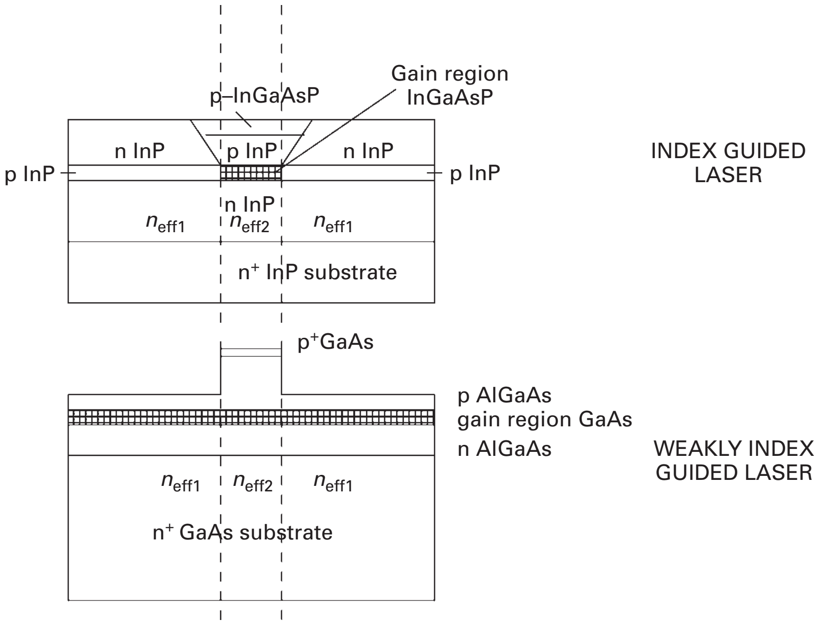


Fig. 4.10 Three-layer slab symmetric guide in which the slabs are specified by the effective indices n_{eff1} and n_{eff2} vertically, with $n_{\text{eff2}} > n_{\text{eff1}}$ for the index-guided laser and the weakly guiding ridge laser.

by the gain. However, if the losses need to be accounted for, this is usually performed by the perturbation technique. Thus, the lateral modes are calculated in a similar manner as the transverse modes, assuming the guide is lossless. In this case, the wave equation to be solved is given by Equation (4.39) and repeated here for convenience:

$$\frac{d^2\psi}{dx^2} + \left[\beta_{\text{eff}}^2(x) - \beta^2\right]\psi = 0. \tag{4.60}$$

Note that $\beta_{\text{eff}}(x) = k_0 n_{\text{eff}}(x)$ which is obtained from Equation (4.59), and thus the propagation constant of the total guide β is obtained. In the case of the index-guided laser in Figure 4.9, the transverse effective index in the main guide region has the largest effective index, and the regions outside have lower effective indices. As shown in Figure 4.10, the symmetric slab guide in the vertical direction with the effective indices calculated from the transverse slab guide equation. Here, $n_{\text{eff2}} > n_{\text{eff1}}$ so that the guide acts with the centre region as the larger index.

In this case, the TE mode from the transverse slab guide becomes a TM mode and the TM mode becomes a TE mode. Based on the earlier approach, the width of the central region is set as w, and the normalised width as W. Then, it follows that

$$W = k_0 \left(n_{\mathrm{eff2}}^2 - n_{\mathrm{eff1}}^2\right)^{0.5} w \tag{4.61}$$

and

$$W = q\pi. \tag{4.62}$$

For the lowest order mode

$$w < \frac{\lambda}{2} \left(n_{\mathrm{eff2}}^2 - n_{\mathrm{eff1}}^2\right)^{-0.5}. \tag{4.63}$$

In this case, the effective indices take different values, but this provides an indication to the width of the guide which usually is in the 3 μm region.

The lateral confinement factor Γ_{L}, following the earlier derivation, is given by

$$\Gamma_{\mathrm{L}} = \frac{W^2}{\left(2 + W^2\right)}. \tag{4.64}$$

The waveguide mode refractive index is now given by

$$n_{\mathrm{eff}}^2 \cong n_{\mathrm{eff1}}^2 + \Gamma_{\mathrm{L}}(n_{\mathrm{eff2}}^2 - n_{\mathrm{eff1}}^2). \tag{4.65}$$

The laser confinement factor Γ is the product of Γ_{T} and Γ_{L} and is given by

$$\Gamma = \Gamma_{\mathrm{T}}\Gamma_{\mathrm{L}}. \tag{4.66}$$

The two most important parameters that are obtained from this analysis are the effective index of the guide and the confinement factor. From the effective index, the guide propagation constant may be obtained.

These results are for the index-guided structures, and the same technique may be used for the weakly guiding ridge structure in Figure 4.9. The analysis of the gain-guided laser is much more complex, and interested readers are referred to the paper by Nash [24].

Emission characteristics

The waveguide which guides the light has facet mirrors on both sides of the device. These may be coated to obtain high reflectivity or to reduce the reflectivity as desired, but generally these facets are uncoated, in which case the reflectivities are equal.

With current pumping, the light intensity against input current is plotted in the L–I characteristic as shown schematically in Figure 4.11. Note the different regions of this curve: the pre-threshold where the laser output is from spontaneous emission, the threshold current at which the device starts to lase, which is the linear region, followed by the saturation region due to gain saturation. The current into the laser is of the form:

$$I = wLJ, \tag{4.67}$$

where w is the width of the active region and L is the cavity length. In practice, index-guided lasers in Figure 4.9 have leakage current through the various reverse-biased junctions, and should be added to this current expression, but this is ignored in the present discussion.

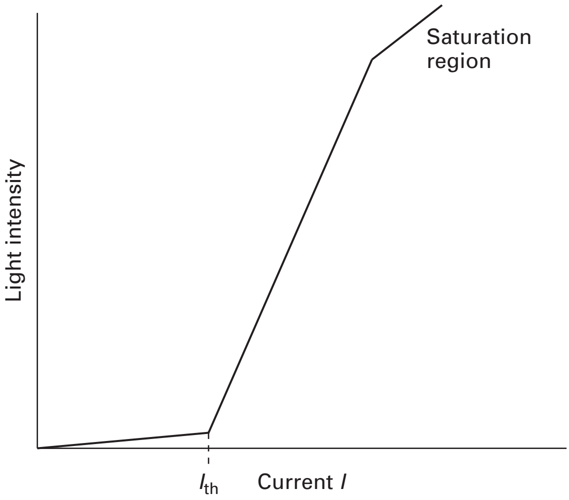


Fig. 4.11 Schematic diagram of the light intensity against current drive, the L–I characteristic.

Assume that the gain is given approximately by the expression

$$g = a(N - N_0), \tag{4.68}$$

where a is the gain coefficient, equal to $(\partial g/\partial N)$, N is the carrier density, and N_0 is the carrier density at which transparency is obtained, when population inversion occurs. The threshold carrier density is defined when the product of the confinement factor and the gain is equal to the loss, or

$$\Gamma a(N_{\mathrm{th}} - N_0) = \alpha_{\mathrm{mir}} + \alpha_{\mathrm{int}}, \tag{4.69}$$

or

$$N_{\mathrm{th}} = N_0 + (\alpha_{\mathrm{mir}} + \alpha_{\mathrm{int}})/(a\Gamma). \tag{4.70}$$

The number of carriers pumped into the active region per second is I/q. However, since the equations are in carrier density, the number of carriers injected per second per unit volume is $I/(q\mathrm{Vol}) = I/(qwLd) = J/qd$. The loss of the carrier density is through recombination at the rate $R(N)$, which also includes the stimulated emission recombination, and this is defined as R_{total}. Thus, the rate equation for the carriers is given by

$$\frac{dN}{dt} = \frac{J}{qd} - R_{\mathrm{total}} = \frac{J}{qd} - \frac{N}{\tau_{\mathrm{e}}} - R_{\mathrm{stim}}N_{\mathrm{ph}}, \tag{4.71}$$

where τ_{e} is the carrier lifetime and the last term is the stimulated emission recombination. At and below threshold, this last term may be omitted as the contribution from stimulated emission is small. At steady state, the time variation is zero, and hence

$$J = qdR_{\mathrm{total}}. \tag{4.72}$$

The stimulated emission recombination rate is $R_{\text{stim}}N_{\text{ph}}$, where N_{ph} is the photon density and R_{stim} is defined by

$$R_{\text{stim}} = \frac{c}{n}g(N), \tag{4.73}$$

where $g(N)$ is the gain defined in Equation (4.68), c/n is the group velocity.

From Equations (4.70) and (4.71), the threshold current density is obtained as

$$J_{\text{th}} = \frac{qdN_{\text{th}}}{\tau_{\text{e}}(N_{\text{th}})}, \tag{4.74}$$

where, neglecting stimulated emission recombination close to threshold,

$$\frac{1}{\tau_{\text{e}}(N)} = (A_{\text{nr}} + B_{\text{r}}N + CN^2), \tag{4.75}$$

where A_{nr} is non-radiative recombination coefficient, B_{r} is the radiative recombination coefficient and C is the Auger non-radiative recombination coefficient, which is important for long wavelength lasers.

When the laser operates beyond threshold, the carrier density is clamped at the threshold value, and further injection results in conversion into photons by stimulated emission. Thus, Equation (4.72) may be written as

$$J = qdR_{\text{total}} = qd\frac{N_{\text{th}}}{\tau_{\text{e}}} + qdR_{\text{stim}}N_{\text{ph}}. \tag{4.76}$$

Substituting from Equation (4.73), and replacing g by $\alpha_{\text{int}} + \alpha_{\text{mir}}$ at threshold,

$$J - J_{\text{th}} = qdv_{\text{g}}(\alpha_{\text{int}} + \alpha_{\text{mir}})N_{\text{ph}}. \tag{4.77}$$

The photon density in the cavity depends on the carriers injected into the cavity and the quantum efficiency of the material. However, the photons in the cavity travel at the group velocity of the medium v_{g}, and are either absorbed due to internal losses or escape from the facets. Thus, the photons have a finite lifetime in the cavity, and the photon lifetime τ_{p} is given by

$$\tau_{\text{p}} = \frac{1}{v_{\text{g}}(\alpha_{\text{mir}} + \alpha_{\text{int}})}. \tag{4.78}$$

The Equation (4.77) becomes

$$J - J_{\text{th}} = qd\frac{N_{\text{ph}}}{\tau_{\text{p}}} \tag{4.79}$$

The relationship between the injected carrier density per second and the photon density in the cavity is now obtained. The carriers recombine at a rate defined by the carrier lifetime, to produce photons, and this is accounted for by the quantum efficiency of the material from Equation (4.3), and the photon density also decays at the photon lifetime. Thus,

$$\eta_{\text{int}}\frac{(J - J_{\text{th}})}{qd} = \frac{N_{\text{ph}}}{\tau_{\text{p}}} \tag{4.80}$$

or

$$N_{\text{ph}} = \eta_{\text{int}} \tau_{\text{p}} \frac{(J - J_{\text{th}})}{qd}. \tag{4.81}$$

This equation shows that the photon density in the cavity increases linearly with current density above the threshold current density.

The output power per facet is obtained by the product of energy of each photon $h\nu$, the photon loss per facet $v_{\text{g}}\alpha_{\text{mir}}/2$, the active laser volume *Vol* and the photon density N_{ph}:

$$P_{\text{facet}} = \frac{1}{2} h\nu v_{\text{g}} \alpha_{\text{mir}} \text{Vol} N_{\text{ph}}, \tag{4.82}$$

where $\text{Vol} = Lwd$. Substituting for N_{ph}

$$P_{\text{facet}} = \frac{h\nu}{2q} \eta_{\text{int}} \frac{\alpha_{\text{mir}}}{\alpha_{\text{mir}} + \alpha_{\text{int}}} Lw(J - J_{\text{th}}). \tag{4.83}$$

Since $I = LwJ$, this equation becomes

$$P_{\text{facet}} = \frac{h\nu}{2q} \eta_{\text{int}} \frac{\alpha_{\text{mir}}}{\alpha_{\text{mir}} + \alpha_{\text{int}}} (I - I_{\text{th}}). \tag{4.84}$$

Note that the threshold current is obtained from Equation (4.74)

$$I_{\text{th}} = \frac{qLwdN_{\text{th}}}{\tau_{\text{e}}}. \tag{4.85}$$

The total output power $P_{\text{out}} = 2P_{\text{facet}}$ and the differential (external) quantum efficiency is given by

$$\eta_{\text{d}} = \frac{d\left(\dfrac{P_{\text{out}}}{h\nu}\right)}{d\dfrac{(I - I_{\text{th}})}{q}} = \eta_{\text{int}} \frac{\alpha_{\text{mir}}}{\alpha_{\text{mir}} + \alpha_{\text{int}}}. \tag{4.86}$$

This is proportional to the slope of the L–I curve. Note that in the laser the η_{int} needs to take into account the stimulated emission rate. Then, the total recombination rate becomes

$$R_{\text{total}} = A_{\text{nr}}N + B_{\text{r}}N^2 + CN^3 + R_{\text{stim}}N_{\text{ph}}. \tag{4.87}$$

The coefficients have been identified in Equation (4.75) except for R_{stim}, which is the stimulated emission rate and given in equation (4.73) Thus, the internal quantum efficiency now is given by

$$\eta_{\text{int}} = \frac{R_{\text{rad}}}{R_{\text{total}}} = \frac{B_{\text{r}}N^2 + R_{\text{stim}}N_{\text{ph}}}{A_{\text{nr}}N + B_{\text{r}}N^2 + CN^3 + R_{\text{stim}}N_{\text{ph}}} \tag{4.88}$$

Above threshold, the $R_{\text{stim}}N_{\text{ph}}$ dominates and is much larger than the other terms, which makes η_{int} almost unity.

The power efficiency of the laser is given by

$$\eta_P = \frac{P_{\text{out}}}{VI} = \frac{h\nu}{qV} \frac{\alpha_{\text{mir}}}{(\alpha_{\text{mir}} + \alpha_{\text{int}})} \frac{(I - I_{\text{th}})}{I}, \tag{4.89}$$

where V is the applied bias in volts.

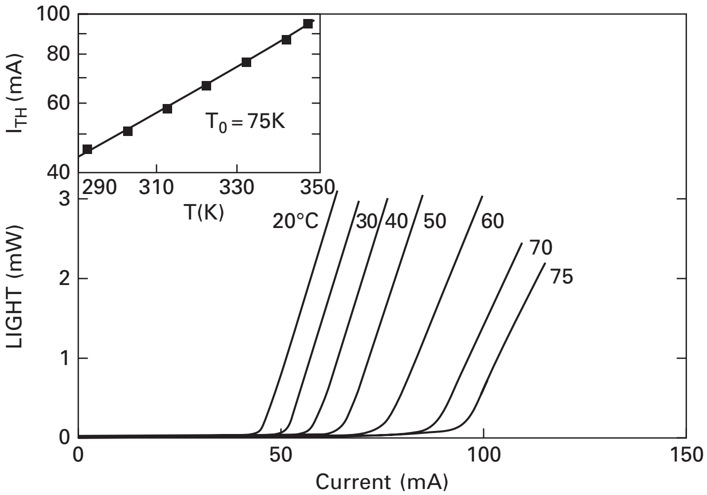


Fig. 4.12 The L–I curves for a buried heterostructure laser at different temperatures. The inset plots the threshold current against temperature to obtain T_0 (R. J. Nelson, R. B. Wilson, P. D. Wright, P. A. Barnes, N. K. Dutta, *IEEE Journal of Quantum Electronics*, Vol. 17, No. 2, pp. 202–207, 1981. ©1981 IEEE).

The temperature dependence of the threshold current density varies as

$$J_{\text{th}} = J_{\text{th}0} e^{T/T_0}. \tag{4.90}$$

A typical example of the variation of the L–I curves for different temperatures is shown in Figure 4.12 [25], and the inset in this figure is also the plot of threshold current with temperature to obtain T_0.

This figure shows that as the temperature rises, the threshold current increases as the recombination becomes increasingly non-radiative.

The edge-emitting laser has a large number of longitudinal modes within the gain region, and at threshold all these are excited. Figure 4.13 shows these modes, and at threshold only those modes that have enough gain to neutralise the loss finally emerge. In this figure, two modes are shown to have this property, and in general it is possible that only one of these modes is likely to be dominant and the other becomes a secondary mode as the current and hence the gain is increased.

In Figure 4.14, the spectrum of the emission is plotted relative to the current excitation along the L–I curve, and this is also from [25].

The spectrum narrows from several modes close to threshold to a dominant mode with a few subsidiary modes at higher drive currents. The ratio of the intensity of the dominant mode to the next highest mode expressed in dBs is a measure of the mode suppression ratio (MSR). Since there is no guarantee that the Fabry–Perot laser with facet mirrors will produce a device with a single dominant mode with a large MSR of at least 20 dB, other techniques for mode selection, such as gratings in the active region or outside, may be used.

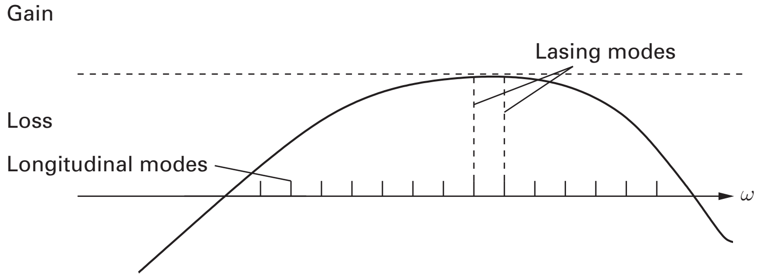


Fig. 4.13 The longitudinal modes of the laser within the gain region of the laser, showing the modes that are likely to lase when the gain is equal to the loss, including the mirror loss.

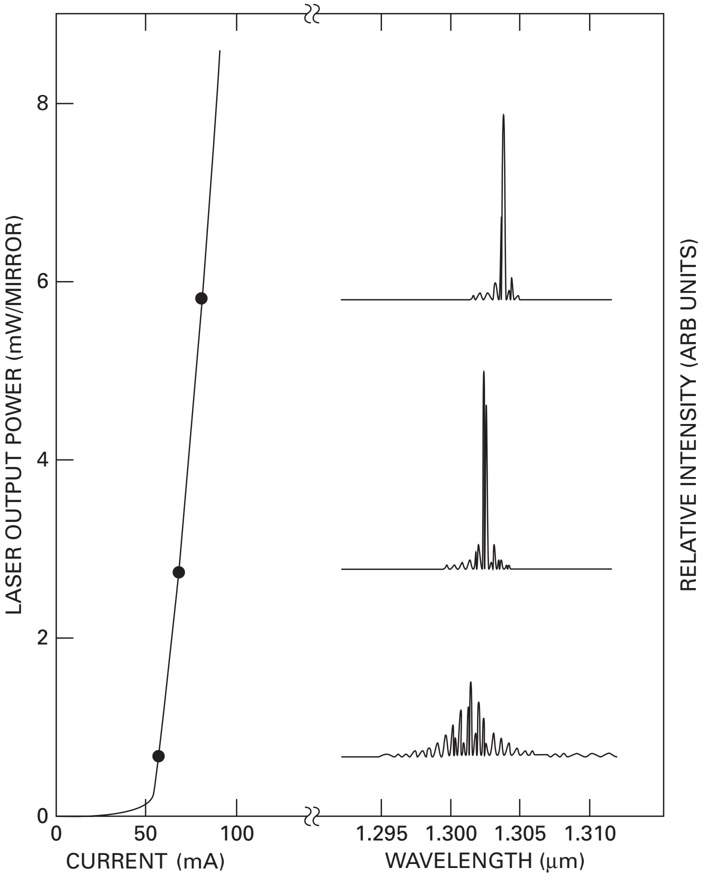


Fig. 4.14 The L–I curve of a buried heterostructure laser, together with the mode spectrum at different current levels of excitation (R. J. Nelson, R. B. Wilson, P. D. Wright, P. A. Barnes, N. K. Dutta, *IEEE Journal of Quantum Electronics*, Vol. 17, No. 2, pp. 202–207, 1981. ©1981 IEEE).

The laser illuminates the facets and these fields determine the near field pattern of the laser. It is usual to approximate the fields as a Gaussian distribution in the transverse direction, and a similar Gaussian distribution in the lateral direction (see [1]). The product of these Gaussians give rise to an elliptic distribution of the E-field on the facet. The far-field beam pattern is obtained using the usual methods by the spatial Fourier transform of the near field pattern [1].

Calculation of absorption, emission and gain

The calculation of the absorption, emission and gain in a semiconductor is a complex process, and will not be given here, as the method has been discussed in several textbooks, for example in [8]. Figure 4.15 shows the calculated gain/absorption spectra of InGaAsP for different levels of carrier injection, with the gain peak shifting as the carrier density increases [9].

Rate equations

Since lasers have a complex relationship between the injected carriers and the photons generated, the calculation of the dynamic response is more involved than that in LEDs. Essentially, the laser operates as an oscillator with a cavity, which is many wavelengths long, except in specific cases. Therefore, the cavity has many resonances corresponding to the expressions in Equation (4.33) in which the resonance frequencies are given by $\nu_{\mathrm{m}} = mc/2nL$, for different values of m. The corresponding radial frequency for the mth mode is assumed to be $\Omega_{\mathrm{m}} = 2\pi\nu_{\mathrm{m}}$, and the corresponding wave number k is given by

$$k_{\mathrm{m}} = \frac{n\Omega_{\mathrm{m}}}{c} = \frac{m\pi}{L}. \tag{4.91}$$

For the present, the subscript m is omitted for convenience. The laser radial frequency ω is undetermined, but nearly coincides with the cavity radial frequency Ω. The effective permittivity ϵ_{reff} is defined by the effective index of the laser guide, but the perturbation of gain and loss has to be added. However, the use of the effective permittivity or index reduces the wave equation to the one-dimensional form.

The injection of carriers into the laser causes a small change in index of the various layers, and for the active layer this takes the form:

$$\Delta n = bN, \tag{4.92}$$

where b is equal to $\partial n/\partial N$, and has a small negative value. Define

$$\beta_{\mathrm{c}} = -\frac{2k_0 b}{a} = -2k_0\left(\frac{\frac{\partial n}{\partial N}}{\frac{\partial g}{\partial N}}\right), \tag{4.93}$$

which results in $\beta_{\mathrm{c}} \propto b$, but is dimensionless, and also positive. β_{c} has been described as the anti-guiding parameter, or the linewidth enhancement factor.

The effective index is now given by

$$n_{\mathrm{eff}} = n + \Gamma\Delta n. \tag{4.94}$$

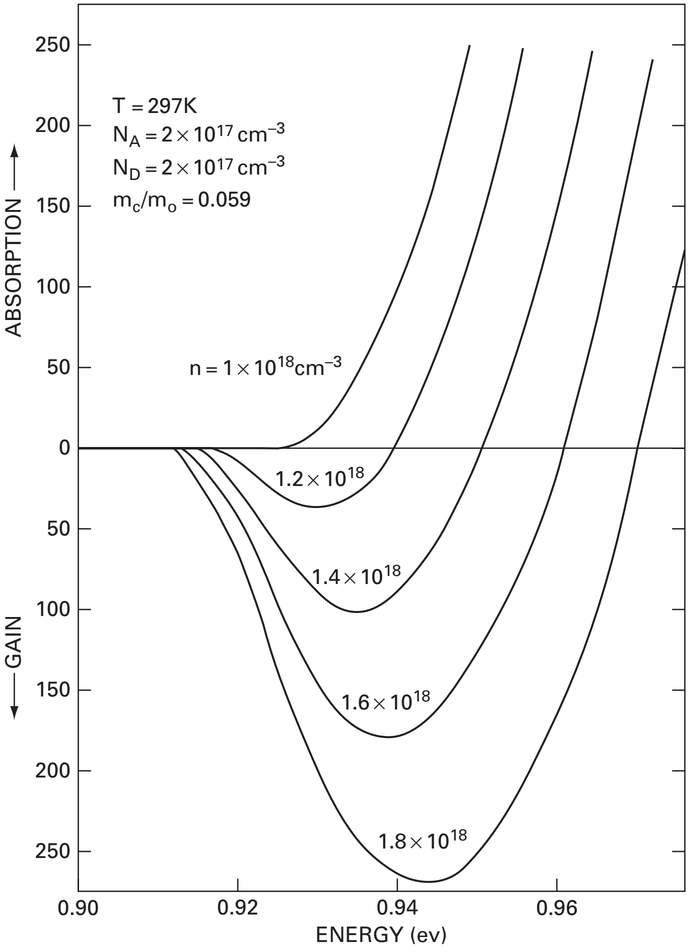


Fig. 4.15 The absorption/gain spectra for different levels of carrier injection. Reprinted with permission from N. K. Dutta, *Journal of Applied Physics*, Vol. 51, pp. 6095–6100, 1980. ©1980, American Institute of Physics.

This assumes that Δn is mainly in the active guide due to carriers, which is strictly not correct, as the carriers have to be injected from the top and bottom contacts. However, both the electrons and the holes end up in the active region to recombine and so this is an acceptable approximation. This assumption also simplifies equations to be derived below.

Also, the effective permittivity is given by

$$\epsilon_{\text{reff}} = (n + \Gamma \Delta n)^2 - j\frac{n\alpha}{k_0} \approx n^2 + 2\Gamma n \Delta n - \frac{jn\alpha}{k_0}. \tag{4.95}$$

The one-dimensional wave equation applied to a cavity requires a solution in the form $\sin k_{\text{m}} z$, with k_{m} defined in Equation (4.91), neglecting the effects of the end mirrors.

The electric field wave equation for the cavity is

$$\frac{d^2\mathcal{E}}{dz^2} - \frac{\epsilon_{\text{reff}}}{c^2}\frac{d^2\mathcal{E}}{dt^2} = 0, \tag{4.96}$$

and the solution is of the form:

$$\mathcal{E}(z,t) = \sum_i \sin(k_j z)\mathcal{A}(t)e^{j\omega t}, \tag{4.97}$$

where $\mathcal{A}(t)$ is assumed to be slowly varying in time compared to the light wave frequency. Assuming this is a single longitudinal mode laser, the summation and the subscripts i and j are dropped. Substituting this in the wave Equation (4.96) and neglecting the second time derivative of A

$$-k^2\mathcal{A} + \frac{\omega^2}{c^2}\epsilon_{\text{reff}}\mathcal{A} - \frac{2j\omega}{c^2}\epsilon_{\text{reff}}\frac{d\mathcal{A}}{dt} = 0. \tag{4.98}$$

Simplifying with $k^2 \simeq \Omega^2 n^2/c^2$:

$$\left(\frac{\omega^2}{c^2}\epsilon_{\text{reff}} - \frac{\Omega^2 n^2}{c^2}\right)\mathcal{A} - \frac{2j\omega}{c^2}n^2\frac{d\mathcal{A}}{dt} = 0. \tag{4.99}$$

However, since the laser frequency ω is very close to the cavity frequency Ω, then $(\omega^2 - \Omega^2) \approx 2\omega(\omega - \Omega)$. With further simplifications, substituting for ϵ_{reff} from Equation (4.95), this equation becomes

$$\frac{d\mathcal{A}}{dt} = -j\frac{n}{n_{\text{g}}}(\omega - \Omega)\mathcal{A} - \frac{j\omega}{n_{\text{g}}}\left(\Gamma\Delta n - j\frac{\alpha}{2k_0}\right)\mathcal{A}, \tag{4.100}$$

where n_{g} is the group index.

Separate this equation into its real and imaginary parts by substituting

$$\mathcal{A} = Ae^{j\phi} \tag{4.101}$$

into (4.100) to obtain for the real and imaginary parts:

$$\frac{dA}{dt} = -\frac{\alpha}{2k_0}A = \frac{1}{2}v_{\text{g}}(\Gamma g - \alpha_{\text{mir}} - \alpha_{\text{int}})A \tag{4.102}$$

and

$$\frac{d\phi}{dt} = -(\omega - \Omega) - \frac{\omega}{n_{\text{g}}}\Gamma\Delta n. \tag{4.103}$$

The rate equations may be written in terms of the photon density N_{ph} and carrier density N, or in terms of total number of photons and total number of carriers in the laser active volume, and both these approaches have been used in the literature. In the following derivations, the total number of photons and carriers in the laser volume are used. The total number of photons in the cavity is obtained from the equation:

$$S = \frac{\epsilon_0 n^2}{h\nu}\int \text{Vol}\,\mathcal{E}^2 dV, \tag{4.104}$$

and since $A^2 \propto S$, then multiplying Equation (4.102) by A, this equation becomes

$$\frac{dS}{dt} = (G - \gamma_{\text{p}})S + R_{\text{sp}}, \tag{4.105}$$

where

$$G = \Gamma v_g g \tag{4.106}$$

is the normalised gain, which is the stimulated emission rate. The photon decay rate is given by

$$\gamma_p = v_g(\alpha_{int} + \alpha_{mir}) = \frac{1}{\tau_p}. \tag{4.107}$$

The term R_{sp} has been added to account for the spontaneous emission in the lasing process. The spontaneous emission takes place over the whole laser cavity but only a small fraction β_{sp} couples into the waveguide mode, which is the integral part of the laser. R_{sp} may be written as

$$R_{sp} = \beta_{sp}\eta_{sp}\gamma_e N_t, \tag{4.108}$$

where

$$\gamma_e = (A_{nr} + B_r N + CN^2) = \frac{1}{\tau_e} \tag{4.109}$$

and β_{sp} is the spontaneous emission factor, which is usually a fitting parameter with values of 10^{-4}–10^{-5}, according to Agrawal [2]. The term $\eta_{sp} = B_r N/\gamma_e$ is the internal quatum efficiency, and is the fraction of carriers that recombine to emit photons through spontaneous emission.

N_t is the total number of carriers in the active volume obtained by integrating the carrier density N over the volume:

$$N_t = \int N dV. \tag{4.110}$$

The phase Equation (4.103) needs further simplification. From Equations (4.92) and (4.93)

$$\Delta n = -\frac{\beta_c}{2k_0} aN \approx -\frac{\beta_c}{2k_0}\Delta g. \tag{4.111}$$

The last term in Equation (4.103) becomes

$$\frac{\omega}{n_g}\Gamma\Delta n = -\frac{1}{2}\beta_c \Gamma v_g \Delta g. \tag{4.112}$$

Now $\Delta G = \Gamma v_g \Delta g$, and replacing ΔG by $G - \gamma_p$, Equation (4.103) becomes

$$\frac{d\phi}{dt} = -(\omega - \omega_{th}) + \frac{1}{2}\beta_c(G - \gamma_p). \tag{4.113}$$

The first term on the right-hand side in the above equation has been expanded in terms of the threshold frequency, and at threshold the cavity frequency Ω is very close to the threshold frequency.

The carrier rate equation given above in Equation (4.71) is

$$\frac{dN}{dt} = \frac{J}{qd} - \frac{N}{\tau_e} - R_{stim}N_{ph}. \tag{4.114}$$

The expression for R_{stim} is given by

$$R_{\mathrm{stim}} = \frac{c}{n_{\mathrm{g}}} g(N) = v_{\mathrm{g}} g(N). \tag{4.115}$$

Equation (4.75) defines the carrier lifetime τ_{e}. Since the volumetric values of N and N_{ph} are used in the photonic rate equations, integrating this equation over the active volume results in

$$\frac{dN_{\mathrm{t}}}{dt} = \frac{I}{q} - \gamma_{\mathrm{e}} N_{\mathrm{t}} - GS, \tag{4.116}$$

where $\gamma_{\mathrm{e}} = 1/\tau_{\mathrm{e}}$. The confinement factor is introduced in the last term to convert the expression to G, and the total photon number S is obtained from the integration of the photon density N_{ph}.

In practice, Fabry–Perot lasers with facet mirrors are longitudinally multimode, and therefore the rate equations apply to each individual mode, and have to be solved simultaneously. Thus, the rate equations become

$$\dot{S}_{\mathrm{m}} = (G_{\mathrm{m}} - \gamma_{\mathrm{p}}) S_{\mathrm{m}} + R_{\mathrm{sp}}(\omega_{\mathrm{m}}) \tag{4.117}$$

$$\dot{N}_{\mathrm{t}} = \frac{I}{q} - \gamma_{\mathrm{e}} N_{\mathrm{t}} - \sum_{\mathrm{m}} G S_{\mathrm{m}}. \tag{4.118}$$

The rate equations developed above allow the calculation of the $L{-}I$ curve, the longitudinal mode spectrum and the MSR, among other properties of the laser, provided the parameters are known. The dynamic behaviour which includes the turn-on delay, the modal behaviour, small- and large-signal modulation may also be calculated from these equations. The calculation of noise requires the addition of the noise sources to these equations, and these may also be obtained. In the following section, some aspects of the calculation methods for the steady-state and dynamic behaviour are discussed.

Steady-state and dynamic characteristics

Under steady-state conditions, for the single longitudinal mode laser in Equation (4.105), the time variation is zero. This makes this equation:

$$(G - \gamma_{\mathrm{p}}) S + R_{\mathrm{sp}}(\omega) = 0, \tag{4.119}$$

which becomes

$$S = \frac{R_{\mathrm{sp}}}{\gamma_{\mathrm{p}} - G}, \tag{4.120}$$

which states that the spontaneous emission photons in the cavity are created by the injected current. When the net stimulation emission rate G is nearly equal to the photon decay rate γ_{p}, then threshold is reached. The value of G is a little below the decay rate at threshold, and as the gain increases, G is asymptotic to γ_{p}, but the denominator should always remain positive.

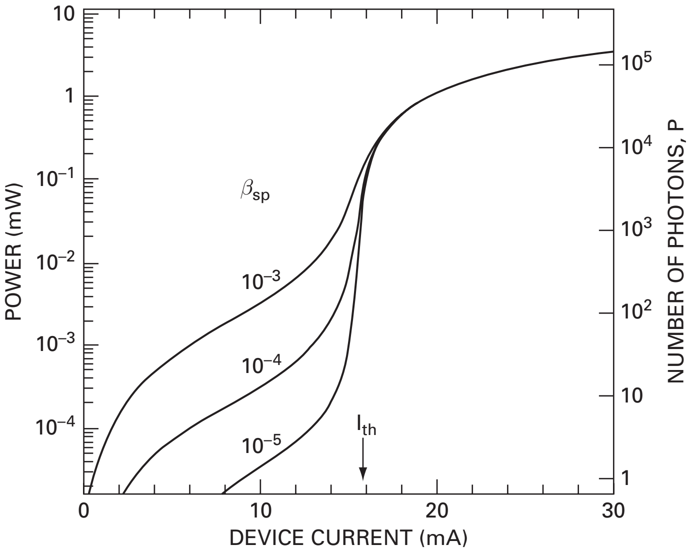


Fig. 4.16 Power output against current (G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*, Van Norstrand Rheinhold, 1993. ©Springer). With kind permission of Springer Science and Business Media.

Substituting in the carrier rate Equation (4.116)

$$\frac{I}{q} = \gamma_e N_t + R_{sp} \frac{G}{\gamma_p - G}. \tag{4.121}$$

This equation may be used to calculate the light intensity output against current, the L–I curve. Lee [19] has discussed the method of solution, as this is a non-linear equation. Figure 4.16 shows the results of the calculation of output power against current from [2] for the particular laser that has been modelled.

The time evolution of the build up in the carrier levels and the photon levels are shown in Figure 4.17 from Marcuse's paper [21] for a multimode laser when the current has been increased as a step function from 0 to 1.5 I_{th}. The rate equations are solved by numerical calculations outlined in [21], and the results are shown in the Figure 4.17.

The time delay of the build up to stimulated emission is typically of the order of 3 ns, and this is the reason why lasers are biased just below threshold current when they are pulsed on and off. The output power at this bias level is extremely low and so for most purposes may be regarded as negligible. Note the oscillations in the light power output at the relaxation oscillation frequency of the laser. This relaxation oscillation frequency f_r determines the maximum small-signal response of the laser, and this may be derived from the above rate equations. Also it may be shown that the relative intensity noise of the laser also peaks at the relaxation oscillation frequency, and thus modulation at or near this frequency needs to be avoided to improve the signal-to-noise ratio of the detected signal.

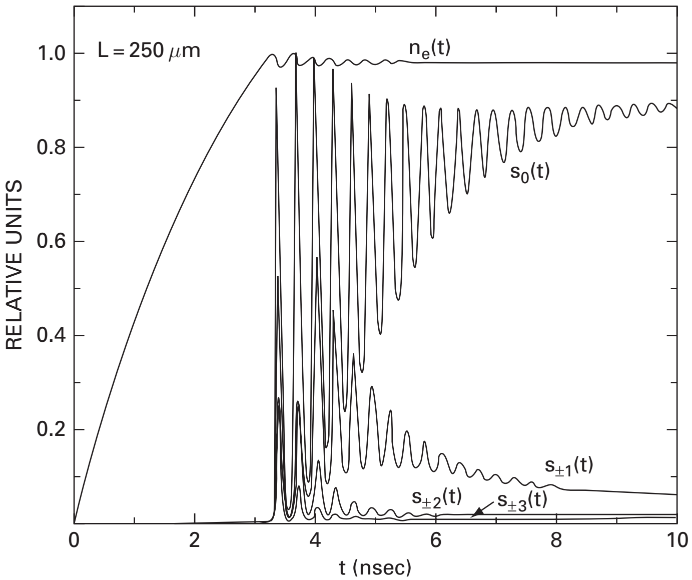


Fig. 4.17 Transient response for a laser in which the current is increased as a step function from 0 to 1.5 I_{th}, showing the carrier density and power in the fundamental mode and other modes (D. Marcuse, T. P. Lee, *IEEE Journal of Quantum Electronics* Vol. QE-19, No. 9, pp. 1397–1406, 1983. ©1983 IEEE).

This figure shows that the oscillations are damped and in the form of $e^{-(\Gamma_r \pm j\Omega_r)t}$, where Γ_r is the decay rate and Ω_r is the frequency of oscillation, which is the radial relaxation oscillation frequency in radians s^{-1}.

To calculate the relaxation oscillation frequency, the rate Equations (4.105) and (4.116), repeated here for convenience, are used.

$$\dot{S} = (G - \gamma_p)S + R_{sp} \tag{4.122}$$

$$\dot{N}_t = \frac{I}{q} - \gamma_e N_t - GS. \tag{4.123}$$

Consider perturbations to S and N_t given by δS and δN_t. Additionally,

$$G(N_t, S) \cong G + G_{N_t}\delta N_t + G_S \delta S, \tag{4.124}$$

where $G_{N_t} = \partial G/\partial N_t$ and $G_S = \partial G/\partial S$. Substituting in the rate equations, separating the perturbed terms, results in

$$\delta\dot{S} = -\Gamma_S \delta S + \left(G_{N_t} S + \frac{\partial R_{sp}}{\partial N_t}\right)\delta N_t \tag{4.125}$$

$$\delta\dot{N}_t = \Gamma_{N_t}\delta N_t - (G + G_S S)\delta S, \tag{4.126}$$

where

$$\Gamma_S = \frac{R_{sp}}{S} - G_S S, \tag{4.127}$$

and

$$\Gamma_{N_t} = \gamma_e + N_t\left(\frac{\partial\gamma_e}{\partial N_t}\right) + G_{N_t}S. \tag{4.128}$$

To solve these equations, the assumption is that these perturbations decay at the rate shown in the time evolution result in Figure 4.16.

Thus, let

$$\delta S = \delta S_0 e^{-\mathrm{ht}} \tag{4.129}$$

$$\delta N_t = \delta N_{t0} e^{-\mathrm{ht}}, \tag{4.130}$$

where δS_0 and δN_{t0} are the initial values of the perturbations, and

$$h = \Gamma_r \pm j\Omega_r. \tag{4.131}$$

The real part is the decay rate

$$\Gamma_r = \frac{1}{2}(\Gamma_{N_t} + \Gamma_S), \tag{4.132}$$

and the radial relaxation oscillation frequency, after some approximations, is given by

$$\Omega_r \approx (GG_{N_t}S)^{1/2}. \tag{4.133}$$

With further substitutions for

$$S = \frac{(I - I_{th})}{(qG)} \tag{4.134}$$

$$I_{th} = q\gamma_e N_{t-th} \tag{4.135}$$

$$G_{N_t} = \frac{\Gamma v_g a}{\mathrm{Vol}}, \tag{4.136}$$

the expression for the relaxation frequency becomes

$$\Omega_r = \left[\frac{1 + \Gamma v_g a N_0 \tau_p}{\tau_p \tau_e}\left(\frac{I}{I_{th}} - 1\right)\right]^{1/2}. \tag{4.137}$$

where N_0 is the transparency carrier density. Derivation of the small-signal response shows that Ω_r is the key parameter for high-speed lasers; the larger this is, the higher the laser response frequency. The term $\Gamma v_g a N_0 \tau_p$ evaluates to a number close to unity, and therefore may not be neglected. However, making τ_p smaller increases the value of Ω_r. This reduction in τ_p may be obtained by decreasing the facet mirror reflectivities by coatings, or alternatively by making the laser guide more lossy. The parameter τ_e which is the carrier recombination rate is initially determined by the material. After threshold this recombination rate becomes shorter, and little can be done to reduce it further. Reducing I_{th} and increasing the ratio I/I_{th} also increases the magnitude of Ω_r.

The small-signal modulation response calculation uses the rate equations, but in this case the phase Equation (4.103) is also featured. The derivation assumes that the laser is biased above threshold at some current I_b, and the small-signal current modulation term, modulation at ω_m, is given by $I_m \sin(\omega_m t)$. The expression for the total current is

$$I(t) = I_\mathrm{b} + I_\mathrm{m}\sin(\omega_\mathrm{m}t), \tag{4.138}$$

with the assumption that $I_\mathrm{m} \ll (I_\mathrm{b} - I_\mathrm{th})$, which implies small-signal modulation.

The total cavity photons may be written as

$$S(t) = S_\mathrm{b} + S_\mathrm{m}\sin(\omega_\mathrm{m}t + \theta_\mathrm{m}), \tag{4.139}$$

where S_b is the cavity photons at the bias current of I_b and the second term is the sinusoidal time varying component, with a phase term which usually lags the current component. The carriers in the cavity have a similar expression given by

$$N_\mathrm{t}(t) = N_\mathrm{t-b} + N_\mathrm{t-m}\sin(\omega_\mathrm{m}t + \xi_\mathrm{m}), \tag{4.140}$$

with the phase delay ξ_m, which differs from the photon number delay.

Substituting in the rate equations, the solutions for the various small-signal terms may be obtained. The small-signal photons in the cavity, which is also a measure of the output power, may be shown to be

$$S_\mathrm{m} = \frac{G_{\mathrm{N_t}}SI_\mathrm{m}/q}{[(\omega_\mathrm{m}^2 - \Omega_\mathrm{r}^2 - \Gamma_\mathrm{r}^2)^2 + 4\omega_\mathrm{m}^2\Gamma_\mathrm{r}^2]^{0.5}}, \tag{4.141}$$

and the phase lag term for the photons is given by

$$\theta_\mathrm{m} = \tan^{-1}\left(\frac{2\Gamma_\mathrm{r}\omega_\mathrm{m}}{\omega_\mathrm{m}^2 - \Omega_\mathrm{r}^2 - \Gamma_\mathrm{r}^2}\right). \tag{4.142}$$

The modulated cavity photons or equivalently the modulated output light power is almost constant when $\omega_\mathrm{m} \ll \Omega_\mathrm{r}$, then peaks to a maximum near Ω_r, and falls off for $\omega_\mathrm{m} > \Omega_\mathrm{r}$. Note that the laser is a forward-biased p–i–n diode, and therefore additional series resistance due to contact resistance and parasitic inductance and capacitance from the bond wire and pads results in further degradation of the response. This is shown [33] in Figure 4.18. It may be shown that the 3 dB bandwidth is given by [1], assuming $\Gamma_\mathrm{r} \ll \Omega_\mathrm{r}$:

$$f_\mathrm{3dB} = \frac{\sqrt{3}\Omega_\mathrm{r}}{2\pi}. \tag{4.143}$$

When the laser current is modulated, the increase and decrease in the current result in the laser effective index varying inversely as the current according to Equation (4.92). Thus, the longitudinal mode cavity changes its electrical length since the index changes with current injection. As the modulation takes place the frequency of emission of the laser keeps changing in synchronism with this modulation. Using the phase rate equation, it may be shown that the change in frequency results in a frequency chirp given by [1]:

$$\delta\nu(t) = \frac{1}{2\pi}\frac{d\phi}{dt} = \frac{\beta_\mathrm{c}}{4\pi}\left[G_{N_\mathrm{t}}(N_\mathrm{t} - N_\mathrm{t-0}) - \frac{1}{\tau_\mathrm{p}}\right], \tag{4.144}$$

where β_c is the linewidth broadening factor of Equation (4.93) and

$$N_\mathrm{t-0} = \int N_0 dV. \tag{4.145}$$

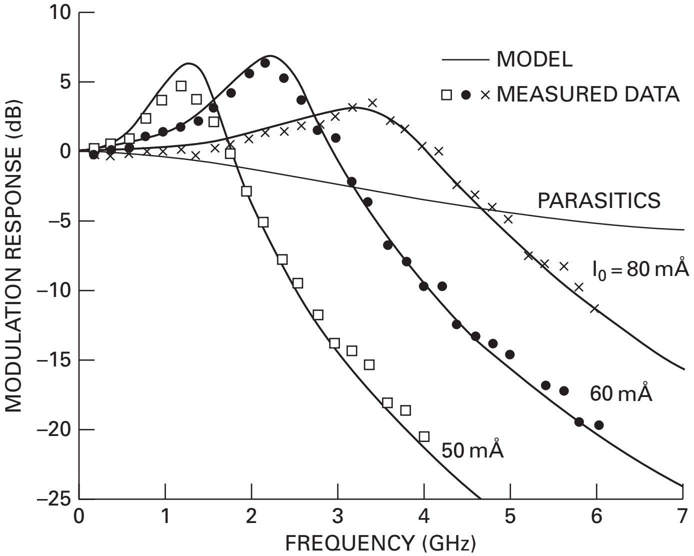


Fig. 4.18 Modulation response for different current drives, showing the effect of parasitics, from [33] with permission (R. S. Tucker and I. P. Kaminow, *Journal of Lightwave Technology*, Vol. 2, No. 4, pp. 385–393, 1984. ©1984 IEEE, OSA).

This linewidth broadening that occurs with chirp is unacceptable for long haul fibre optic systems, and therefore an external modulator is used with dc current into the laser.

The process of obtaining a single longitudinal mode requires a grating in the waveguide either in the whole of the gain section which makes it the distributed feedback (DFB) laser or at the ends as the grating acts as a frequency selective reflector which makes it a distributed Bragg reflector (DBR) laser. This does ensure that a single emission line emerges from the laser, and feeds into an external modulator.

Laser noise

The semiconductor laser has noise associated with the emission, and the primary source of noise is the spontaneous emission, followed by the carrier recombination noise, which is essentially shot noise. Since this is in a cavity, the oscillations are affected by an amplitude or intensity noise component called the *Relative intensity noise* (RIN), and a phase noise component which affects the linewidth of the emission. The method of analysis is to include the noise terms in the rate equations, and traditionally these are the Langevin noise components, which added to the rate equations. These noise terms are assumed to be Gaussian, with a zero mean, and the correlations are assumed to be Markovian, which simplifies the equations. Thus, the rate equations become

$$\frac{dS}{dt} = (G - \gamma_{\mathrm{p}})S + R_{\mathrm{spon}} + F_{\mathrm{S}}(t) \tag{4.146}$$

$$\frac{dN_{\mathrm{t}}}{dt} = \frac{I}{q} - \gamma_{\mathrm{e}}N_{\mathrm{t}} - GS + F_{\mathrm{N_t}}(t) \tag{4.147}$$

$$\frac{d\phi}{dt} = -(\omega - \omega_{\mathrm{th}}) + \frac{1}{2}\beta_{\mathrm{c}}(G - \gamma_{\mathrm{p}}) + F_{\phi}(t), \tag{4.148}$$

where $F_S(t)$, F_{N_t} and $F_\phi(t)$ are the Langevin noise components. The mean of each of these components is zero, but their auto- and cross-correlation terms result in so-called *diffusion components* which have specific values. Thus,

$$< F_i(t) > = 0 \tag{4.149}$$

$$< F_i(t), F_j(t') > = 2D_{ij}\delta(t - t') \tag{4.150}$$

where i, j, k are S, N_t, ϕ, and it may be shown [2] that the dominant contributions are from the auto-correlation terms $D_{SS} = R_{sp}S$ and $D_{\phi\phi} = R_{sp}/4S$. The method of solution is by perturbations of all three variables; these lead to equations which give rise to the auto-correlation factors for the perturbations, and consequently noise spectral density is obtained through the Fourier transform. Thus, the spectral density of the photon number or the light intensity leads to the relative intensity noise. The intensity auto-correlation is given by A_{SS}, and thus

$$A_{SS}(\tau) = < \delta S(t)\delta S(t + \tau) > \bar{S}^2, \tag{4.151}$$

where $\bar{S}$ is the time average photon intensity and $\delta S = S - \bar{S}$, represents the fluctuations, and hence the noise. Fourier transform of A_{SS} gives the RIN

$$\mathrm{RIN} = \int_{-\infty}^{\infty} A_{SS}(\tau)e^{j\omega t}dt, \tag{4.152}$$

and it may be shown that [2]

$$\mathrm{RIN} = \frac{2R_{sp}[(\Gamma_{N_t}^2 + \omega^2) + G_{N_t}S^2(1 + \gamma_e N_t/R_{sp}S) - 2\Gamma_{N_t}G_{N_t}S]}{S\left[(\Omega_r - \omega)^2 + \Gamma_r^2\right]\left[(\Omega_r + \omega)^2 + \Gamma_r^2\right]}. \tag{4.153}$$

From this expression, the RIN may be plotted as a function of frequency as shown in Figure 4.19 [2]. The RIN peaks at the relaxation oscillation frequency, and therefore it follows that direct modulation of lasers should not be performed close to f_r.

It may also be shown that the signal-to-noise ratio is given by

$$\mathrm{SNR} = \left(\frac{2\Gamma_r S}{R_{sp}}\right)^{0.5}. \tag{4.154}$$

Using the phase rate equation with the Langevin noise terms (Equation (4.148)), it may be shown [2] that the frequency noise spectral density is given by the expression in the following equation. The frequency noise is the phase noise integrated over time:

$$A_{ff} \cong \frac{R_{sp}}{2S}\left(1 + \frac{\beta_c^2\Omega_r^4}{\left[(\Omega_r^2 - \omega^2)^2 + (2\omega\Gamma_r)^2\right]}\right). \tag{4.155}$$

The above expression shows that the term is flat in the region $\omega \ll \Omega_r$, and similar to the RIN, it peaks at the relaxation oscillation frequency Ω_r, and then falls off rapidly.

The linewidth fluctuation is obtained by considering the electric field fluctuations, and including both the amplitude and phase fluctuations. After considerable simplifications, the FWHM $\Delta\omega = 2\pi\Delta f$ of the line, assumed to be a Lorentzian, is given by [2]

$$\Delta f = \frac{R_{sp}(1 + \beta_c^2)}{4\pi S}. \tag{4.156}$$

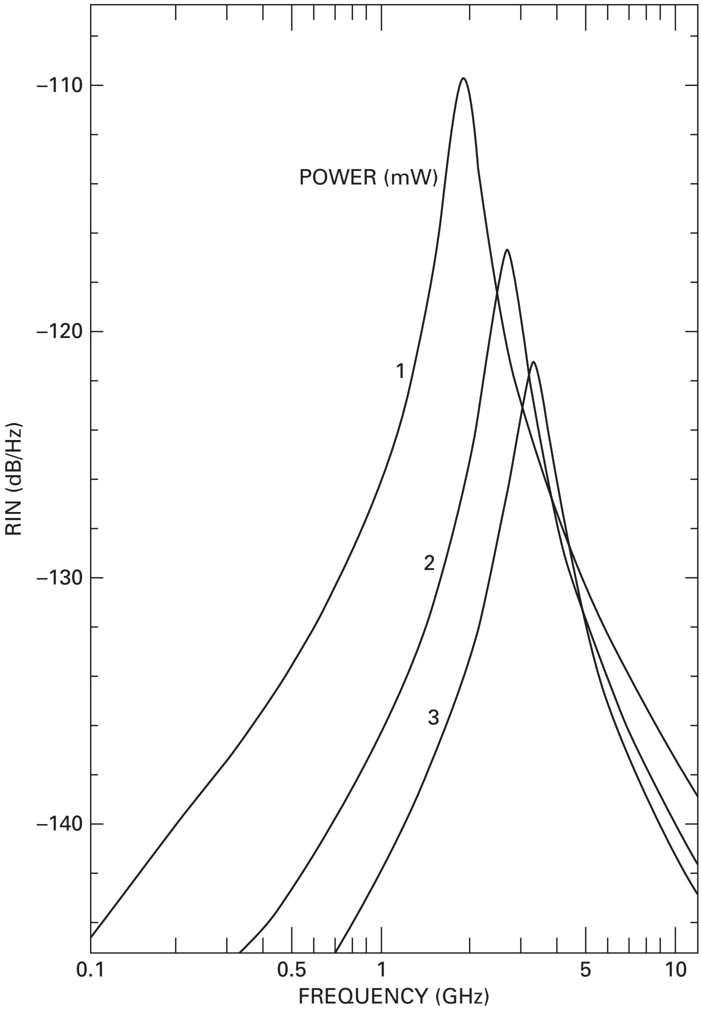


Fig. 4.19 Calculated RIN for a typical 1.3 μm InGaAsP laser at different output power levels (G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*, Van Norstrand Rheinhold, 1993. ©Springer). With kind permission of Springer Science and Business Media.

This may be written as

$$\Delta f = (1 + \beta_c^2)\Delta f_0, \tag{4.157}$$

where

$$\Delta f_0 = \frac{R_{sp}}{4\pi S}, \tag{4.158}$$

which is the unperturbed linewidth of the laser.

Quantum well and quantum dot lasers

The gain medium considered so far has been assumed to be bulk material, where gain is in the region of $100\,\mathrm{cm}^{-1}$. The use of quantum wells or quantum dots as the gain

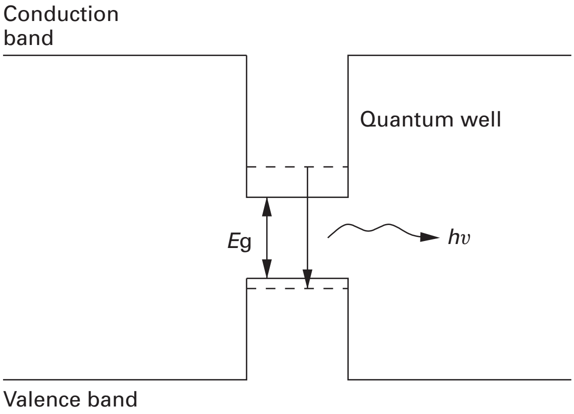


Fig. 4.20 A single quantum well, showing the lowest energy levels in the conduction and valence bands.

medium results in considerable reduction in the number of carriers injected to obtain population inversion and thus the threshold current.

In traditional double heterostructure lasers, the active region is between higher bandgap materials, and the holes and electrons recombine in this region to emit light. To obtain stimulated emission, the energy difference between quasi-Fermi levels for the holes and electrons in this active region needs to be greater than the band gap of the active region. Then population inversion takes place and the light output is dominated by stimulated emission. When the active region becomes very narrow, comparable to the de Broglie wavelength in the material, quantum confinement occurs, and this region becomes a quantum well.

In a quantum well the energy level in the direction across the well, the x direction in Figure 4.20, is quantised, and is a continuum in the other directions, y and z directions. The depth of the well in the conduction band is ΔE_{c} and in the valence band is ΔE_{v}, and the emission energy for the recombination from $E1c$ to $E1v$ is greater than (E_{g}). These levels are calculated by solving the time-independent Schrödinger equation of the well with finite barriers for the conduction band electrons, and for the valence band heavy holes and light holes. The bulk material degeneracy of the heavy hole and light hole bands is removed in the quantum well with the heavy hole band being uppermost, followed by the light hole band, and at a lower level, the split-off band as shown schematically in Figure 4.21. The effective masses of the heavy and light holes are different and therefore the levels are also different. The normal wavelength of the lasing level is from $E1c$ to $E1hh$ which is a TM mode wave.

The density of states per unit volume for the quantum well is given by

$$\rho_{\mathrm{qw-ci}} = \frac{\pi m_{\mathrm{ci}}}{h^2 L_{\mathrm{qw}}}, \tag{4.159}$$

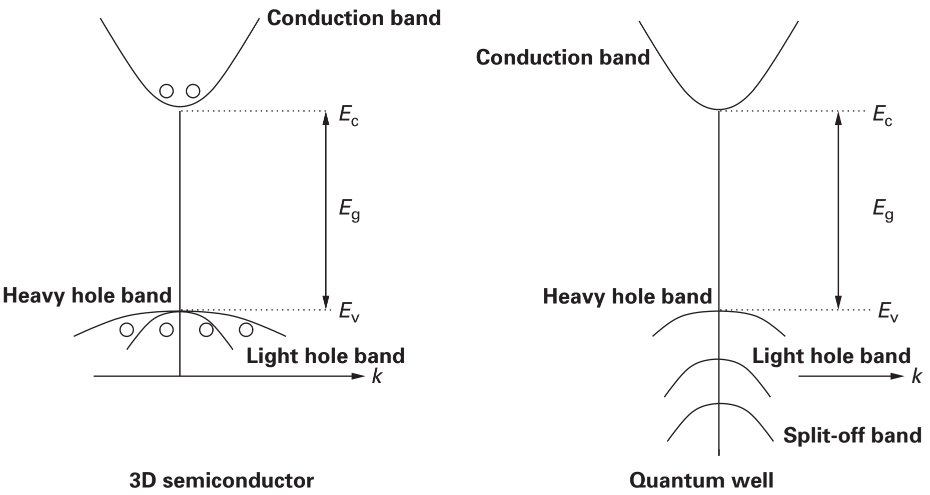


Fig. 4.21 Schematic band diagram of a bulk semiconductor and a quantum well.

where m_{ci} is the effective mass of the electrons in ith sub-band of the quantum well and L_{qw} is the width of the quantum well. The density of states for the valence band is identical except that the appropriate effective masses are used.

The density of states for the conduction band of bulk material, assuming parabolic bands, is of the form:

$$\rho_{\mathrm{c}}(E) = 4\pi\left(\frac{2m_{\mathrm{c}}}{h^2}\right)^{3/2} E^{1/2}. \tag{4.160}$$

Comparing the two, it is apparent that the density of states for the quantum well is independent of energy. Furthermore, it may be shown that the density of states for the quantum well is much smaller than that of the bulk material. Thus, the current density for threshold is generally much smaller than that for bulk material. The current state of the art structures have current densities of the order of 100 A · cm^{-2} per quantum well, in contrast to bulk structures which have threshold current densities of almost ten times this figure. The maximum gain of quantum wells is estimated at between 1000 cm^{-1} and almost 10 000 cm^{-1}. However, the confinement factor of a single quantum well is of the order of 0.01–0.02, which makes the net gain of the order of 100 cm^{-1}. In contrast, bulk structures have gains of the order of 100 cm^{-1}, and confinement factors of about 0.20–0.40. In multiple quantum well structures, the spacing between the wells is chosen to be sufficiently large so that the wells are not coupled. Typical well widths are about 10 nm, and the barriers between them are also of similar widths.

Quantum well lasers come in different forms, from the single quantum well laser to the multi-quantum well laser, and these may be incorporated in the waveguides for lasers discussed earlier. Figure 4.22 sketches some of these structures from the graded index single quantum well laser to the separate confinement single and multiple quantum well lasers.

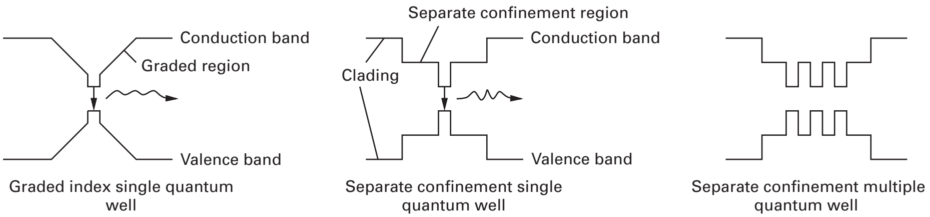

Conduction band
Graded region
Valence band
Graded index single quantum well
Separate confinement region
Conduction band
Clading
Valence band
Separate confinement single quantum well
Separate confinement multiple quantum well

Fig. 4.22 Schematic band diagrams of the graded index single quantum well laser structure, and separate confinement single and multiple quantum well laser structures.

Strained quantum well lasers have the quantum wells with compressive or tensile strain, and sometimes the barriers between wells are also strained. The compressive strain increases the band gap, and also the separation between the heavy hole and light hole bands, and the emission is in TM mode. With tensile strain the band gap is reduced, and the gap between the light hole and heavy hole bands is reduced, and may even push the light hole band above the heavy hole band. With the light hole band above the heavy hole band, the lasers emit in the TE mode. Strained quantum well lasers have lower threshold current densities and are sometimes preferred over unstrained well structures.

Quantum well lasers show better performance compared to the bulk lasers with reduced threshold current densities, better linewidths and better chirp performance. With the distributed Bragg grating to obtain single longitudinal mode with further modifications, these have become the lasers of choice for fibre optical systems.

Quantum dot lasers are a class of lasers that use confinement to improve on the performance of the quantum well lasers. In these devices, the gain is through current injection into quantum dots. A quantum dot is a three-dimensional structure in which every dimension is less than the de Broglie wavelength in the material, and quantum confinement occurs. The density of states in this case is a delta function, and in principle the number of carriers to be injected is very small. Quantum dots are created by the Stranski–Krastinov growth technique [5], with further modifications, by allowing a strained layer to relax, and this results in self-organised dots. These dots need to have wetting layers to make contacts, and since single layer of dots does not provide sufficient gain this composite of dot layers and wetting layers need to be repeated many times. Lasers made of these active layers show reduced threshold current densities, linewidth reduction, and reduced chirp [11] compared to quantum well lasers. The major problem with these lasers is that the dots vary in size and so the distribution affects the linewidth.

Vertical cavity surface emitting lasers (VCSELs) are another class of lasers that use quantum wells for the gain medium. In this case, the cavity is usually one wavelength long, and the quantum well is placed in the middle so as to provide maximum gain to the standing wave within the cavity. Since the gain is so small, the mirror reflectivities need to be extremely high, as close to unity as possible. This is obtained by a multi-layer dielectric stack mirror with reflectivities of 0.997 for the top output mirror and 0.999 for the lower mirror. A schematic diagram of a VCSEL is shown in Figure 4.23. Currently, VCSELs have been built in the visible, at 840 nm, 980 nm, and also at the longer wavelengths of 1300 nm and 1550 nm. These devices may be built with sub-milliampere

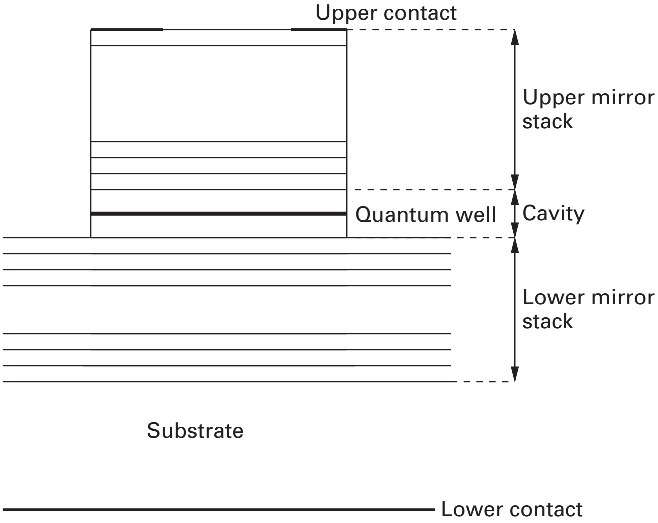


Fig. 4.23 Schematic diagram of a single quantum well vertical cavity laser.

threshold current. These lasers are also one of the most efficient devices with very high wall plug efficiencies. The 840 nm VCSELs are used in data-communication applications where data is transferred between computers, processors and storage media. The linewidth of these lasers is generally high, of the order of 0.1 nm, and direct modulation of these results in linewidth broadening. These lasers find application in coarse wavelength division multiplexed systems, where channel spacing may be as high as 40 nm, and in the 10 GHz Ethernet applications.

High-speed lasers

High-speed lasers need to have high relaxation frequencies, Ω_r or f_r. The expressions for Ω_r are repeated here:

$$\Omega_r \cong (G G_{N_t} S)^{1/2}. \tag{4.161}$$

With the substitution for S,

$$S = \frac{(I - I_{th})}{(qG)} \tag{4.162}$$

results in

$$\Omega_r = \left[\frac{G_{N_t}(I - I_{th})}{q}\right]^{0.5}. \tag{4.163}$$

Also note that

$$I_{th} = q\gamma_e N_{t-th} \tag{4.164}$$

$$G_{N_t} = \Gamma v_g a/\mathrm{Vol}, \tag{4.165}$$

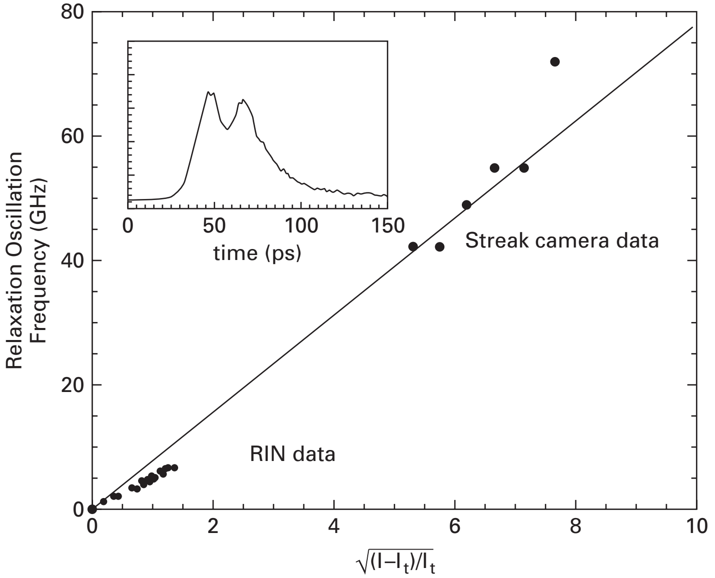


Fig. 4.24 Plot of relaxation oscillation frequency against square root of normalised current. Reprinted with permission from D. Tauber, G. Wang, R. S. Geels, J. E. Bowers, L. A. Coldren, *Applied Physics Letters*, Vol. 62, No. 4, 1993. Copyright 1993, American Institute of Physics.

the expression for the relaxation frequency becomes

$$\Omega_r = \left[\frac{1+\Gamma v_g N_0 \tau_p}{\tau_p \tau_e}\left(\frac{I}{I_{th}}-1\right)\right]^{1/2}. \tag{4.166}$$

Equation (4.161) suggests that high values of G_{N_t} are required, which is essentially that $\partial g/\partial N$ is high per unit volume, and S needs to be high, which implies that the photon density N_{ph} should be high in the cavity. From Equation (4.163), it follows that I_{th} should be as low as possible. The small-signal derivation in Equation (4.145) shows again that f_{3dB} is determined by Ω_r. Thus, this equation may be written as

$$f_{3dB} \approx \frac{\sqrt{3}\Omega_r}{2\pi}. \tag{4.167}$$

Ideally, the quantum dot lasers should have the highest relaxation oscillation frequency. However, the problem of the parasitics of the wetting layer and other issues [11] result in these lasers not being as fast as the quantum well tunneling injection lasers discussed below.

Of the lasers discussed above, the lowest threshold current device is the VCSEL, and early measurements have shown that the relaxation oscillation frequency measured by streak camera [31] was 84 GHz shown in Figure 4.24. However, the problem is that the parasitics of the device in its present form, with the current passing through the entire top and bottom mirrors, are large. Thus, the device parasitics should be made as low as possible, as well as the packaging parasitics, as otherwise the modulation signal will be attenuated severely at high frequencies. A high-speed VCSEL operating at 35 GHz

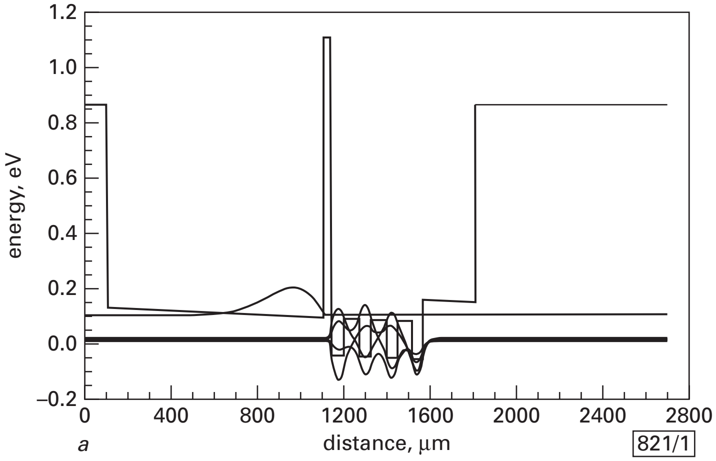


Fig. 4.25 Plot of the band diagram of the conduction band of the tunnelling injection laser (X. Zhang, A. L. Gutierrez-Aitkens, D. Klotzkin, P. Bhattacharya, C. Caneau, R. Bhatt, *Electronics Letters*, Vol. 32, No. 18, pp. 1715–1717, 1996. ©1996 IEEE).

[6] has the laser drive and modulation current pass through the top mirror but only through a small number of layers of the lower mirror to achieve this result.

The approach by Bhattacharya's group [35] has resulted in tunneling injection quantum well lasers that have the potential for very high modulation rates. The claim is that when the electrons travel from the cladding and fall into the separate confinement region, they gain energy and become hot. The electrons diffuse across the separate confinement region (SC), and when they fall into the quantum wells, they become even more hot. The holes having a large mass are not as mobile, and therefore are largely unaffected by these changes in potential. The approach is to have a reservoir of electrons in the SC region, and these tunnel out into the SC region, and also into the quantum wells with only small gains in energy, to allow high modulation rates. The theory discussed above for Ω_r does not account for the temperature of the carriers in determining its value. A more elaborate theory would do this, but only qualitative explanations are therefore available.

Figure 4.25 shows the conduction band diagram of the device. Careful design of the reservoir, its width and its level with respect to the SC region and the quantum wells are necessary. Again the device parasitics need to be low. The results are indeed impressive, 3 dB bandwidths of 100 GHz, and these are probably the fastest lasers built to date. Note that the structure used in high-speed quantum dot lasers has also used the tunnelling structure, but these are not as fast as the quantum lasers discussed above.

4.3 Photodetectors

Photodetectors are devices used to convert light signals into electrical versions. The performance of the different types of detectors is determined by their quantum efficiency, their frequency response and responsivity. In this chapter only solid state detectors are considered, and the slower detectors are only briefly discussed. The photomultiplier, which is widely used as a sensitive and fast (of the order of 100 MHz) detector, will not be discussed here. Solid state detectors considered here are the extrinsic kind, in which the photon energy is close to the band gap of the semiconductor material used for the detector. Intrinsic photodetectors are used for detection of light with energies below the band gap, and depend on deep level traps, or with different energy levels in a quantum well, but these are not discussed here. In this chapter, the detectors considered are photoconductors, and the junction devices which include the p–i–n diode, the metal–semiconductor–metal (MSM) photodetector and the avalanche photodetector (APD). The photoconductor is a slow device and also noisy, but its simplicity is an attractive feature.

4.3.1 Preliminaries

Consider a semiconductor photodetector which absorbs photons with energies at or above E_g, the bandgap energy of the semiconductor. Suppose the incident optical power is given by P_{in}, and it is assumed that all the incident photons enter the semiconductor. Suppose the photocurrent generated as a result of this incident optical power is given by I_p, then the relationship between P_{in} and I_p is

$$I_p = RP_{in}, \tag{4.168}$$

where R is the responsivity of the photodetector in units of AW^{-1}.

The quantum efficiency of the detector η may be defined as ratio of the number of hole–electron pairs generated to the number of incident photons, and is given by

$$\eta = \frac{\frac{I_p}{q}}{\frac{P_{in}}{h\nu}} = \frac{h\nu}{q}R. \tag{4.169}$$

Thus, R may be written as

$$R = \frac{q\eta}{h\nu} = \frac{\eta\lambda q}{hc}. \tag{4.170}$$

Suppose the thickness of the semiconductor is w, and the absorption coefficient is α Np m^{-1}, then the transmitted optical power escaping from the semiconductor is given by

$$P_{tr} = P_{in}e^{-\alpha w}. \tag{4.171}$$

Thus, the absorbed power is given by

$$P_{abs} = P_{in} - P_{tr} = P_{in}(1 - e^{-\alpha w}). \tag{4.172}$$

Since every absorbed photon creates one hole–electron pair, the quantum efficiency is given by

$$\eta = \frac{P_{abs}}{P_{in}} = (1 - e^{-\alpha w}), \tag{4.173}$$

which assumes that all the incident photons enter the semiconductor with no reflection.

4.3.2 Photoconductor detectors

The photoconductor detector depends on the increase in conductivity of a semiconductor when illuminated with photons of energy above the band gap. The absorbed light creates hole–electron pairs, which increases the conductivity, and with an applied bias, the excess carriers drift to the appropriate electrodes, and constitute an increase in current. Holes and electrons created by the light may be swept out before they recombine by the applied bias field. Alternatively, they recombine as they drift towards the appropriate electrode. The electrons are swept out faster than the holes, and to maintain charge neutrality, more electrons are injected, and this constitutes gain.

A typical photoconductor detector takes the form of a slab of material of thickness a, width b and length L, with ohmic contacts at the sides as shown in Figure 4.26.

The dark current flowing in the slab is given by

$$I = qab(n\mu_{\rm n} + p\mu_{\rm p})\frac{V}{L}, \tag{4.174}$$

where n and p are the free electron and hole number densities, $\mu_{\rm n}$ and $\mu_{\rm p}$ are the mobilities of the electrons and holes respectively, and V is the applied voltage across the slab. When illuminated, the conductivity increases due to the electron–hole pairs created by the photons. Thus, the current when the slab is illuminated is given by

$$(I + \Delta I) = qba[(n + \Delta n)\mu_{\rm n} + (p + \Delta p)\mu_{\rm p}]\frac{V}{L}, \tag{4.175}$$

where I is the dark current, assumed to be small. The increase in current due to the illumination is

$$\Delta I = qab(\Delta n\mu_{\rm n} + \Delta p\mu_{\rm p})\frac{V}{L}. \tag{4.176}$$

Suppose the minority carriers are electrons, then the rate equation for the electrons takes the form:

$$\frac{d\Delta n}{dt} = \frac{\eta P}{h\nu abL} - \frac{\Delta n}{\tau}, \tag{4.177}$$

where η is the quantum efficiency, which determines the number of hole–electrons generated per photon, usually taken to be unity, and P is the optical power absorbed. Note that an equal number of holes as electrons are generated, but remain as the majority carrier. Under steady state, the time variation is set to zero, and then

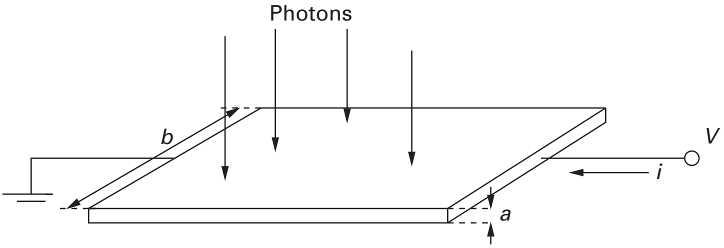


Fig. 4.26 Schematic diagram of a photoconductor slab: length between contacts L, width b and thickness a.

$$\Delta n = \frac{\eta P \tau}{h\nu abL}. \tag{4.178}$$

This assumes that the photon-generated hole-electron pairs are in the volume of the slab, and the volume is abL. The electron current due to these excess electrons is given by

$$\Delta I_{\mathrm{n}} = q\Delta n\mu_{\mathrm{n}}ba\frac{V}{L} = \frac{q\eta PG}{h\nu}, \tag{4.179}$$

where G is the gain of the device. Substituting for Δn from Equation (4.178), G is defined as

$$G = \frac{\mu_{\mathrm{n}}\tau V}{L^2} = \frac{\tau}{t_{\mathrm{tr,n}}}, \tag{4.180}$$

where the electron transit time $t_{\mathrm{tr,n}}$ is given by

$$t_{\mathrm{tr,n}} = \frac{L}{v_{\mathrm{n}}} = \frac{L}{\mu_{\mathrm{n}}\mathcal{E}} = \frac{L^2}{\mu_{\mathrm{n}}V}. \tag{4.181}$$

In addition to the electrons, adding the motion of the holes results in this equation becoming

$$G = \frac{(\mu_{\mathrm{n}} + \mu_{\mathrm{p}})\tau V}{L^2} = \frac{\mu_{\mathrm{n}} + \mu_{\mathrm{p}}}{\mu_{\mathrm{n}}}\frac{\tau}{t_{\mathrm{tr,n}}}. \tag{4.182}$$

Note that since $\mu_{\mathrm{h}} \ll \mu_{\mathrm{n}}$, the expression for gain is that given in Equation (4.180).

Making L as small as possible increases the gain and also reduces the response time. If the carriers are swept out before they recombine, then the electrons reach the ohmic contact before the holes. To maintain charge neutrality, extra electrons are injected into the photoconductor. If sweep out of the carriers occurs, then the lifetime is the transit time of the carriers and is given by

$$t_{\mathrm{tr,n,p}} = \frac{L}{\mu_{\mathrm{n,p}}\left(\frac{V}{L}\right)} = \frac{L^2}{\mu_{n,p}V}. \tag{4.183}$$

Since the mobility of holes is lower than that of electrons, the transit time of electrons is smaller. If $\mu_{\mathrm{p}} \ll \mu_{\mathrm{n}}$, then the gain term is dominated by the electron mobility term and in turn the transit time

$$G = \frac{\tau}{t_{\mathrm{tr,n}}}. \tag{4.184}$$

Thus, the photon-induced current from Equation (4.179) becomes

$$I_{\mathrm{ph}} = \Delta I_{\mathrm{n}} = \frac{\tau}{t_{\mathrm{tr,n}}}\frac{q\eta P}{h\nu}. \tag{4.185}$$

In Equation (4.177), the time variation of Δn may be as $e^{\mathrm{j}\omega\mathrm{t}}$, and in the small-signal case the response time varies inversely as $(1 + j\omega\tau)$, which defines the bandwidth. Thus, if the optical power is given by

$$P(\omega) = P_{\mathrm{opt}} + P_1 e^{\mathrm{j}\omega\mathrm{t}}, \tag{4.186}$$

then the ac current term is

$$\Delta I_{\mathrm{ac}} = \frac{q\eta P_1}{h\nu}\frac{\tau}{t_{\mathrm{tr,n}}}\frac{1}{1 + j\omega\tau}. \tag{4.187}$$

The rms current magnitude is given by

$$\Delta I_{\mathrm{ac,rms}} = \frac{q\eta P_1}{\sqrt{2}h\nu}\frac{\tau}{t_{\mathrm{tr,n}}}\frac{1}{(1+\omega^2\tau^2)^{1/2}}. \tag{4.188}$$

Noise in these detectors is from several sources and these are thermal or Johnson noise, generation-recombination noise, and at low frequencies, the $1/f$ flicker noise. If the resistance of the device is R_{cond}, then the thermal noise current is given by

$$\left\langle |\imath_{\mathrm{th}}|^2 \right\rangle = \frac{4k_{\mathrm{B}}T\Delta f}{R_{\mathrm{cond}}}, \tag{4.189}$$

where Δf is the bandwidth. The generation-recombination noise term is given by

$$\left\langle |\imath_{\mathrm{GR}}|^2 \right\rangle = \frac{4qGI_{\mathrm{o}}\Delta f}{1+\omega^2\tau^2}. \tag{4.190}$$

where $I_{\mathrm{o}} = q\eta P_{\mathrm{opt}}G/\omega\tau$. Neglecting the contribution of 1/f noise at low frequencies, the signal-to-noise ratio is given by

$$\frac{S}{N} = \frac{\Delta I^2_{\mathrm{ac,rms}}}{\left\langle |\imath_{\mathrm{th}}|^2 \right\rangle + \left\langle |\imath_{\mathrm{GR}}|^2 \right\rangle} \tag{4.191}$$

$$= \frac{\eta P_1^2}{8h\nu P_{\mathrm{opt}}\Delta f}\left[1+\frac{k_{\mathrm{B}}T}{Gq}(1+\omega^2\tau^2)\frac{1}{R_{\mathrm{cond}}I_{\mathrm{o}}}\right]^{-1}. \tag{4.192}$$

The noise equivalent power (NEP) is defined as the incident rms optical power required to produce a signal-to-noise ratio of unity in a bandwidth of 1 Hz, [3], and the detectivity of a detector is the inverse of NEP or D = (NEP)$^{-1}$. The normalised detectivity D^* is defined as

$$D^* = \frac{A^{1/2}(\Delta f)^{1/2}}{NEP} \quad \left(\mathrm{cm.Hz^{1/2}\,W^{-1}}\right), \tag{4.193}$$

where A is the area of the photoconductor on which light is incident. The bandwidth is usually set to 1 Hz, and the reference area is set to 1 cm^2. Note that D^* is usually expressed as $D^*(\lambda, f, 1)$, where λ is the wavelength of light, f is its frequency of modulation, and 1 is the bandwidth in Hz and may be obtained from Equation (4.192).

The structure of these devices may take various forms, including the interdigitated surface contact structure shown in Figure 4.27.

4.3.3 P–I–N diodes

Detection of photons with energies at or above the band gap of a semiconductor requires that they are absorbed and create hole–electron pairs, and a current be induced due to this absorption. The depletion layer of a reverse-biased p–n junction of the semiconductor causes the holes and electrons to separate and be collected by the appropriate contact/collection region. The photons entering this device, schematically shown in Figure 4.28, create hole–electron pairs as the light is absorbed, from the top contact region, through the depletion layer on both sides of the junction and possibly beyond, to the lower contact region. The fields in the depletion layer are high enough to separate the

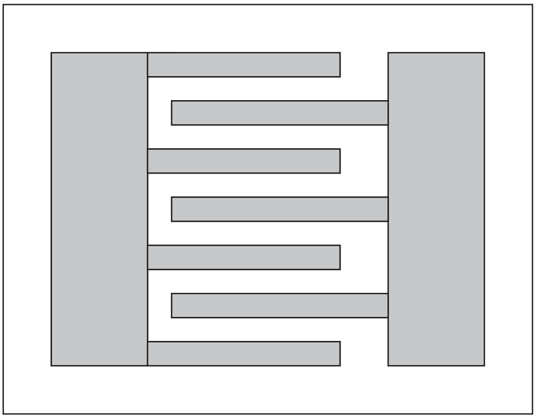


Fig. 4.27 Schematic diagram of a photoconductor detector with interdigitated fingers.

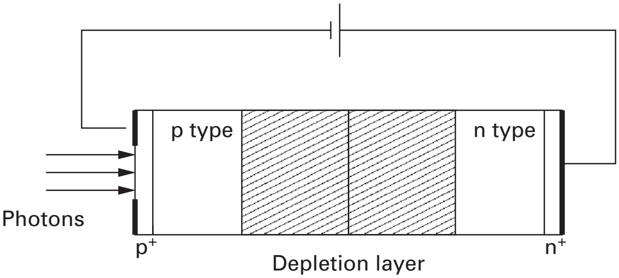


Fig. 4.28 Schematic diagram of a p–n junction photodiode.

holes and electrons, and send them to the respective majority carrier region, holes to the p region and electrons to the n region, because of the reverse bias. Holes and electrons generated in the p and n contact regions need to be considered differently. The minority carriers, electrons in the p region and holes in the n region, about one diffusion length from the depletion layer, diffuse towards the depletion layer and are accelerated to the appropriate majority carrier region. This diffusion is a slow process, which degrades the response of the diode detector.

The alternative to the simple p–n junction detector is to use a p–i–n structure, the 'i' region is either an 'i' layer or a p^- or n^- layer. In this device, the depletion layer extends through the whole of the 'i' region with no bias or with a negative bias. The usual top p^+ layer is made thin to ensure that absorption in the top contact layer is negligibly small. Most of the absorption is in the 'i' or n^- layer, with some in the lower n^+ layer, as shown in Figures 4.29 and 4.30.

Suppose the photon flux per unit area is given by ϕ_0, and the absorption coefficient is given by α, then the hole–electron generation rate is given by

$$Gen(z) = \phi_0 \alpha e^{-\alpha z}, \tag{4.194}$$

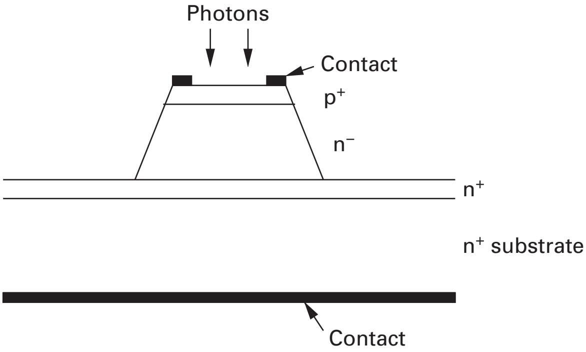


Fig. 4.29 Schematic diagram of a p–n junction photodiode in mesa form, with a ring top contact.

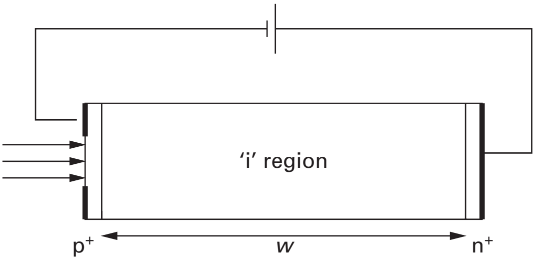


Fig. 4.30 Schematic diagram of a p–i–n junction photodiode.

and the photon flux density ϕ_0 is given by

$$\phi_0 = \frac{P_{\text{inc}}(1-R)}{Ah\nu} \tag{4.195}$$

where the incident optical power is given by P_{inc}, R is the reflectivity of the surface and A is the device area. Note that the extra α introduced in the generation term is to normalise the current density. The drift current density is given by

$$J_{\text{drift}} = -q\int_0^w Gen(z)dz = q\phi_0(1-e^{-\alpha\text{w}}). \tag{4.196}$$

The tail end of the optical power also enters the lower n$^+$ region, and generates hole–electron pairs. In this region, the hole, minority carrier, motion is determined by the diffusion equation

$$D_\text{p}\frac{\partial^2 p_\text{n}}{\partial z^2} - \frac{p_\text{n}-p_{\text{n}0}}{\tau_\text{p}} - Gen(z) = 0, \tag{4.197}$$

where D_p is the diffusion coefficient for holes, τ_p is the lifetime for holes beyond $p_{\text{n}0}$, the equilibrium hole density for the n^+ doped layer. The boundary conditions are $p_\text{n} = p_{\text{n}0}$ for $z = \infty$ and $p_\text{n} = 0$ for $z = w$; the former has some validity since the structure is generally grown on an n^+ substrate. The latter boundary condition has the excess hole density to be zero since those generated at the boundary are acted upon by the depletion layer and accelerated away. The solution is given by

$$p_{\mathrm{n}} = p_{\mathrm{n0}} - (p_{\mathrm{n0}} + Ce^{-\alpha \mathrm{w}})e^{(\mathrm{w}-\mathrm{z})/\mathrm{L_p}} + Ce^{-\alpha \mathrm{z}}, \tag{4.198}$$

where $L_{\mathrm{p}} = \sqrt{D_{\mathrm{p}}\tau_{\mathrm{p}}}$ and

$$C = \frac{\phi_0 \alpha L_{\mathrm{p}}^2}{D_{\mathrm{p}}\left(1 - \alpha^2 L_{\mathrm{p}}^2\right)}. \tag{4.199}$$

The diffusion current density is

$$J_{\mathrm{diff}} = -qD_{\mathrm{p}}\left(\frac{\partial p_{\mathrm{n}}}{\partial z}\right) \tag{4.200}$$

$$= q\phi_0 \frac{\alpha L_{\mathrm{p}}}{1 + \alpha L_{\mathrm{p}}} e^{-\alpha \mathrm{w}} + qp_{\mathrm{n0}}\frac{D_{\mathrm{p}}}{L_{\mathrm{p}}}, \tag{4.201}$$

and the total current density is the sum of the drift current density and the diffusion current density, and is given by

$$J_{\mathrm{tot}} = q\phi_0\left(1 - \frac{e^{-\alpha \mathrm{w}}}{1 + \alpha L_{\mathrm{p}}}\right) + qp_{\mathrm{n0}}\frac{D_{\mathrm{p}}}{L_{\mathrm{p}}}. \tag{4.202}$$

The value of p_{n0} is small in the n^+ region, and therefore under illumination, this last term is small and is usually omitted, and the current is proportional to the photon flux.

The quantum efficiency may be obtained from these expressions:

$$\eta_{\mathrm{ext}} = \frac{J_{\mathrm{tot}}A/q}{P_{\mathrm{opt}}/h\nu} = (1 - R_{\mathrm{r}})\left(1 - \frac{e^{-\alpha \mathrm{w}}}{1 + \alpha L_{\mathrm{p}}}\right), \tag{4.203}$$

where P_{opt} is the optical power absorbed and R_{r} is the reflectivity. To make the quantum efficiency high, the diode reflectivity needs to be made small using an anti-reflection coating, so that $R_{\mathrm{r}} \approx 0$. Also $\alpha w \gg 1$ makes η_{ext} closer to unity but at the expense of transit time delay, and hence the frequency response becomes small.

The frequency response of the photodiode is governed by several factors, and these are discussed below. The diffusion of the minority carriers generated outside the depletion region and their transit to the appropriate electrode region are major limitations to the response time of the diode. However, careful design may minimise this effect, including the use of heterostructures so that the p^+ and the n^+ regions are in higher bandgap material, which prevents the generation of hole–electron pairs in these contact regions. The RC time constant of the diode, where C is its capacitance and R is shunt resistance added to extract the signal, determines the circuit response, and is minimised by suitable choice of R, and the area of the diode. The carrier transit time across the depletion region is also a major factor, and this is analysed below.

Suppose the incident optical flux is modulated to have a time varying component of the form $\phi_1 e^{j\omega_{\mathrm{m}}t}$. At any point z in the depletion layer, the carriers generated have to drift to the appropriate electrode region at the saturation velocity v_{s}. Thus, the conduction current density due to the carriers generated at z has a phase delay of $e^{-\mathrm{j}\omega_{\mathrm{s}}\mathrm{z}/\mathrm{v_s}}$ and is given by

$$J_{\mathrm{cond}}(z) = q\phi_1 e^{\mathrm{j}\omega_{\mathrm{m}}\mathrm{t}} e^{-\mathrm{j}\omega_{\mathrm{m}}\mathrm{z}/\mathrm{v_s}}. \tag{4.204}$$

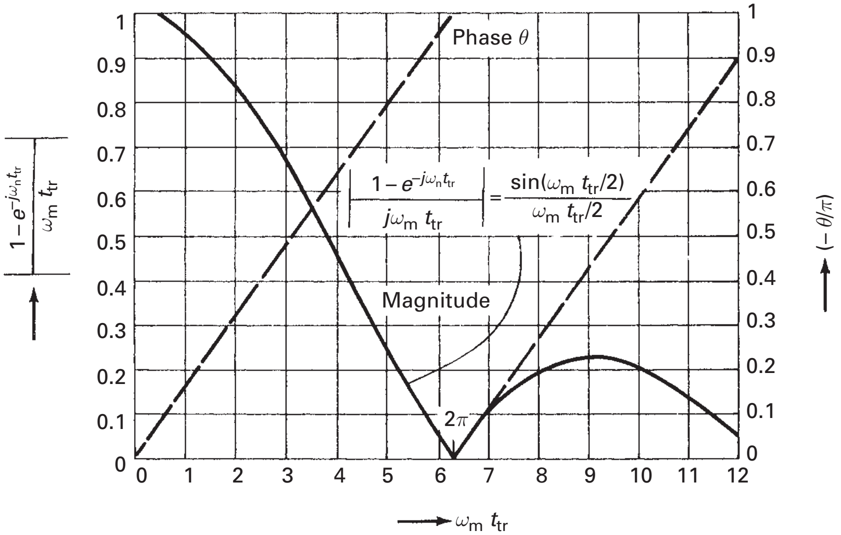


Fig. 4.31 Plot of the transit time factor in Equation (4.206) against the product of the transit time and the modulation frequency (A. Yariv, *Optical Electronics in Modern Communications*, Fifth Edition, Oxford University Press 1997. Figure 4.11–4.17. By permission of Oxford University Press, Inc.).

Considering the carriers generated over the entire depletion layer, neglecting the absorption coefficient, the total conduction current density is given by

$$J_{\text{cond}} = \frac{1}{w}\int_0^w q\phi_1 e^{\mathrm{j}\omega_{\mathrm{m}}\mathrm{t}} e^{-\mathrm{j}\omega \mathrm{z}/\mathrm{v_s}} dz \tag{4.205}$$

$$= q\phi_1\left(\frac{1-e^{-\mathrm{j}\omega_{\mathrm{m}}\mathrm{t_{tr}}}}{j\omega t_{\mathrm{tr}}}\right)e^{\mathrm{j}\omega_{\mathrm{m}}\mathrm{t}}, \tag{4.206}$$

where t_{tr} is the transit time equal to w/v_{s}. In addition to the conduction current density, the displacement current density is given by

$$J_{\text{disp}} = j\omega_{\mathrm{m}}\epsilon\mathcal{E} = \frac{j\omega_{\mathrm{m}}\epsilon V}{w}. \tag{4.207}$$

The total current density is the sum of these two terms. The conduction current density is reduced by the transit time factor as shown in Equation (4.206). Figure 4.31 shows a plot of this factor $(1 - e^{-\mathrm{j}\omega_{\mathrm{m}}\mathrm{t_{tr}}})/(\omega_{\mathrm{m}} t_{\mathrm{tr}})$ against $\omega_{\mathrm{m}} t_{\mathrm{tr}}$. Note that when the denominator becomes 2π, this factor goes to zero.

The expression for the transit time effect in Equation (4.206) becomes $1/\sqrt{2}$ when the value of $\omega_{\mathrm{m}} t_{\mathrm{tr}}$ is 2.4, and also has a phase shift of $2\pi/5$. Thus, the 3 dB response is given by

$$f_{3\,\mathrm{dB}} = \frac{2.4}{2\pi t_{\mathrm{tr}}} = \frac{0.382 v_{\mathrm{s}}}{w} = 0.382 v_{\mathrm{s}}\alpha, \qquad \text{for } \alpha w = 1. \tag{4.208}$$

Note that including the effect of the absorption coefficient in the above derivations, Equations (4.204)–(4.208) in fact reduces the transit time factor further, and therefore these results are optimistic.

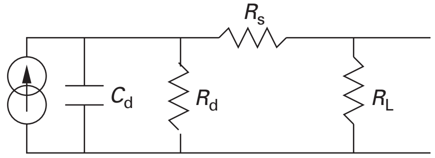


Fig. 4.32 Equivalent circuit of a p–i–n photodiode.

The equivalent circuit of the p–i–n photodiode is shown in Figure 4.32, and consists of a current generator, the diode incremental resistance R_d in parallel, the diode capacitance C also in parallel, the series resistance R_s, which includes the contact resistance and the load resistance R_L also in parallel. The inductance of the wire bond or contact line in series with R_s has been omitted. Note that R_L is generally small compared to R_d.

The analysis of this section shows that to make frequency response of the photodiodes high, the capacitance of the diode should be low so that the R_LC time constant is small and the 'i' layer thickness should be between $1/\alpha$ and $2/\alpha$, where α is the absorption coefficient, and this comes at the expense of responsivity. Using higher bandgap materials for the p^+ and n^+ regions allows higher fields to be used as the breakdown voltage is increased. The usual choice is to make R_LC time constant equal to the transit time across the depletion region. To reduce the capacitance, the area is made small, of the order of 50 μm^2 or less. High-speed p–i–n diodes have been designed and built for many years, and currently diodes with responses in the 50 GHz region are available [10]. To reduce the transit time effects, the absorbing layer is made as thin as possible, of the order of 0.15 µm, and this is at the expense of responsivity. To overcome the reduced absorption in the thin 'i' layer, the use of a dielectric mirror above the contact layer and an epitaxial mirror below the absorbing layer causes the light to have multiple passes, and improves the responsivity [10]. The problem of having this in the n^+ layer requires careful design of the doping of the heterostructure mirror, since current has to flow through it. Note that the lower mirror is similar to those used in VCSELS and there the mirror series resistance problem has been alleviated considerably.

Heterostructure diodes in which the p^+ and n^+ regions are of a larger bandgap material, with the absorption 'i' or n^- layer, have also been designed. The advantage of this structure is that the incoming light is not absorbed in the heavily doped regions, and the depletion fields may be higher. However, the heterojunctions may have to be graded to prevent carrier trapping.

The noise output of the p–i–n diode is dominated by the thermal or Johnson noise. Then the input signal photocurrent is given by

$$I_{ph} = \frac{q\eta_{int}P_{in}}{h\nu} \equiv RP_{in}, \tag{4.209}$$

where R is the diode responsivity in A W^{-1}. The shot noise is due to the background optical power generating a current given by I_B, the dark current I_d due to the reverse-biased p–i–n diode and the photocurrent I_{ph}, and is given by

$$\left\langle |\sigma_s|^2 \right\rangle = 2q(I_{ph} + I_B + I_d)\Delta f. \tag{4.210}$$

The diode is usually connected to a preamplifier, noise figure F_n, input resistance R_a, which is in parallel with the diode output resistance R_d and the load resistance across the diode R_L, and neglecting the series resistance R_s which is small, the equivalent resistance given by

$$\frac{1}{R_\mathrm{eq}} = \left(\frac{1}{R_\mathrm{d}} + \frac{1}{R_\mathrm{L}} + \frac{1}{R_\mathrm{a}}\right) \approx \frac{1}{R_\mathrm{L}}, \tag{4.211}$$

and the corresponding Johnson or thermal noise current squared, including the preamplifier noise figure, is given by

$$\left\langle |\sigma_\mathrm{t}|^2 \right\rangle = \frac{4k_\mathrm{b} T F_\mathrm{n} \Delta f}{R_\mathrm{eq}}. \tag{4.212}$$

Thus, the signal-to-noise ratio is given by

$$\frac{S}{N} = \frac{\left(\dfrac{q\eta_\mathrm{int} P_\mathrm{in}}{h\nu}\right)^2}{2q(I_\mathrm{ph} + I_\mathrm{B} + I_\mathrm{d})\Delta f + \dfrac{4k_\mathrm{b} T F_\mathrm{n} \Delta f}{R_\mathrm{eq}}}. \tag{4.213}$$

Usually, the shot noise current is small, R_eq is approximately equal to R_L, and the signal-to-noise ratio becomes

$$\frac{S}{N} = \frac{\left(\dfrac{q\eta_\mathrm{int} P_\mathrm{in}}{h\nu}\right)^2}{\dfrac{4k_\mathrm{b} T \Delta f}{R_\mathrm{L}}} \tag{4.214}$$

Thus, to make the signal-to-noise ratio large, R_L needs to be made as large as possible, but the problem here is that the RC time constant then becomes large and the response time becomes slow. Thus, the values of $\mathrm{R_L}$ are of the order of 50–100 Ω to ensure that the RC time constant approaches the transit time limit.

4.3.4 Avalanche photodiodes

The p–i–n diode has no gain, and adding gain enhances its performance. Increasing the reverse bias close to or at the breakdown of the diode, the electric field in the 'i' or n^- region becomes high, and results in the carriers being accelerated to a higher velocity before a collision with the lattice occurs. If the velocity is high, then this collision may be inelastic and causes ionisation, resulting in an additional electron and hole being generated. This additional electron and hole together with the original electron are also accelerated in turn to have further collisions to create additional hole–electron pairs. A schematic diagram of this process is shown in Figures 4.33 and 4.34 in which an electron is injected, at the start of the high field region of the depletion layer. This figure assumes that the ionisation coefficient for the electrons α_e is equal to that for the holes α_h. These coefficients are probabilistic, and are the reciprocal of the average distance that the carrier travels before an ionising collision occurs. These coefficients are a function of the electric field, and vary with the material parameters. In Silicon,

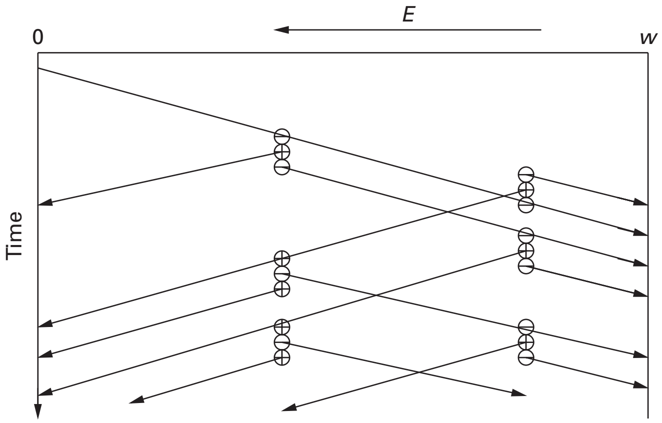


Fig. 4.33 Schematic diagram of the avalanche process with injection of an electron into the avalanche region, which after collision creates an additional hole–electron pair, which, together with the original electron, is also accelerated to have further ionising collisions, resulting in additional hole–electron pairs for each carrier. In this diagram, the ionisation coefficients of the electrons and holes are assumed to be nearly equal.

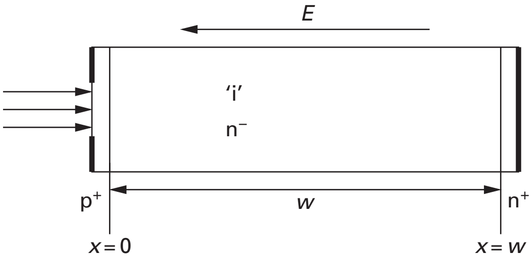


Fig. 4.34 Schematic diagram of the avalanche diode.

$\alpha_e > \alpha_h$; but in III–V material, GaAs and InP and others they are almost equal. The collisional increase in carriers results in an increase in the current which implies current gain.

The expression for the ionisation coefficient is given by [3]:

$$\alpha_{e,h} = \alpha_\infty \exp\left[-\left(\frac{b}{\mathcal{E}}\right)^{m}\right]. \tag{4.215}$$

In GaAs, α_∞ is $\sim 1.3 \times 10^6\ \mathrm{cm}^{-1}$, b is $2 \times 10^6\ \mathrm{V\,cm}^{-1}$ and m is 2. The multiplication coefficient or factor M is the ratio of the output current density to the input current density, and may be subdivided for the electron component and the hole component.

Suppose that the total current density in the diode is given by

$$J_{\text{tot}} = J_e(x) + J_h(x), \tag{4.216}$$

where $J_{\mathrm{e}}(x)$ is the electron current density, $J_{\mathrm{h}}(x)$ is the hole current density and J_{tot} is a constant at any plane of the diode. The multiplication factors are defined as

$$M_{\mathrm{e}} = \frac{J_{\mathrm{e,out}}}{J_{\mathrm{e,in}}} = \frac{J_{\mathrm{e}}(w)}{J_{\mathrm{e}}(0)} \tag{4.217}$$

for the electron current density, and

$$M_{\mathrm{h}} = \frac{J_{\mathrm{h,out}}}{J_{\mathrm{h,in}}} = \frac{J_{\mathrm{h}}(0)}{J_{\mathrm{h}}(w)} \tag{4.218}$$

for the hole current density.

Suppose a p–i–n diode with an 'i' region is considered, the field across the 'i' region is uniform. Also assume that the ionisation coefficients $\alpha_{\mathrm{e}} = \alpha_{\mathrm{h}}$, and are constant at the fixed electric field in the 'i' region, and the width of the avalanching region is w. The holes and electrons are accelerated in opposite directions, and the collisions result in ionisation and multiplication. Assume that at $x = 0$, there is only electron injection, hole injection is zero at $x = w$, and therefore the avalanching is initiated by electrons. In this case, the multiplication factor is given by

$$M = 1 + \alpha_{\mathrm{e}}w + (\alpha_{\mathrm{e}}w)^2 + (\alpha_{\mathrm{e}}w)^3 + (\alpha_{\mathrm{e}}w)^4 + \cdots \tag{4.219}$$

$$= \frac{1}{1 - \alpha_{\mathrm{e}}w} \text{ for } \alpha_{\mathrm{e}}w < 1. \tag{4.220}$$

This final result implies that the multiplication factor goes to ∞ when $\alpha_{\mathrm{e}}w$ is unity, or that each carrier has one ionising collision over the distance w.

For the case of the p–i–n, with an n$^-$ region instead of the 'i' region, the field increases linearly from the n$^+$ region to the p$^+$ region, and the above assumptions of constant ionisation coefficients no longer hold. The total current density flowing in the diode is the sum of the electron current density and the hole current density.

The current densities in the avalanching region satisfy the following equations:

$$\frac{dJ_{\mathrm{e}}}{dx} = \alpha_{\mathrm{e}}J_{\mathrm{e}} + \alpha_{\mathrm{h}}J_{\mathrm{h}} + qGen(x) \tag{4.221}$$

$$-\frac{dJ_{\mathrm{h}}}{dx} = \alpha_{\mathrm{e}}J_{\mathrm{e}} + \alpha_{\mathrm{h}}J_{\mathrm{h}} + qGen(x), \tag{4.222}$$

where $qGen(x)$ is the optical generation rate. Subtracting these two equations results in

$$\frac{d}{dx}[J_{\mathrm{e}}(x) + J_{\mathrm{h}}(x)] = 0, \tag{4.223}$$

or

$$J_{\mathrm{e}}(x) + J_{\mathrm{h}}(x) = \text{constant} = J_{\mathrm{tot}}. \tag{4.224}$$

This implies that the current density is a constant at any plane in the diode, as claimed earlier in Equation (4.216). Substituting for J_{h} from Equation (4.216) by $J_{\mathrm{tot}} - J_{\mathrm{e}}$ in Equation (4.221)

$$\frac{dJ_{\mathrm{e}}}{dx} - (\alpha_{\mathrm{e}} - \alpha_{\mathrm{h}})J_{\mathrm{e}} = \alpha_{\mathrm{h}}J_{\mathrm{tot}} + qGen(x). \tag{4.225}$$

Consider the case when $\alpha_e = \alpha_h$, neglecting the *Gen* term, this becomes

$$\frac{dJ_e}{dx} = \alpha_e J_{tot}, \tag{4.226}$$

and the solution is

$$J_e(w) = J_{tot} \int_0^w \alpha_e dx + J_e(0). \tag{4.227}$$

Assuming that there is no hole injection at $x = w$, then $J_e(w) = J_{tot}$, and this equation becomes

$$J_e(w) = J_e(w) \int_0^w \alpha_e dx + J_e(0). \tag{4.228}$$

Dividing by $J_e(0)$, this equation becomes

$$\frac{J_e(w)}{J_e(0)} = M_e = M_e \int_0^w \alpha_e dx + 1, \tag{4.229}$$

or

$$M_e = \frac{1}{1 - \int_0^w \alpha_e dx}. \tag{4.230}$$

For avalanching,

$$\int_0^w \alpha_e dx = 1. \tag{4.231}$$

When $\alpha_e \neq \alpha_h$, then Equation (4.225) needs to be solved. This is first-order differential equation of the form:

$$\frac{dy}{dx} + P(x)y = Q(x), \tag{4.232}$$

and the solution is given by

$$y = \frac{\int_0^x Q(x')e^{\int_0^{x'} P(x'')dx''} + y(0)}{e^{\int_0^x P(x')dx'}}. \tag{4.233}$$

Consider the case of electron injection at $x = 0$, with $Gen(x) = 0$ and no hole injection at $x = w$, which implies that $J_h(w) = 0$, and $J_{tot} = Je(w)$. Following the above procedure, with these boundary conditions, the solution of Equation (4.225) is obtained, and the electron current multiplication factor becomes [8]:

$$M_e = \frac{J_e(w)}{J_e(0)} = \frac{e^{\int_0^w (\alpha_e - \alpha_h)dx'}}{1 - \int_0^w dx' \alpha_h(x') e^{\int_{x'}^w (\alpha_e - \alpha_h)dx''}}. \tag{4.234}$$

When $\alpha_e = \alpha_h$, then this equation becomes, as shown in Equation (4.230)

$$M_e = \frac{1}{1 - \int_0^w \alpha_e dx}. \tag{4.235}$$

When the denominator of these multiplication factors is zero, the diode avalanches. A similar expression to that in Equation (4.234) may be derived for M_h, if the avalanche is initiated by holes. For this case, it is assumed that at $x = w$, $J_e(w) = 0$, $Gen(x) = 0$ and $J_h(0) = J_{tot}$, and the variable is $J_h(x)$ to give

$$M_h = \frac{J_h(w)}{J_e(0)} = \frac{e^{-\int_0^w (\alpha_e - \alpha_h) dx'}}{1 - \int_0^w dx' \alpha_e(x') e^{-\int_0^{x'} (\alpha_e - \alpha_h) dx''}}, \qquad (4.236)$$

and again the avalanching occurs when the denominator goes to zero. The avalanching is dependent on the field in the depletion layer, carrier population and the collision frequency, and not dependent on the carrier initiating the avalanche process. By careful control of the bias, the avalanche gain may be controlled to values as desired.

Notice that the generation term has been omitted in these expressions, but may be included as necessary. Silicon avalanche photodiodes (APDs) may use the absorption region as the avalanche region. In this case, the derivations above with the generation term would be the appropriate equations. Near infrared wavelengths have used GaAs/AlGaAs APDs, and for wavelengths in the telecommunications band, 1300 nm and 1550 nm bands, the absorption layer is generally InGaAs which is lattice-matched to InP. The avalanche region for these diodes is separate and usually InP, since large leakage currents occur due to tunneling with high reverse bias in InGaAs. This type of diode is termed the *separate absorption and multiplication* (SAM) APDs [26], and is currently the usual APDs at these wavelengths, as shown schematically in Figure 4.35. Accumulation of holes occurs in the InP–InGaAs valence band junction region, and a graded junction alleviates this problem [22].

The multiplication factor is a random variable and therefore the excess noise due to this avalanche process may be estimated by a noise factor F_M. This factor is a function of the multiplication factor M and k_A, where k_A is the ratio of the ionisation coefficients and k_A lies in the range $0 < k_A < 1$. Thus, $k_A = \alpha_h/\alpha_e$, provided $\alpha_h < \alpha_e$, or alternatively $k_A = \alpha_e/\alpha_h$, provided $\alpha_e < \alpha_h$. The expression for the noise factor is

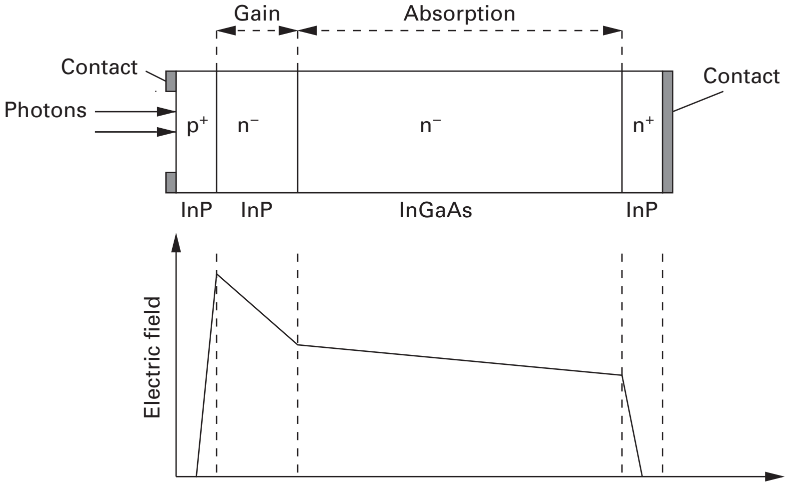


Fig. 4.35 Schematic diagram of the separate absorption and multiplication avalanche photodiode, also showing the field in the different regions.

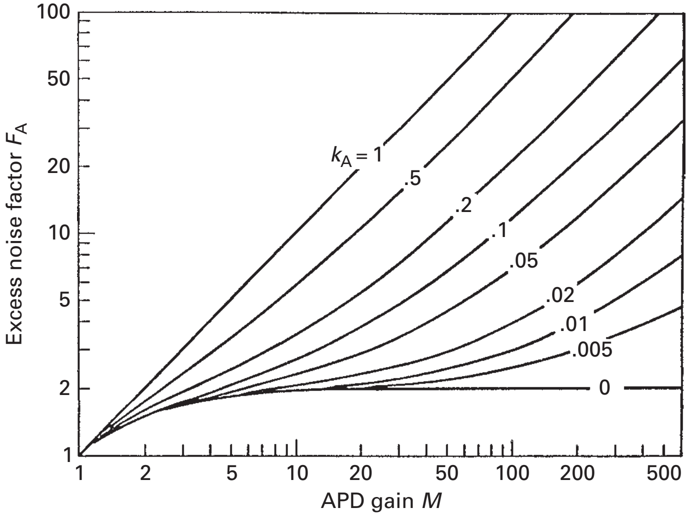


Fig. 4.36 Excess noise factor F_A as a function of the multiplication factor M and the ratio of the ionisation coefficients k_A (G. P. Agrawal, *Fiber-Optic Communication Systems*, 3rd edition, John Wiley & Sons, 2002. ©2002 John Wiley & Sons). Reprinted with permission of John Wiley & Sons, Inc.

given by [23]:

$$F_A = k_A M + (1 - k_A)\left(2 - \frac{1}{M}\right). \tag{4.237}$$

Figure 4.36 shows the excess noise factor F_A plotted against the multiplication factor M, as a function of k_A, from the Equation (4.237) [1].

The responsivity of the avalanche photodiode includes the multiplication factor M and is given by

$$R_{APD} = M\frac{\eta q}{h\nu}, \tag{4.238}$$

and therefore the input signal photocurrent is given by

$$I_{ph,APD} = R_{APD}P_{opt} = M\frac{\eta q}{h\nu}P_{opt}. \tag{4.239}$$

The shot noise current squared of the avalanche diode is also enhanced by the multiplication factor and the noise factor and is given by

$$\left\langle |\sigma_s|^2 \right\rangle = M^2 F_A (R_{APD}P_{in} + I_D + I_B)\Delta f, \tag{4.240}$$

where I_D is the dark current and I_B the background current. The thermal noise current squared is given by, as in Equation (4.212), including the amplifier noise figure F_n,

$$\left\langle |\sigma_t|^2 \right\rangle = \frac{4k_b T F_n \Delta f}{R_{eq}}. \tag{4.241}$$

Hence the signal-to-noise ratio is given by

$$\frac{S}{N} = \frac{\left(\dfrac{Mq\eta_{\text{int}}P_{\text{in}}}{h\nu}\right)^2}{2qM^2F_{\text{A}}(I_{\text{ph}} + I_{\text{B}} + I_{\text{d}})\Delta f + \dfrac{4k_{\text{b}}TF_{\text{n}}\Delta f}{R_{\text{eq}}}}. \tag{4.242}$$

In this case, the shot noise term dominates and therefore the signal-to-noise ratio becomes

$$\frac{S}{N} = \frac{\left(\dfrac{Mq\eta_{\text{int}}P_{\text{in}}}{h\nu}\right)^2}{2qM^2F_{\text{A}}(I_{\text{ph}} + I_{\text{B}} + I_{\text{d}})\Delta f}. \tag{4.243}$$

The APD is the preferred photodiode, but the introduction of the gain stage reduces the frequency response. However, recent results have shown that these devices have gain bandwidth products of over 300 GHz, with a response of 28 GHz [18]. Lower gain results show a higher response of over 30 GHz [29]. These devices are much more expensive than p–i–n diodes, although the performance is better because of the gain that arises due to avalanching.

4.3.5 Metal–semiconductor–metal detectors

A planar version of the p–i–n diode is the metal–semiconductor–metal (MSM) detector, in which the contacts are metal Schottky barrier diodes to a thin undoped semiconductor layer, and the region between the metal contacts, usually in the form of an interdigitated structure, shown in Figure 4.37, is completely depleted. The gap between the fingers is made small, and the transit time limitation is small. The capacitance has two components: one in parallel with the gap across the fingers, and the second is the capacitance to the ground of both electrodes. However, this latter may be made small so that the RC time constant of the diode is small. Consequently, these detectors are extremely fast,

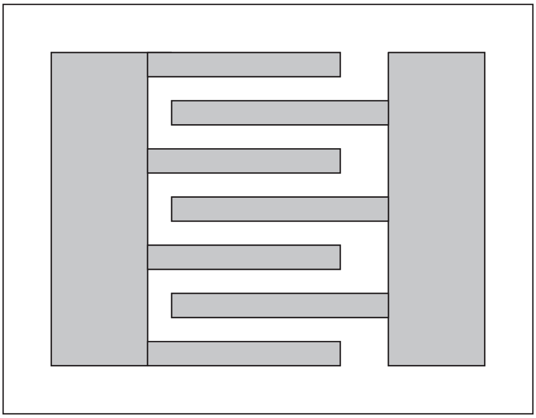


Fig. 4.37 Interdigitated form of the MSM detector, in which the contacts are Schottky diodes to the thin epitaxial absorption layer.

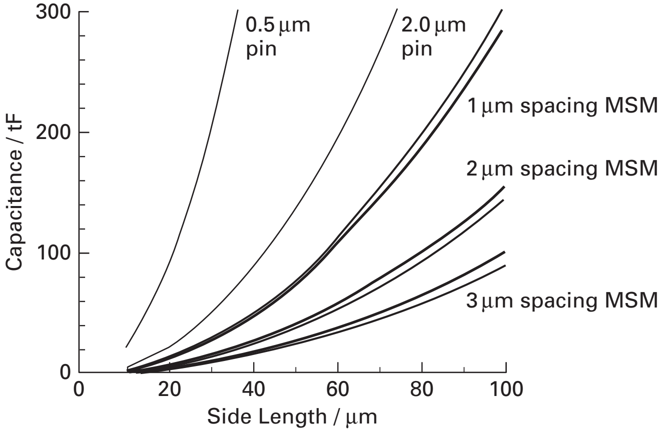


Fig. 4.38 Capacitance against the edge length of the fingers of the interdigitated MSM device for gaps of 1, 2 and 3 μm spacing, with finger widths of 0.5 and 1.0 μm. The parallel plate capacitance of the p–i–n diode is plotted for comparison, with the 'i' or n$^-$ layer thickness of 0.5 and 1 μm. The electrode width for the p–i–n diode has not been given in the paper (J. B. D. Soole and H. Schumacher, *IEEE Journal of Quantum Electronics*, Vol. 27, No. 3, March 1991. ©1991 IEEE).

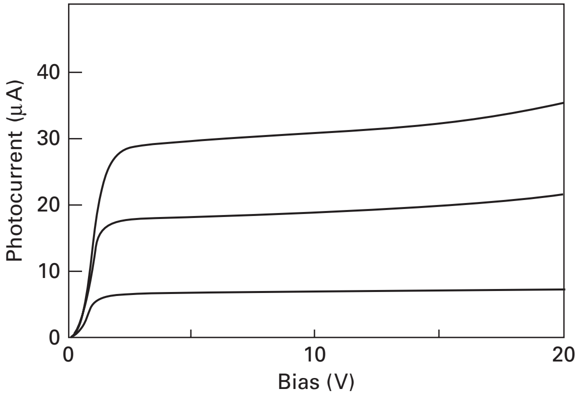


Fig. 4.39 Response of the MSM detector showing the detected photocurrent against the bias voltage for different values of light intensity for the 1 μm electrode width and a 2 μm interelectrode gap device (J. B. D. Soole and H. Schumacher, *IEEE Journal of Quantum Electronics*, Vol. 27, No. 3, March 1991. ©1991 IEEE).

but their responsivity is affected by the shadowing effect of the fingers. However, it was found that the quantum efficiency increases with increasing bias, which suggests that there is some gain that arises due to trap densities at the interfaces between layers, at electrode interfaces. However, [30] suggests that with careful growth this gain may be much reduced as seen in Figure 4.39. Note that one of the diodes is forward-biased and the other is reverse-biased, and again careful processing results in symmetric response and soft breakdown.

MSM photodetectors of InGaAs on an InP substrate [30] for operation at 1.3 μm and 1.5 μm wavelengths are discussed here. The Schottky barrier on InGaAs is of the

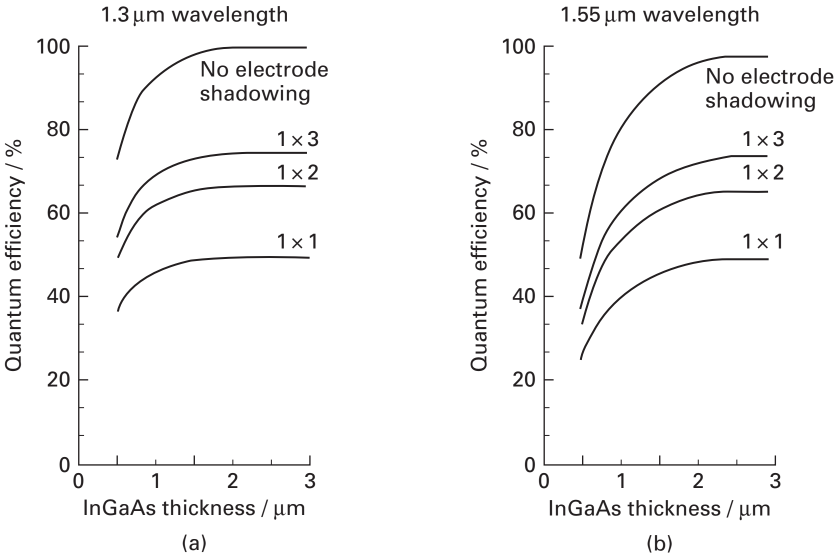


Fig. 4.40 Quantum efficiency of the InGaAs MSM device the electrode widths of 1 μ and different spacings from 1, 2, 3 μm inter-electrode spacing. The zero width electrode result is also plotted here (J. B. D. Soole and H. Schumacher, *IEEE Journal of Quantum Electronics*, Vol. 27, No. 3, March 1991. ©1991 IEEE).

order of 0.2 V, and is leaky in reverse bias, and therefore is enhanced by a layer of lattice-matched InAlAs under the electrodes. The thickness of the InGaAs layer was 1.3 μm, and the InAlAs layer was 80 nm. The capacitance between the fingers has been calculated by means of the usual Schwarz–Christofel transformation in this paper, and it is shown that the edge capacitance of the fingers is much less than the capacitance of comparable area mesa-type p–i–n diodes, as shown in Figure 4.38. Provided the series resistance is comparable, the RC time constant is much less for the MSM detector.

The device with a 1 μm wide electrode with the interelectrode gap of 2 μm between them has a response shown in Figure 4.39 of the detected photocurrent against the bias voltage across the electrodes for different values of light intensity on the device. Note that the response is not flat but rises, indicating that there is gain in the device; this may be in part due to the photoconductor effect and also may be at the higher bias due to avalanching near the electrodes. The quantum efficiency of the MSM detectors is hampered by the shadowing effect of the electrodes and this is plotted for this InGaAs device in Figure 4.40, which shows that this may rise to as high as about 70% for the 1 μm electrode width and a 2 μm interelectrode gap device. The bandwidth of the device is in the 20 GHz region, and other devices on GaAs show similar or better frequency response.

4.3.6 Travelling wave p–i–n photodiodes

As discussed above, the frequency response limitation of p–i–n diodes arises from the transit time of photogenerated carriers reaching their respective electrode contact regions, and also from the RC time constant of the device. Making the absorption layer small, 0.3 μm, results in poor responsivity; placing mirrors underneath to obtain a double pass improves the response, but the RC time constant rises. The MSM device reduces the RC time constant but the transit time limitation remains. The second limitation is the device saturation that occurs when the light intensity becomes high. The alternative is the waveguide photodetector (WPD), which is an edge-fed p–i–n diode that is made very narrow and long, and the absorption layer may be thin to overcome the transit time limitation, making it highly efficient. However, the RC time constant limits the device response to about 55 GHz bandwidth–efficiency value (product of bandwidth and quantum efficiency) [16].

To overcome the problem of the RC limitation of the waveguide detector, and also be able to handle high power detection, the travelling wave detector was proposed by [8, 32]. An optical waveguide with a thin absorption layer is the basis for this device. The absorption layer is made thin enough to only absorb a small fraction of the light per unit length so as not to saturate the device, thus high power signals may be detected without saturation. However, the limitation of the travelling wave detector arises from the velocity mismatch between the electrical wave on the electrode structure and the optical waveguide. The electrode on top of the guide with the accompanying ground electrodes on the sides form a coplanar waveguide, in which the detected signal travels in the form of a voltage/current wave, see Figure 4.41. The optical signal travels at the guide layer group velocity, usually at c/n_g, and n_g is the group index of the guide layer. The detected signal forms electrical forward and backward waves on the electrode structure. The backward wave reflects at the input of the electrode structure, if it is an open circuit at the start of the detector, or alternatively, the wave is absorbed in the load if matched. Thus, the electrical wave travelling on the electrodes is a combination of the forward wave and the reflected component of the backward wave travelling in the forward direction, and also the backward wave traveling in the reverse direction. The lack of velocity match between these two optical and electrical waves leads to walk off, and the frequency response drops [15]. However, bandwidth of 172 GHz and bandwidth–efficiency product of 76 GHz [12] and pulse response with transform bandwidth as high as 560 GHz [28] have been reported. Careful choice of the absorbing layer, absorbing coefficient of α_0, and its thickness and location in the waveguide

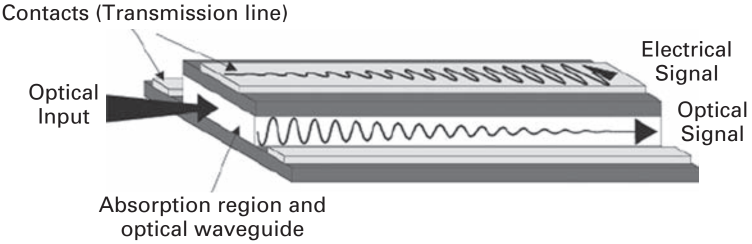

Contacts (Transmission line)
Electrical Signal
Optical Input
Optical Signal
Absorption region and optical waveguide

Fig. 4.41 Schematic diagram of a travelling wave photodiode, with coplanar electrodes (G. Rengel-Sharp, R. E. Miles, S. Iezekiel, *The Radio Science Bulletin*, Vol. 311, No. 12, pp. 55–64, December 2004. ©2004 URSI).

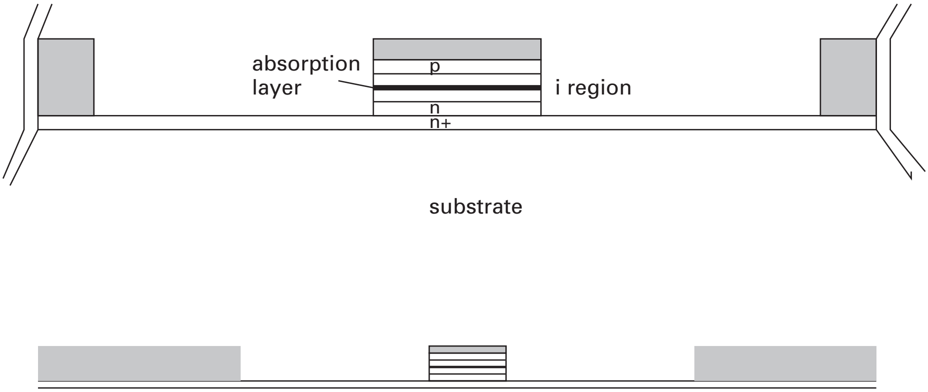


Fig. 4.42 Schematic diagram of a travelling wave photodiode, showing the p–i–n structure with the absorbing layer, and the electrodes.

allow the confinement factor Γ to be determined, and thus the decay factor becomes $e^{-\alpha_0 \Gamma z}$ for the optical wave intensity. These types of travelling wave devices are the fully distributed detectors. A second design has used a passive guide between MSM detectors [20], in which each detector only detects a small fraction of the power carried by the optical wave. Velocity matching allows the 3dB frequency response to rise to 49 GHz. Both these types of travelling wave devices are the fastest photodetectors built at the present time. The usual optical waveguides are of the ridge type, with the top electrode and adjacent ground electrodes of the coplanar waveguide (CPW) type, as shown in Figure 4.41 [27].

The theory of the distributed detector has been developed by [12] and [15]. Essentially, the p–i–n diode has to support the optical mode, and the electrode structure supports the electrical mode. Figure 4.42 shows a typical structure of this travelling wave photodetector.

The optical mode travels in the central p–i–n region, and the electrical mode is supported by the electrode structure transmission line shown in this figure. The equivalent circuit of the elemental section of the line is now the usual R–L in series and C–G in parallel, but now with the elemental p–i–n structure in parallel; the modified circuit is shown in Figure 4.43. Note that the usual capacitance is that of the depletion layer, and therefore is across the current source.

In this circuit, R_{cpw} is the series resistance of the central electrode and varies due to the skin effect, increasing as $\sqrt{f}$. The semiconductor series resistance R_{semi} is in series with the depletion layer, and is determined by the doping levels of the p^+ and n^+ layers, shown as p and n in Figure 4.42. The capacitance of the central electrode is dominated by the depletion layer capacitance of the p–i–n diode, which is much larger than the edge capacitance. The depletion layer capacitance gives rise to a slow wave effect, similar to that suggested by Hasegawa [14]. The inductance is the usual CPW value, and is unaffected by the presence of the p–i–n structure. The series resistance G_s is the top contact layer in parallel with the electrodes. The distributed current source is

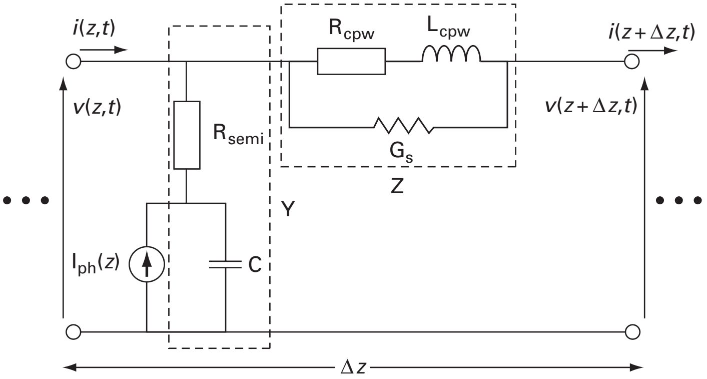


Fig. 4.43 Schematic diagram of the equivalent circuit of a travelling wave photodiode (G. Rengel-Sharp, R. E. Miles and S. Iezekiel, *The Radio Science Bulletin*, Vol. 311, No. 12, pp. 55–64, December 2004. ©2004 URSI).

related to the input optical power P_o, and is given by

$$I_{\text{ph}}(z) = P_o \frac{\eta q \lambda}{hc} \Gamma \alpha_o e^{-\Gamma \alpha_o} e^{-j\beta_o z}, \tag{4.244}$$

where λ is the wavelength of the optical power, α_o is the absorption layer coefficient, Γ is the confinement factor in this absorption layer, β_o is the optical propagation constant and η is the quantum efficiency of the distributed p–i–n diode. The terminations at the input may be an open circuit or a matched load, and at the output is matched to extract the signals on the electrical wave. This equivalent circuit may be used to analyse the performance of the travelling wave detector.

The major concern in the treatment of the travelling wave detector is the velocity mismatch between the optical wave and the electrical wave on the electrode structure. Giboney *et al.* [12] have analysed the transmission line using an impulse excitation, to account for the velocity mismatch. Performing the Fourier transform, the following expression for the normalised photocurrent is obtained for the case when $\Gamma\alpha_o\ell \ll 1$ for the frequency ω:

$$\imath_\nu(\omega) = \frac{1}{2}\left[\frac{\omega_f}{\omega_f - j\omega} + \gamma(\omega)\frac{\omega_r}{\omega_r + j\omega}\right] e^{-j\omega(\ell/v_e)}, \tag{4.245}$$

where

$$\omega_f = \frac{\Gamma\alpha_o v_e}{\left(\dfrac{1 - v_e}{v_o}\right)}$$

and

$$\omega_r = \frac{\Gamma\alpha_o v_e}{\left(\dfrac{1 + v_e}{v_o}\right)}.$$

γ is the electrical reflection coefficient at the device input, $\gamma = 1$ for an open circuit and $\gamma = 0$ for matched load. The term v_e is the electrical wave velocity and v_o is the

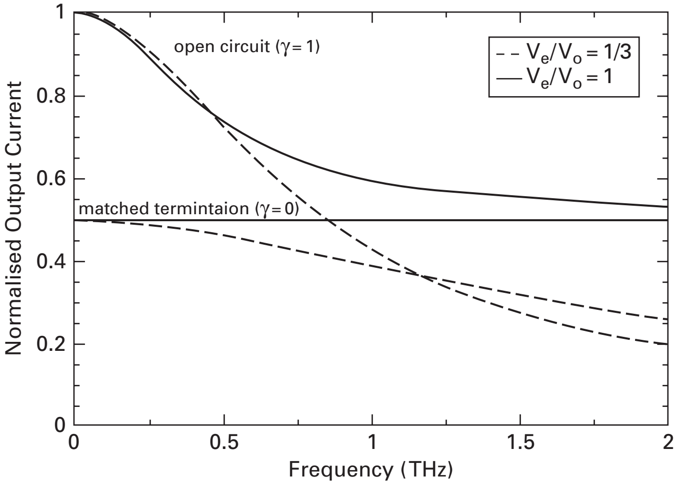


Fig. 4.44 Normalised current response for both the velocity-matched and -mismatched cases for both the open circuit and the matched terminations at the input (G. Rengel-Sharp, R. E. Miles and S. Iezekiel, *The Radio Science Bulletin*, Vol. 311, No. 12, pp. 55–64, December 2004. ©2004 URSI).

optical wave velocity. The results of this calculation are shown in Figure 4.44. Notice that the normalised current for the matched case has no frequency variation, but since half of it goes to the input-matched load, the output current is reduced by half. For the velocity-mismatched case, for $\gamma = 0$, again the frequency variation is lower than for the case when the input has an open circuit, when $\gamma = 1$. The open circuit case has considerable frequency variation, but the current is higher at all frequencies for the velocity-matched case, and becomes asymptotic to the $\gamma = 0$ case. For the velocity-mismatched case, the current falls off faster at the higher frequencies, and becomes asymptotic to the velocity-mismatched $\gamma = 0$ case.

For the velocity mismatched case with matched input termination, $\gamma = 0$, when $\ell\Gamma\alpha_{\mathrm{o}} \gg 0$, the 3 dB bandwidth is given by [27]

$$f_{3\mathrm{dB}} = \frac{\Gamma\alpha_{\mathrm{o}}}{2\pi}\frac{v_{\mathrm{o}}v_{\mathrm{e}}}{(v_{\mathrm{o}} - v_{\mathrm{e}})}, \tag{4.246}$$

and for the case when open circuit input termination, $\gamma = 1$, the bandwidth is

$$f_{3\mathrm{dB}} \approx \frac{\Gamma\alpha_{\mathrm{o}}v_{\mathrm{e}}}{3\pi}. \tag{4.247}$$

This last result holds for velocity mismatch in the range $0 \leq v_{\mathrm{e}}/v_{\mathrm{o}} \leq 1.47$ [13]. Additional losses occur due to the optical guide scattering loss and microwave transmission line loss, and carrier transit time effects, and these will change the results and hence the performance of the photodetector. Since the optical waveguide is in the heterostructure form, carrier trapping may be a problem, but graded interface junctions would alleviate this.

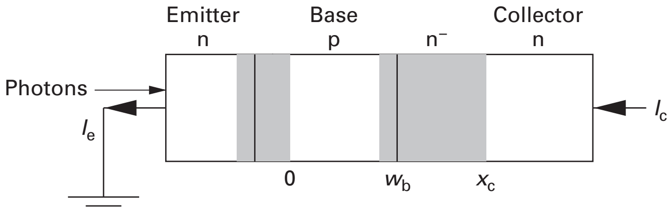


Fig. 4.45 Schematic diagram of a photo-HBT showing the depletion layers.

4.3.7 Heterostructure bipolar transistor photodetector

The bipolar transistor acts as excellent photodetector with gain. In the HBT, Figure 4.45, the absorption of light occurs solely in the base, and possibly in the subcollector region. In a double HBT in which the emitter and collector are of larger bandgap material, the absorption is confined to the base region, even though the base is not fully depleted. The usual modus operandi is to leave the base open circuit, as the light-generated carriers constitute the base current, though in some case to avoid carrier trapping contact is made to the base. The transit time of the electrons in the base of an n–p–n transistor determines the frequency limit of operation of the HBT, in addition to the base resistance effects. For a reasonable fraction of the incident to be absorbed, the base and the base–subcollector regions should be of the order of 1 μm thick, which makes these devices extremely slow, typically about 1 GHz bandwidth.

According to [3], the optical gain G of the single heterojunction phototransistor is the ratio of the number of carriers in the collector current due to the photogenerated–base and subcollector carriers to the number of photons incident on the base-subcollector:

$$G = \frac{h\nu I_{\text{ph-coll}}}{q P_{\text{inc}}}. \tag{4.248}$$

For a thick subcollector depletion width, which is greater than $1/\alpha$, then

$$G = \eta \beta_T, \tag{4.249}$$

where η is the quantum efficiency and β_T is optical current gain of the photo-transistor.

4.4 Problems

(1) A GaAs LED has a structure shown in Figure 4.1, and injection takes place from the n$^+$ layer into the p layer where the recombination takes place. Suppose the lifetime of the electrons in this layer is 10 ns, the electron mobility is 3000 cm^2 (V.s)$^{-1}$ and hole mobility is 300 cm^2 (V.s)$^{-1}$, determine the thickness of the p layer if it is to be one diffusion length.

(2) Calculate, numerically, the correct value of the external quantum efficiency of a LED radiating into air. Assume that the semiconductor index is 3.5, the polarisation random which implies that half the radiation is perpendicular polarisation and the other half is parallel polarisation. Assume that the transmission coefficient for the perpendicular polarisation from medium 1, index n_1, into medium 2, index n_2, is given by

$$\tau_\perp = \frac{2n_1 \cos\theta_1}{(n_1 \cos\theta_1 + n_2 \cos\theta_2)},$$

and the transmission coefficient for parallel polarisation is given by

$$\tau_\| = \frac{2n_1 \cos\theta_2}{(n_2 \cos\theta_1 + n_1 \cos\theta_2)}.$$

(3) A ridge laser is to operate at 800 nm, with a guide index of 3.41 and cladding index of 3.43. Determine the maximum thickness and width of the laser waveguide.

(4) An InGaAsP laser operating at 1300 nm has an effective guide index of 3.4. What is the reflectivity of its cleaved facets? Suppose the internal loss is $30\,\text{cm}^{-1}$, what is the photon lifetime, assuming the group index is 3.5? Suppose the laser length is $300\,\mu$m, width $3\,\mu$m and thickness of $0.2\,\mu$m, the gain coefficient a is $2 \times 10^{-16}\,\text{cm}^2$, the value of the transparency number density is $10^{18}\,\text{cm}^{-3}$ and carrier lifetime is 1 ns. What is the threshold carrier density and the threshold current? What is the relaxation oscillation frequency at twice the threshold current?

(5) A p–i–n diode receiver at 1300 nm has a preamplifier noise figure of 2 dB. The signalling rate is 1 GHz, bandwidth of 100 MHz, the load resistance is $100\,\Omega$, dark current is 1 nA, the diode quantum efficiency is 0.95 and a signal of $5\,\mu$W illuminates this diode. Determine the signal-to-noise ratio at the output.

(6) Suppose the p–i–n diode is replaced by an avalanche photodetector, multiplication factor of 10, noise factor of 3, dark current of 10 nA and background current of 1 nA. This APD is illuminated with $1\,\mu$W of optical power; determine the signal-to-noise ratio.

References

[1] Agrawal G. P. (2002). *Fiber-Optic Communication Systems*, 3rd edn. John Wiley & Sons.

[2] Agrawal G. P., Dutta N. K. (1993). *Semiconductor Lasers*. Van Norstrand Rheinhold.

[3] Bhattacharya P. (1997). *Semiconductor Optoelectronic Devices*, 2nd edn. Prentice Hall.

[4] Burrus C. A., Dawson R. W. (1970). Small-area high-current-density GaAs electroluminescent diodes and a method of operation for improved degradation characteristics. *Appl. Phys. Lett. 17*, 3, 97–99.

[5] Chang K. P., Yang S. L., Chuu D. S., Hsiao R. S., Chen J. F. (2005). Characterization of self-assembled InAs quantum dots with InAlAs/InGaAs strain-reduced layers by photoluminescence spectroscopy. *J. Appl. Phys. 97*, 83511–83514.

[6] Chang Y.-C., Wang C. S., Coldren L. A. (2007). High-efficiency, high-speed VCSELs with 35 Gbit/s error-free operation. *Electron. Lett. 43*, 19, 1022–1023.

[7] Cheng D. K. (1989). *Field and Wave Electromagnetics*, 2nd edn. Addison-Wesley.

[8] Chuang S. L. (1995). *Physics of Optoelectronic Devices*. John Wiley & Sons.

[9] Dutta N. K. (1980). Calculated absorption, emission, and gain in $In_{0.72}Ga_{0.28}As_{0.6}P_{0.4}$. *J. Appl. Phys. 52*, 6095–6100.

[10] Effendberger F. J., Joshi A. M. (1996). Ultrafast, dual-depletion region, InGaAs/InP p-i-n detector. *J. Lightw. Technol. 14*, 8 (August), 1859–1864.

[11] Fathpour K. P., Mia Z., Bhattacharya P. (2005). High speed quantum dot lasers. *J. Phys. D: Appl. Phys. 38*, 2103–2111.

[12] Giboney K. S., Nagarajan R. L., Reynolds T. E., *et al.* (1995). Traveling-wave photodetectors with 172-GHz bandwidth and 76-GHGz bandwidth-efficiency product. *IEEE Photonic Technol. Lett. 7,* 4 (April), 412–414.

[13] Giboney K. S., Rodwell M. J. W., Bowers J. E. (1997). Traveling-wave photodetector theory. *IEEE Trans. Microw. Theory Tech. 45*, 8 (August), 1310–1319.

[14] Hasegawa H., Furukawa M., Yani H. (1971). Properties of microstrip line on Si-SiO_2 system. *IEEE Trans. Microw. Theory Tech. 19*, 11 (November), 869–881.

[15] Hietala V., Vawter G. A., Brennam T. M., Hammons B. E. (1995). Traveling-wave photodetectors for high-power, large-bandwidth applications. *IEEE Trans. Microw. Theory Tech. 43*, 9 (September), 2291–2298.

[16] Kato K. (1999). Ultrawide band/high-frequency photodetectors. *IEEE Trans. Microw. Theory Tech. 47*, 7 (July), 1265–1281.

[17] Keiser G. (2000). *Optical Fiber Communications*, 2nd edn. McGraw-Hill.

[18] Kinsey G. S., Campbell J. C., Dentai A. G. (2001). Waveguide avalanche photodiode operating at 1.55 μm with a gain-bandwidth product of 320 GHz. *IEEE Photonic Technol. Lett. 13*, 8 (August), 842–844.

[19] Lee T. P., Burrus C. A., Copeland J. A., Dentai A. G., Marcuse D. (1982). Short cavity InGaAsP injection lasers: dependence of mode spectra and single-longitudinal-mode power on cavity length. *IEEE J. Quan. Electron. QE-18*, 7, 1101–1113.

[20] Lin L. Y., Wu M. C., Itoh T., *et al.* (1997). High-power high-speed photodetectors – design, analysis, and experimental demonstration. *IEEE Trans. Microw. Theory Tech. 45*, 8 (August), 1320–1331.

[21] Marcuse D., Lee T. P. (1983). On approximate analytical solutions of the rate equations for studying transient spectra of injection lasers. *IEEE J. Quan. Electron. QE-19*, 9, 1101–1113.

[22] Matsushima Y., Sasaki K., Noda Y. (1981). New type InGaAs/InP heterostructure avalanche photodiode with buffer layer. *IEEE Electron Device Lett. 2*, 7 (July), 179–181.

[23] McIntyre R. J. (1972). The distribution of gains in uniformly multiplying avalance photodiodes: theory. *IEEE Trans. Electron Devices 2*, 6 (June), 703–713.

[24] Nash F. R. (1973). Mode guidance of double-heterostructure GaAs lasers. *J. Appl. Phys. 44*, 10, 4696–4707.

[25] Nelson R. J., Wilson R. B., Wright P. D., Barnes P. A., Dutta N. K. (1981). CW electrooptical properties of InGaAsP (λ = 1.3 μm) buried-heterostructure lasers. *IEEE J. Quan. Electron. QE-17*, 2, 202–207.

[26] Nishida K., Taguchi K., Matsumo Y. (1979). InGaAsP heterostructure avalanche photodiodes with high avalanche gain. *Appl. Phys. Lett. 35*, 3 (August), 251–253.

[27] Rengel-Sharp G., Miles R. E., Iezekiel S. (2004). Traveling-wave photodetectors: a review. *The Radio Science Bulletin, URSI 311*, 12 (December), 54–64.

[28] Shi J.-W., Gan K.-G., Chiu Y.-J., *et al*. (2001). Metal-semiconductor-metal traveling-wave photodetectors. *IEEE Photonic Technol. Lett. 16*, 6 (June), 623–625.

[29] Shi J.-W., Liu Y.-H., Liu C.-W. (2004). Design and analysis of separate-absorption-transport-charge-multiplication traveling-wave avalanche photodetectors. *J. Lightw. Technol. 22*, 6 (June), 1583–1590.

[30] Soole J. B. D., Schumacher H. (1991). InGaAS metal-semiconductor-metal photodetector for long wavelength optical communications. *IEEE J. Quan. Electron. 27*, 3 (March), 737–752.

[31] Tauber D., Wang G., Geels R. S., Bowers J. E., Coldren L. A. (1993). Large and small signal dynamics of vertical cavity surface emitting lasers. *Appl. Phys. Lett. 62*, 4, 325–327.

[32] Taylor H. F., Ekonoyan O., Park C. S., Choi K. N., Chang K. (1990). Traveling-wave photodetectors. *SPIE Optoelectronics Signal Processing for Phased Array Antennas II*, 59–63.

[33] Tucker R. S., Kaminow I. P. (1984). High frequency characteristics of directly modulated InGaAsP ridge waveguide and buried heterostructure lasers. *J. Lightw. Technol. 2*, 4, 385–393.

[34] Yariv A. (1997). *Optical Electronics in Modern Communications*, 5th edn. Oxford University Press.

[35] Zhang X., Gutierrez-Aitkens A. L., Klotzkin D., Bhattacharya P., Caneau C., Bhatt R. (1996). 0.98 μm multiquantum well tunnelling injection lasers with ultra–high modulation bandwidths. *Electron. Lett. 32*, 18, 1715–1717.

Part Two

Circuits

5 Building blocks for high-speed analogue circuits

5.1 Executive summary

This chapter is about the use of electronic devices in elementary circuit blocks found in any micro- or millimetre-wave system – or in the analogue portions of fibre-optic communications systems. An introductory section describes linear two-ports on the basis of scattering parameters, discusses different gain definitions and treats important aspects of stability as well as noise in two-ports, with special emphasis on noise reduction by proper choice of generator impedance.

Following this, amplifiers, oscillators and mixers are described in sequence. In the amplifier section, small-signal parameters are used to derive fundamental properties of common topologies, from the simplest, one-transistor circuits to more complex gain cells, such as the cascode and differential amplifiers. Tuned amplifiers are covered, as well as broadband amplifier techniques, including distributed amplification. Finally, low-noise and power amplifiers are being treated, as well as non-linearities in amplifiers.

The oscillator section discusses how small-signal instability and non-linear gain compression effects combine to create stable sinusoidal oscillations. Important oscillator topologies and noise phenomena affecting the phase stability of oscillators are also covered.

Mixer circuits show how specifically designed non-linear circuits provide frequency-translating capabilities. Mixing principles are discussed first, followed by several mixer topologies using field effect and bipolar transistors.

5.2 Basic relations for two-port networks

5.2.1 Scattering parameter theory

Small signal equivalent circuits for semiconductor devices and circuits are usually represented in two-port form as shown in Figure 5.2. At low frequencies, two-port networks are represented by an impedance matrix, an admittance matrix, a hybrid matrix or a chain matrix. These matrices are described by Z, Y, *h* or ABCD parameters. Such representations are suitable at low frequencies where the parameters may be measured by placing short or open circuits at the input and output terminals of the two-port. At high (microwave) frequencies where there are travelling waves, short and open circuits

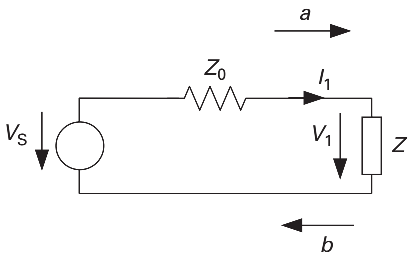


Fig. 5.1 Source and load circuit for the S-parameter discussion. *a* and *b* are normalised power waves.

cannot be precisely placed and the above-mentioned matrix representations cannot be accurately determined. The development of the vector network analyser made it possible to perform measurements of high frequency travelling wave circuits. Scattering parameters were introduced by Kurokawa [24]. He defined a set of *normalised power waves* a and b and introduced a normalising impedance Z_0:

$$a = \frac{V + Z_0 I}{2\sqrt{Z_0}} \tag{5.1}$$

$$b = \frac{V - Z_0 I}{2\sqrt{Z_0}}, \tag{5.2}$$

where V and I are the voltage and the current, respectively, at a load Z.

These normalised power waves are chosen in this way so that they relate to the power delivered to the load. Refer to Figure 5.1 where a source with a real source impedance Z_0 is connected to an arbitrary load $\underline{Z}$. The maximum power is delivered to the load if $\underline{Z} = Z_0^* = Z_0$. Then,

$$P_{Z,\max} = \frac{V_1^2}{Z_0} = \frac{V_S^2}{4Z_0} = |a|^2.$$

$|a|^2$ is hence the available power from a generator with source impedance Z_0. This also tells us that the unit of a (and b) is $\sqrt{W}$.

Now consider:

$$|a|^2 - |b|^2 = \Re\{V_1 I_1^*\}. \tag{5.3}$$

This is the power delivered to the load for arbitrary $\underline{Z}$. We can hence interpret $|a|^2$ as the power travelling towards the load, and $|b|^2$ as the power travelling from the load back to the generator, with the difference dissipated in the load. The ratio of b to a is the reflection coefficient:

$$\Gamma = \frac{b}{a} = \frac{V_1 - Z_0 I_1}{V_1 + Z_0 I_1} = \frac{\underline{Z} - Z_0}{\underline{Z} + Z_0}, \tag{5.4}$$

because $\underline{Z} = V_1/I_1$.

We now expand the normalised power wave concept to a two-port. The incident waves at each port are again designated as a and the reflected waves are designated as b, while the subscript denotes the port where the power waves are measured. For the two-port, the normalised power waves are a_1, b_1, a_2 and b_2, as shown in Figure 5.2. The

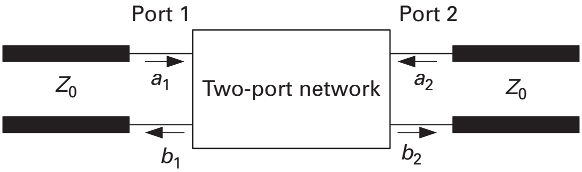


Fig. 5.2 Two-port network embedded in a transmission line with characteristic impedance Z_0, with incident waves a_1, a_2 and reflected waves b_1, b_2.

scattering parameters are the coefficients of linear equations relating the reflected waves b to the incident waves a:

$$b_1 = S_{11}a_1 + S_{12}a_2 \tag{5.5}$$

$$b_2 = S_{21}a_1 + S_{22}a_2. \tag{5.6}$$

The scattering parameters can therefore be expressed as the ratio of two power waves, provided that all the ports are terminated in a non-reflective fashion ($a = 0$ at all other ports). For a two-port,

$$S_{11} = \frac{b_1}{a_1}|_{a_2=0} \tag{5.7}$$

$$S_{12} = \frac{b_1}{a_2}|_{a_1=0} \tag{5.8}$$

$$S_{21} = \frac{b_2}{a_1}|_{a_2=0} \tag{5.9}$$

$$S_{22} = \frac{b_2}{a_2}|_{a_2=0}. \tag{5.10}$$

These relations can be written in matrix form as follows:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}. \tag{5.11}$$

Let Z_0 be the characteristic impedance of the transmission lines connected to ports 1 and 2 of the two-port. If port 2 is terminated by Z_0, there is no reflection at the load and hence the wave incident at port 2, a_2, is zero. Similarly, if port 1 is terminated by Z_0 and the stimulus is fed to port 2, a_1, is zero. If port 1 is designated to be the input and port 2 the output, then S_{11} is the input reflection coefficient with Z_0 the output, S_{22} is the output reflection coefficient with the input terminated by Z_0, S_{21} is the forward transmission coefficient with Z_0 as the output load, and S_{12} is the reverse transmission coefficient with Z_0 at the input.

In general, the scattering parameters are complex. The polar form of the scattering parameter is useful in many applications:

$$S = \mid S \mid \exp^{j\theta}, \tag{5.12}$$

where $\mid S \mid$ is the magnitude of S and θ is the phase.

Properties of scattering parameters

The following properties of the scattering parameters are important in two-port network applications. Subsequently, it is assumed that the transmission line is lossless with negligible attenuation, such that the line's complex propagation constant $\gamma = \alpha + \jmath\beta \approx \jmath\beta$.

(i) *Reciprocity*. Passive networks are reciprocal (unless they contain non-reciprocal components like isolators or circulators), and the S-parameters satisfy

$$S_{\mathrm{jk}} = S_{\mathrm{kj}}. \tag{5.13}$$

This property can be written in general form as

$$[S] = [S^{\mathrm{T}}]. \tag{5.14}$$

It states that the matrix is equal to its transpose denoted by $[S^{\mathrm{T}}]$.

(ii) *Lossless networks.* An important property of lossless networks is that the product of the transposed complex conjugate scattering matrix and the scattering matrix is equal to the unitary matrix.

$$[S][S^{\mathrm{T}}]^{*} = [I], \tag{5.15}$$

where

$$[I] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{5.16}$$

defines the unitary matrix.

(iii) *Lossy networks.* In lossy networks, the network itself dissipates power, hence

$$\sum \mid a_{\mathrm{k}} \mid^2 \; > \sum \mid b_{\mathrm{k}} \mid^2 . \tag{5.17}$$

The scattering matrix satisfies the property

$$[I] - [S][S^{\mathrm{T}}]^{*} > 0. \tag{5.18}$$

(iv) *Reference planes.* Measurements can be made at different planes along the transmission lines connected to the two-port network; this changes the results due to the signals' travelling wave nature. The reference plane is the position where the actual measurements are made. If the positions of the ports are shifted by electrical distances $\beta\ell$ away from the reference planes, the S-parameters in these shifted planes can be related to the initial S-parameters in the reference plane.

If the S-parameters were measured originally at the planes $z_1 = 0$ and $z_2 = 0$ and if the reference planes are now at $z_1 = \ell_1$ and $z_2 = \ell_2$ as in Figure 5.3, the resulting S-matrix is given by

$$\begin{bmatrix} S'_{11} & S'_{12} \\ S'_{21} & S'_{22} \end{bmatrix} = \begin{bmatrix} S_{11}\exp^{-\mathrm{j}2\theta_1} & S_{12}\exp^{-\mathrm{j}(\theta_1+\theta_2)} \\ S_{21}\exp^{-\mathrm{j}(\theta_1+\theta_2)} & S_{22}\exp^{-\mathrm{j}2\theta_2} \end{bmatrix}, \tag{5.19}$$

where

$$\theta_1 = \beta\ell_1$$
$$\theta_2 = \beta\ell_2.$$

The expression (5.19) assumes that the transmission lines are lossless.

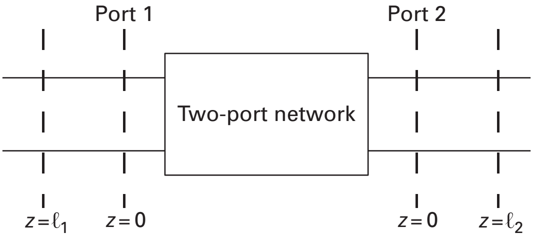


Fig. 5.3 Change of reference planes.

The scattering parameters can be converted to current–voltage parameters such as impedance ($[Z]$) parameters as well as admittance ($[Y]$) parameters. These conversions are given by Gonzalez [14].

5.2.2 The Smith chart

The analysis of two-port networks at microwave frequencies was tedious and time-consuming before speedy computation methods were available with computer-aided design software. A graphical aid to calculate various network properties such as impedances was developed by Smith [36, 37] and referred to as the *Smith chart*. The accuracy of results obtained from the Smith chart is quite adequate in most cases.

When a transmission line is terminated in an arbitrary impedance Z, there are reflections along the line, and the reflection is defined as the ratio of the voltage in the wave reflected from the terminating load to the voltage in the wave incident on the terminating load. In Equation (5.4), we had already defined the reflection coefficient Γ, which can be expressed as

$$\Gamma = \frac{Z - Z_0}{Z + Z_0}. \tag{5.20}$$

In most applications, it is convenient to use the value of the impedance normalised to the characteristic impedance or other reference impedance. The normalised value is given by

$$z = \frac{Z}{Z_0}. \tag{5.21}$$

Equation (5.20) can now be written as

$$\Gamma = \frac{z - 1}{z + 1}. \tag{5.22}$$

Both z and Γ are complex quantities and they are written in terms of their real and imaginary parts:

$$\Gamma = u + jv \tag{5.23}$$

$$z = r + jx. \tag{5.24}$$

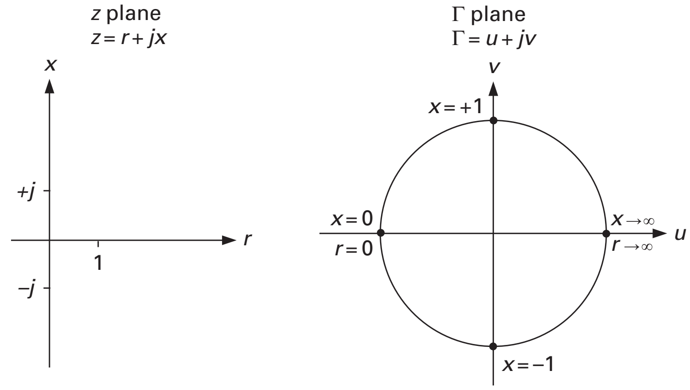


Fig. 5.4 Mapping between the z plane and the Γ plane.

It is often useful to express the reflection coefficient in polar coordinates.

$$\Gamma = \mid \Gamma \mid e^{j\theta}, \tag{5.25}$$

where $\mid \Gamma \mid$ is the magnitude, and θ is the phase of the reflection coefficient.

It follows that z and Γ are defined in two complex planes. Equation (5.22) gives the relationship between points in the two complex planes. The relationship is known as *mapping*. Equation (5.22) is a bilinear transformation where orthogonal lines in the z plane map into orthogonally intersecting circles in the Γ plane. Furthermore, it is a conformal mapping whereby the angle between the two line segments is maintained in mapping between the z and Γ planes. It is to be noted that a straight line is simply a degenerate circle. Figure 5.4 shows the correspondence between points in the z and Γ planes. Expanding Equation (5.22) we have

$$u + jv = \frac{r + jx - 1}{r + jx + 1}. \tag{5.26}$$

Equating real and imaginary parts on both sides,

$$u = \frac{r^2 - 1 + x^2}{(r+1)^2 + x^2} \tag{5.27}$$

$$v = \frac{2x}{(r+1)^2 + x^2}. \tag{5.28}$$

By eliminating x from Equations (5.27) and (5.28), we have

$$\left(u - \frac{r}{r+1}\right)^2 + v^2 = \left(\frac{1}{r+1}\right)^2. \tag{5.29}$$

By eliminating r from Equations (5.27) and (5.28), we obtain

$$(u-1)^2 + \left(v - \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2. \tag{5.30}$$

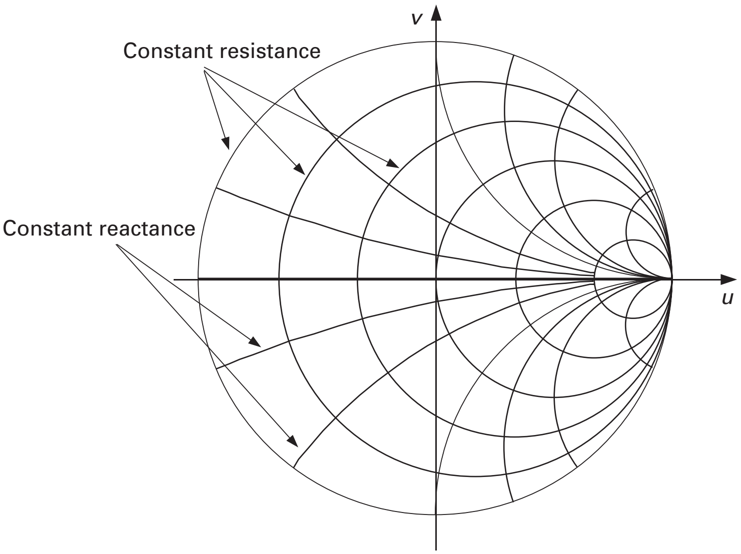


Fig. 5.5 Constant resistance and reactance circles in the Γ plane.

Equation (5.29) is the equation of a family of circles with their centres at $u = r/r + 1$, $v = 0$ and radii equal to $1/r + 1$, while Equation (5.30) is the equation of a family of circles with their centres at $u = 1$, $v = 1/x$ and radii equal to $1/x$. Equation (5.29) represents constant resistance circles; each value of $r \geq 1$ represents a circle. Equation (5.30) represents constant reactance circles which are plotted for all values of z when $\mathrm{Re}(z) \geq 0$. Both constant resistance and constant reactance circles are shown in Figure 5.5. A typical Smith chart representation for practical use is shown in Figure 5.6.

The Smith chart can also be used as an admittance chart. The constant resistance circles become the constant conductance circles and the constant reactance circles become the constant susceptance circles. The bilinear transformation in this case is

$$\Gamma = \frac{1 - y}{1 + y}. \tag{5.31}$$

The impedance and admittance representations of the Smith chart are symmetric with respect to the origin of the Smith chart. Because in typical impedance-matching problems, both impedance and admittance representations are needed, they are frequently plotted in the same diagram (see Figure 5.7).

5.2.3 Impedance matching

To maximise gain in an amplifier, its input and output impedances must be chosen to be the complex conjugate of the generator and load impedances, respectively. To achieve this, impedance-matching networks are connected at the input and the output of the amplifier, which convert the true generator and load impedances (frequently 50 Ω) to the necessary values. Figure 5.8 shows the block diagram of the complete network.

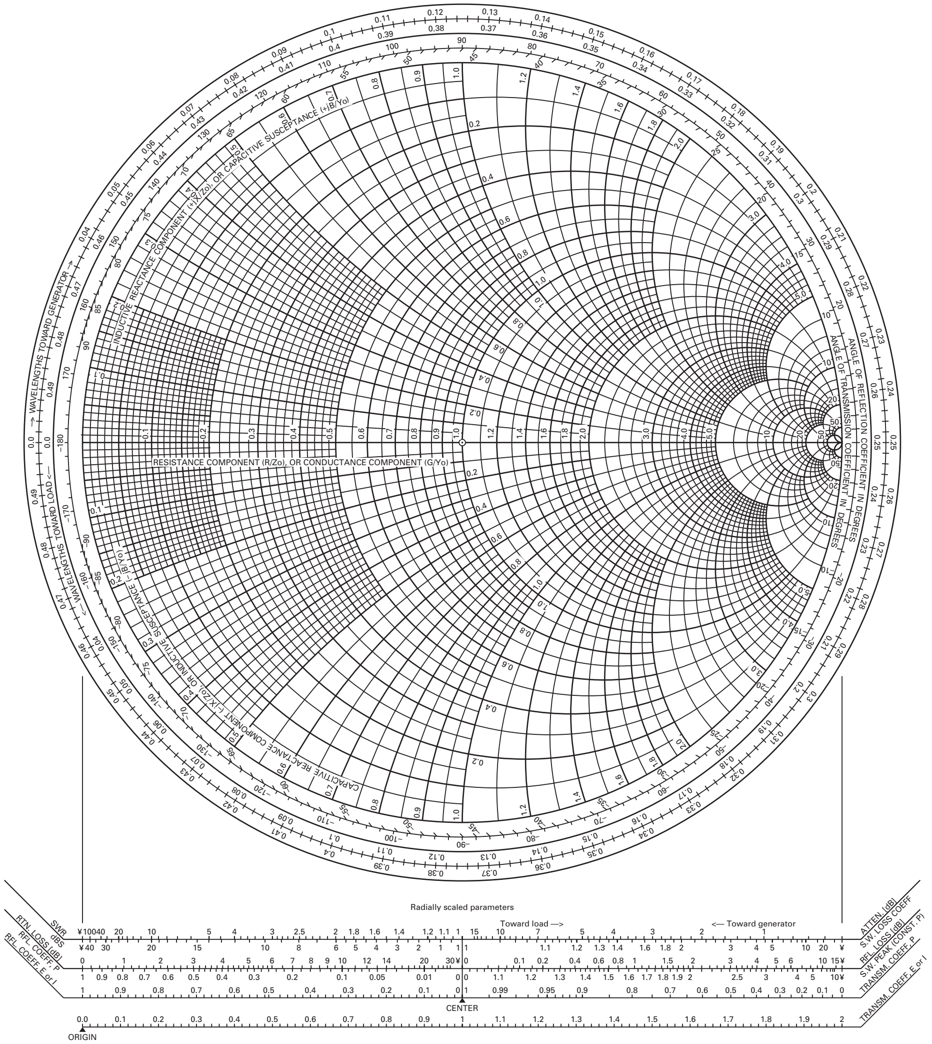

WAVELENGTHS TOWARD GENERATOR →
← WAVELENGTHS TOWARD LOAD ←
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (−jX/Zo), OR INDUCTIVE SUSCEPTANCE (−jB/Yo)
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
Radially scaled parameters
Toward load —>
<— Toward generator
SWR
dBS
RTN. LOSS [dB]
RFL. COEFF, P
RFL. COEFF, E or I
ATTEN. [dB]
S.W. LOSS COEFF
RFL. LOSS [dB]
S.W. PEAK (CONST. P)
TRANSM. COEFF, P
TRANSM. COEFF, E or I
CENTER
ORIGIN

Fig. 5.6 Practical Smith chart.

Aside from the complex conjugate match, also referred to as *power match*, impedance transformation may be necessary to achieve minimum noise figure, or maximum output power. Detailed discussions of these techniques (with diverging goals) are given later in this chapter.

5.2.4 Power gains for amplifier design

The power gain in microwave circuits is expressed in terms of the scattering parameters for convenience in performing calculations using network analyser measurements.

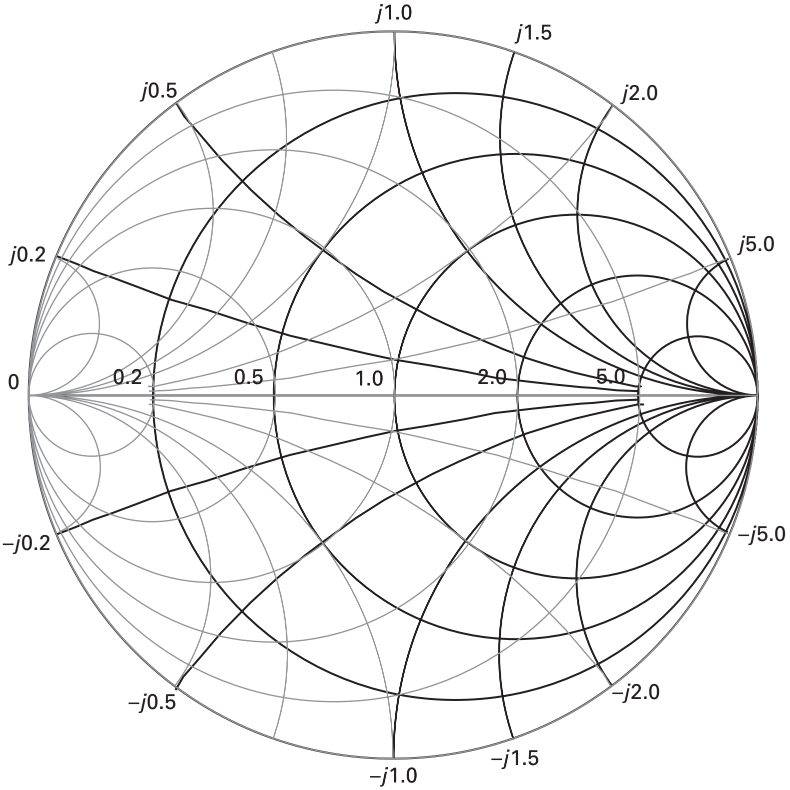


Fig. 5.7 Smith chart showing impedance circles (bold lines) and admittance circles (thin lines).

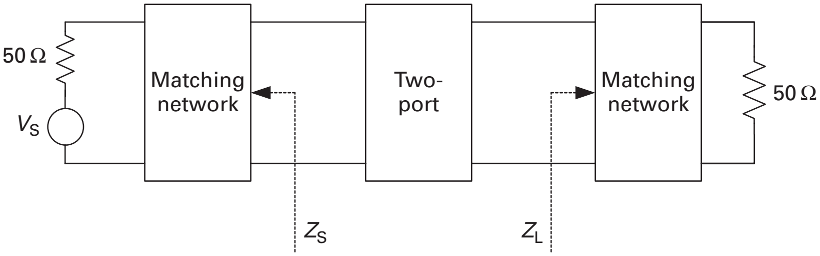


Fig. 5.8 Block diagram of a two-port embedded in impedance matching networks.

Input and output reflection coefficients for arbitrary terminations

The two-port network in Figure 5.9 is now assumed to be an amplifier circuit. We will first calculate the input and output reflection coefficients Γ_{in} and Γ_{out}, respectively, for arbitrary source and load reflection coefficients.

If Z_L is the load impedance in a transmission line system of characteristic impedance Z_0, the reflection coefficients at the source and load (Figure 5.9) are given by

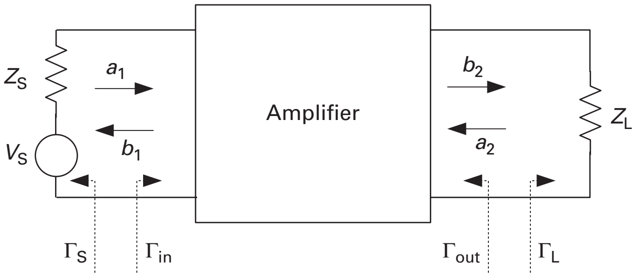


Fig. 5.9 Amplifier representation.

$$\Gamma_S = \frac{Z_S - Z_0}{Z_S + Z_0} \tag{5.32}$$

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0}. \tag{5.33}$$

Writing Equation (5.11) in expanded form, we have

$$b_1 = S_{11}a_1 + S_{12}a_2 \tag{5.34}$$

$$b_2 = S_{21}a_1 + S_{22}a_2. \tag{5.35}$$

It is evident from Figure 5.9 that reflection coefficients are related to the incident and reflected waves by the following equations:

$$\Gamma_{in} = \frac{b_1}{a_1} \tag{5.36}$$

$$a_2 = \Gamma_L b_2. \tag{5.37}$$

By substitution into Equation (5.11), it can be shown that

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}. \tag{5.38}$$

The output reflection coefficient is defined as

$$\Gamma_{out} = \frac{b_2}{a_2} |_{V_S=0} \ . \tag{5.39}$$

By substitution in Equation (5.11), it can be also shown that

$$\Gamma_{out} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S}. \tag{5.40}$$

Powers at input and load

When designing amplifiers, different definitions of power gain are applied depending on the application. It is necessary to define the power at the input and the output of the amplifier in order to obtain the gain.

The power delivered to the input port of the amplifier is written in terms of the incident and reflected waves at the input. The reflected wave can be expressed in terms of the incident wave and the reflection coefficient, see Equation (5.36).

$$P_{\text{in}} = \mid a_1 \mid^2 - \mid b_1 \mid^2 \tag{5.41}$$

$$= \mid a_1 \mid^2 \left(1 - \mid \Gamma_{\text{in}} \mid^2\right). \tag{5.42}$$

Similarly, the power delivered to the load Z_L is

$$P_L = \mid b_2 \mid^2 - \mid a_2 \mid^2 \tag{5.43}$$

$$= \mid b_2 \mid^2 \left(1 - \mid \Gamma_L \mid^2\right). \tag{5.44}$$

The power travelling towards the load is partly due to the wave originated by the generator electromotive force b_S, and partly by the reflection from the source impedance Γ_S:

$$a_1 = b_S + \Gamma_S b_1 \tag{5.45}$$

with

$$b_S = \frac{V_S\sqrt{Z_0}}{Z_S + Z_0} \tag{5.46}$$

$$\Gamma_S = \frac{Z_S - Z_0}{Z_S + Z_0}. \tag{5.47}$$

Detailed analyses are given by Gonzalez [14].

The power available from the source is labelled P_{avs} and is defined as

$$P_{\text{avs}} = \frac{\mid b_S \mid^2}{1 - \mid \Gamma_S \mid^2} \tag{5.48}$$

$$= |a_1|^2 \frac{|1 - \Gamma_S \Gamma_{\text{in}}|^2}{1 - |\Gamma_S|^2}.$$

Note that P_{avs} is the input power under conjugate match, i.e. when $\Gamma_{\text{in}} = \Gamma_S^{\star}$.

The power available from the network P_{avn} is defined as the power delivered by the network when the load is conjugately matched to the output impedance, or

$$P_{\text{avn}} = \frac{|S_{21}|^2 |b_S|^2}{|1 - S_{11}\Gamma_S|^2 (1 - |\Gamma_{\text{out}}|^2)}. \tag{5.49}$$

Power gain definitions

Let us now discuss the power gain definitions which are important in amplifier design. The operating power gain G_P is the ratio of the power delivered to the load to the power delivered to the input of the amplifier:

$$G_P = \frac{P_L}{P_{\text{in}}}$$

$$= \left|\frac{b_2}{a_1}\right|^2 \frac{1 - \mid \Gamma_L \mid^2}{1 - \mid \Gamma_{\text{in}} \mid^2}.$$

Since

$$\frac{b_2}{a_1} = \frac{S_{21}}{1 - \Gamma_L S_{22}},$$

and using Equation (5.38), we obtain

$$G_P = |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - \Gamma_L S_{22}|^2 - |S_{11} - \Gamma_L \Delta(S)|^2}, \tag{5.50}$$

where $\Delta(S)$ is the determinant of the scattering matrix.

The transducer power gain G_T is the ratio of the power delivered to the load to the power available from the source:

$$\begin{aligned} G_T &= \frac{P_L}{P_{avs}} \\ &= \left|\frac{b_2}{a_1}\right|^2 \frac{(1 - |\Gamma_L|^2)(1 - |\Gamma_S|^2)}{|1 - \Gamma_{in}\Gamma_S|^2} \\ &= |S_{21}|^2 \frac{(1 - |\Gamma_L|^2)(1 - |\Gamma_S|^2)}{|1 - \Gamma_L S_{22} - \Gamma_S\,(S_{11} - \Gamma_L \Delta(S))\,|^2}. \end{aligned} \tag{5.51}$$

In the absence of deliberate feedback, the reverse transmission S_{12} is frequently very small and can be neglected. $S_{12} = 0$ simplifies the denominator in Equation (5.51) and we obtain the unilateral transducer power gain

$$G_{TU} = |S_{21}|^2 \frac{(1 - |\Gamma_L|^2)(1 - |\Gamma_S|^2)}{|(1 - S_{11}\Gamma_L)(1 - S_{22}\Gamma_S)|^2}. \tag{5.52}$$

The available power gain G_A, finally, is the ratio of the available gain from the network P_{avn} to the available power from the source P_{avs}:

$$\begin{aligned} G_A &= \frac{P_{avn}}{P_{avs}} \\ &= |S_{21}|^2 \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2 (1 - |\Gamma_{out}|^2)} \\ &= |S_{21}|^2 \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2 - |S_{22}(1 - S_{11}\Gamma_S) + S_{12}S_{21}\Gamma_S|^2}, \end{aligned} \tag{5.53}$$

using Equation (5.40).

5.2.5 Stability

G_A still contains Γ_S as a variable. We know already that the maximum power transfer from the source to the load occurs for $\Gamma_S = \Gamma_{in}^{\star}$, so this appears to be an optimum choice. Before we proceed, however, let us again investigate Equation (5.53). If the denominator becomes zero, the available gain would grow beyond all bounds. This happens if

$$|1 - S_{11}\Gamma_S| = |S_{22}(1 - S_{11}\Gamma_S) + S_{12}S_{21}\Gamma_S|,$$

or, written differently,

$$\left| S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \right| = 1 = |\Gamma_{out}|,$$

using Equation (5.40). Likewise, we can show from the available gain in the reverse direction that it would grow beyond all bounds for $|\Gamma_{\text{in}}| = 1$.

Both these conditions are considered as the *instability* of the amplifier, a potentially dangerous situation which may lead to malfunction or even fatal failure. In most cases (the notable exception are oscillators), it needs to be avoided.

Unconditional stability

From the above, we can deduce that a two-port will be *unconditionally stable* if

$$\left| S_{11} + \frac{S_{12}S_{21}\Gamma_{\text{L}}}{1 - S_{22}\Gamma_{\text{L}}} \right| < 1, \tag{5.54}$$

for all $|\Gamma_{\text{L}}| \leq 1$, and

$$\left| S_{22} + \frac{S_{12}S_{21}\Gamma_{\text{G}}}{1 - S_{11}\Gamma_{\text{G}}} \right| < 1, \tag{5.55}$$

for all $|\Gamma_{\text{G}}| \leq 1$.

Stability circles

The following discussion follows Hoffmann [20]. For conditionally stable two-ports – where at least one of the conditions in Equations (5.54) and (5.55) is violated – we can still find generator and load admittances which allow stable operation. If we plot the locus of

$$\left| S_{22} + \frac{S_{12}S_{21}\Gamma_{\text{G}}}{1 - S_{11}\Gamma_{\text{G}}} \right| = 1$$

in the complex Γ_{G} plane, we obtain a circle with centre vector

$$\Gamma_{\text{G,C}} = \frac{\Delta(S)^* S_{22} - S_{11}^*}{|\Delta(S)|^2 - |S_{11}|^2}. \tag{5.56}$$

The radius is

$$r_{\text{G}} = \frac{|S_{12}\, S_{21}|}{\left| |\Delta(S)|^2 - |S_{11}|^2 \right|}. \tag{5.57}$$

Likewise, plotting the locus of

$$\left| S_{11} + \frac{S_{12}S_{21}\Gamma_{\text{L}}}{1 - S_{22}\Gamma_{\text{L}}} \right| = 1$$

in the complex Γ_{L} plane, we obtain a circle with centre vector

$$\Gamma_{\text{L,C}} = \frac{\Delta(S)^* S_{11} - S_{22}^*}{|\Delta(S)|^2 - |S_{22}|^2}, \tag{5.58}$$

and radius

$$r_{\text{G}} = \frac{|S_{12}\, S_{21}|}{\left| |\Delta(S)|^2 - |S_{22}|^2 \right|}. \tag{5.59}$$

Examples for stability circles in the generator and load planes are shown in Figure 5.10. The stability circles provide the boundaries between the stable and the unstable regions; however, we still need to determine whether the inside or the outside of the circle is stable. Let us do this for the Γ_{G} plane first.

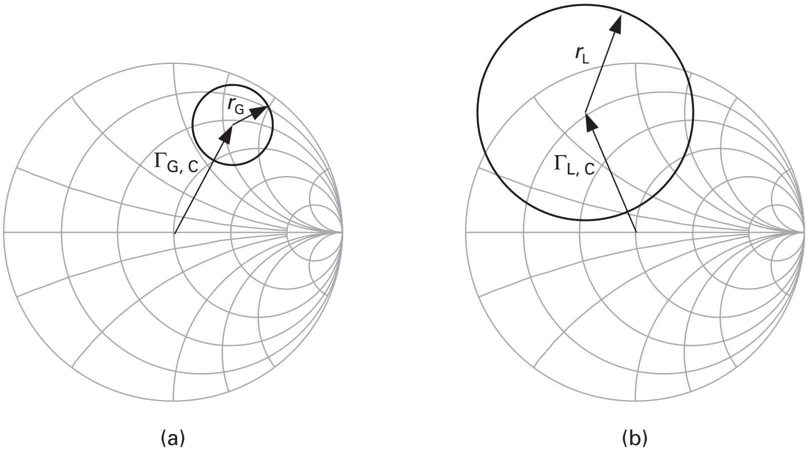


Fig. 5.10 Stability circle examples: (a) in the Γ_G plane and (b) in the Γ_L plane.

Using Equation (5.40), we conclude that $\Gamma_{out}(\Gamma_G = 0) = S_{22}$. Now we locate the area which contains $\Gamma_G = 0$ (the centre of the Smith chart). We can say:

- If $|S_{22}| < 1$, then the region which contains $\Gamma_G = 0$ is the stable region.
- Otherwise, if $|S_{22}| > 1$, the region containing $\Gamma_G = 0$ is the unstable region.

The same procedure applies for finding the stable region in the Γ_L plane.

Using the stability circles, we can look at stability in a different way. A two-port will be unconditionally stable, if all of the following conditions are fulfilled:

(i) $|S_{11}| < 1$
(ii) $|S_{22}| < 1$
(iii) $|\Gamma_{L,C}| > 1 + r_L$
(iv) $|\Gamma_{G,C}| > 1 + r_G$.

In this case, both stability circles are fully outside of the Smith chart unity circles in the reflection coefficient plane (which contain all $|\Gamma_G|$, $|\Gamma_L| < 1$), and the outer regions of the circles are the stable ones.

Rollet's stability factor

Conditions (iii) and (iv) above can be combined into the following:

$$k > 1 + \max\left[0, \frac{|\Delta(S)|^2 - 1}{2|S_{12}||S_{21}|}\right], \tag{5.60}$$

where

$$k = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta(S)|^2}{2|S_{12}||S_{21}|}. \tag{5.61}$$

k is the *Rollet factor*. $k > 1$ is often used as a stability criterion; however, note that it is a necessary, but not sufficient requirement for unconditional stability.

A very common, and potentially fatal, mistake is to assess stability only for the intended frequency of operation. It must be investigated over the full frequency range where instability may conceivably occur – parasitic oscillations at very low frequencies are extremely common, and instabilities may also increase with increasing frequencies such as in cascode amplifiers (see p. 333).

5.2.6 Maximum available gain and maximum stable gain

Let us now reassess the case of an amplifier with simultaneous complex conjugate match at the input and output ports, i.e. $\Gamma_\mathrm{S} = \Gamma^\star_\mathrm{in}, \Gamma_\mathrm{L} = \Gamma^\star_\mathrm{out}$. The available gain in this case is the *maximum available gain* and can be written using the Rollet factor Equation (5.61) as

$$MAG = \left|\frac{S_{21}}{S_{12}}\right| \left(k - \sqrt{k^2 - 1}\right). \tag{5.62}$$

Obviously, a real solution exists only if $k \geq 1$.

For $k < 1$, the *maximum stable gain* is frequently quoted. This is the MAG in the limit of $k = 1$, which can be obtained by the resistive loading of an otherwise not conditionally stable amplifier:

$$MSG = \left|\frac{S_{21}}{S_{12}}\right|. \tag{5.63}$$

5.2.7 Mason's unilateral gain

The concept of unilateral gain for a two-port network was first introduced by Mason in his paper [28], which is now considered a classic. A comprehensive review of the paper and its relevance today is given by Gupta [15].

The unilateral gain is defined as the maximum power gain obtained by a two-port when it is made unilateral – unilateralised. A two-port network that includes an active device is made unilateral by a lossless and reciprocal four-port network connected to input and output of the two-port under investigation. This network provides the necessary feedback to impose the unilateral condition. Mason's unilateral gain is not to be confused with the unilateral transducer power gain G_TU, Equation (5.52), which had been derived by *neglecting* the reverse transmission. Here, reverse transmission is eliminated by a unilateralising network.

The unilateral gain is a figure of merit which is intrinsic to the device and hence independent of the circuit in which the device is placed. The unilateral gain is, therefore, an invariant property of the device. Hence, it can be expressed in terms of the device's small-signal parameters. The scattering parameter representation of U is the most useful in microwave applications. Ku [23] has given an expression in terms of the S-matrix:

$$U = \frac{\mid S_{12} - S_{21} \mid^2}{\Delta([I] - [S^*]^\mathrm{T}[S])}, \tag{5.64}$$

where $[I]$ is the identity matrix. U can also be expressed with the help of the stability factor k [15]:

$$U = \frac{\left|\left(\frac{S_{21}}{S_{12}}\right) - 1\right|^2}{2k\left|\frac{S_{21}}{S_{12}}\right| - 2\mathrm{Re}\left[\frac{S_{21}}{S_{12}}\right]}. \tag{5.65}$$

5.2.8 Maximum frequency of oscillation

The maximum frequency of oscillation is a criterion for a device's ability to amplify power. Because the maximum power gain at any frequency is obtained by conjugately matching the input and output ports, the MAG (see Equation (5.62)) is an obvious choice. The maximum frequency of oscillation $f_{\max}$ is then the frequency where MAG drops to unity:

$$\mathrm{MAG}(f_{\max}) = 1. \tag{5.66}$$

However, we had seen that MAG only exists when Rollet's stability factor $k \geq 1$. This raises a practical problem – for many high-performance microwave transistors, $k < 1$ in the whole measurement range, and so $f_{\max}$ cannot be determined using Equation (5.66).

Another customary definition therefore makes use of Mason's unilateral gain U (see Equation (5.65)), which does not have this limitation. $f_{\max}$ is then understood as the frequency where U drops to unity:

$$U(f_{\max}) = 1. \tag{5.67}$$

Note, however, that Equations (5.66) and (5.67) will generally not yield the same result, so it is important to check the definition used when comparing $f_{\max}$ values.

5.3 Noise in two-ports

5.3.1 Noise phenomena

Any electronic component exhibits electronic noise, provided that the absolute temperature is $T > 0$. There are several physical origins, which have been discussed in Chapter 2 in the context of active devices, so only a brief summary shall be given here.

Noise occurs in all contexts where carrier motion or carrier density is stochastic:

- As there is always a random thermal motion superimposed on any charge carrier movement, *thermal* or *Johnson noise* is omnipresent in all conductors with non-zero resistance. The rms voltage generated by the thermal noise in a resistor of value R is $V_{\mathrm{rms}} = \sqrt{4kTRB}$, where k is Boltzmann's constant and B is the measurement bandwidth.
- Emission of charge carriers over energy barriers is equally a stochastic process, and its associated noise mechanism is called *shot noise*. The rms current generated by a current of magnitude I flowing over an energy barrier is $I_{\mathrm{rms}} = \sqrt{2qIB}$, where q is the elementary charge and B again the measurement bandwidth.

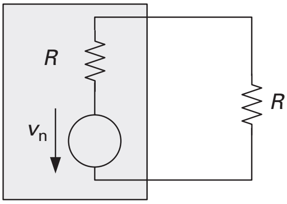


Fig. 5.11 Thévenin equivalent circuit of a noise resistor terminated by a load of equal value.

- Examples of noise due to random fluctuations in carrier density is *generation-recombination noise* or noise due to *trapping and de-trapping processes*. Unlike the first two noise phenomena, which can be considered to have a spectral density independent of frequency (‘white noise’), these processes produce noise spectra with low-pass behaviour, and cutoff frequencies in the Hz–MHz range.
- Another example of noise generated by random changes in carrier densities is *avalanche noise*, due to carrier multiplication effects in high-field regions. This process also produces a low-pass limited noise spectrum with a cutoff frequency in the GHz range.

For the discussion of noise in linear two-ports, however, the physical origin of noise is irrelevant. In fact, we will occasionally assume in the following that noise is always thermal in nature. Thermal noise has an interesting property. Consider that the squared magnitude of the noise voltage phasor generated in a resistor R in a bandwidth B is

$$\left\langle |v_\mathrm{n}|^2 \right\rangle = 8kTRB, \tag{5.68}$$

which corresponds to the rms value of $\sqrt{4kTRB}$ mentioned above. The Thévenin equivalent circuit of the noise resistor can then be drawn as in the box of Figure 5.11. The source is terminated by the same resistance R, so that the *available power* is delivered. As power is related to the peak voltage as $0.5\widehat{V}^2/R$, and the voltage drop across the resistor is $v_\mathrm{n}/2$, the resulting available power due to the thermal noise in resistor R in a bandwidth B is

$$N = \frac{\left\langle |v_\mathrm{n}|^2 \right\rangle}{8R} = kTB. \tag{5.69}$$

The available noise power of a resistor is hence independent of the resistor value and depends only on the absolute temperature T and the measurement bandwidth B.

5.3.2 Noise figure

We will now, in a general form, consider what happens to a signal as it traverses a noisy two-port. Figure 5.12 shows an arrangement where the two-port is connected to a generator (source) with generator resistance R_G and a load R_L. For now, we assume that we deal only with real impedances and that both generator and load constitute a power match, i.e. $R_\mathrm{in} = R_\mathrm{G}$, $R_\mathrm{out} = R_\mathrm{L}$. The condition of power match at the input will be dropped further down.

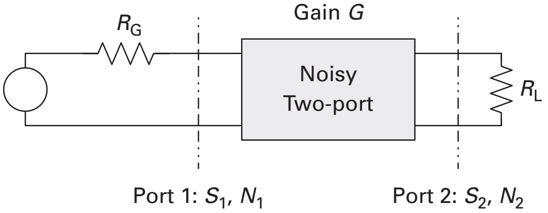


Fig. 5.12 Generic noisy two-port connected to source and load.

We further assume that the noise at the input is only due to the thermal noise of the generator resistance, and that this resistor is at a temperature of $T_0 = 290\,\text{K}$.[1] The noise power at the input port is then

$$N_1 = kT_0B. \tag{5.70}$$

In a 1 Hz bandwidth, this amounts to $4 \cdot 10^{-21}$ W or -174 dBm.[2]

The two-port will also contribute noise. To simplify things, let us assume that all noise sources inside the two-port combine into a single noise source with power N_{eq}, also located at port 1. Because the two-port has a gain of G, the noise power at the output is $N_2 = G \cdot (N_1 + N_{\text{eq}})$.

If there is an additional signal component S_1 present at the input, it equally is magnified by G. $S_2 = G \cdot S_1$. We can now define the ratio of the *signal-to-noise ratios* at the input and the output:

$$\frac{S_1/N_1}{S_2/N_2} = \frac{S_1G(N_1 + N_{\text{eq}})}{N_1GS_1} = 1 + \frac{N_{\text{eq}}}{N_1} = F \tag{5.71}$$

This defines the *noise figure F* of the two-port as the ratio of the signal-to-noise ratios at the input and the output, provided that the input carries thermal noise at a temperature $T_0 = 290\,\text{K}$ only. Note that $F > 1$ under all circumstances.

We also found the relationship between the equivalent noise power at the input and the noise figure:

$$N_{\text{eq}} = (F - 1)N_1 = (F - 1)kT_0B. \tag{5.72}$$

If we assume that N_{eq} is also thermally generated, $N_{\text{eq}} = kT_nB$, we can define an *equivalent noise temperature* for the two-port:

$$T_n = (F - 1)T_0. \tag{5.73}$$

Both F and T_n can be used interchangeably to characterise the noise performance of two-ports. The use of T_n is customary in cases where F is very close to 1, for example in low-noise amplifiers (LNAs) for satellite applications, while F is more popular for general applications.

[1] 290 K is the standard temperature for noise calculations.

[2] 1 dBm is one decibel relative to a power of 1 mW.

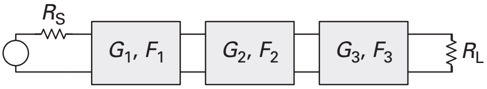


Fig. 5.13 Cascaded noisy two-ports.

Noise figure of cascaded two-ports

The situation depicted in Figure 5.12 is too simplistic for practical applications, because amplifiers or receivers generally consist of several stages. Let us consider next what happens when several noisy two-ports are being cascaded. This is shown in Figure 5.13. Again we assume that all ports are power-matched, which is an important restriction, but is made here to simplify calculations.

First, we calculate the noise at the output (delivered to the load R_L), assuming that for each two-port, the noise sources are combined into equivalent noise sources at its input. Then,

$$N_2 = G_3 \left\{N_{\mathrm{eq},3} + G_2 \left[N_{\mathrm{eq},2} + G_1 \left(N_{\mathrm{eq},1} + N_1\right)\right]\right\}. \tag{5.74}$$

Now we assume that all noise sources are being transferred to the input of the cascade. The equivalent noise source there will have a noise power of

$$N_{\mathrm{eq,tot}} = N_{\mathrm{eq},1} + \frac{N_{\mathrm{eq},2}}{G_1} + \frac{N_{\mathrm{eq},3}}{G_1 G_2}. \tag{5.75}$$

Recalling Equation (5.72), we finally calculate the noise figure of the cascaded two-ports:

$$F_{\mathrm{tot}} = 1 + \frac{N_{\mathrm{eq,tot}}}{N_1} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2}. \tag{5.76}$$

In a more generalised form, the total noise figure of a cascade of n stages is

$$F_{\mathrm{tot}} = F_1 + \sum_{i=2}^{n} \frac{F_{\mathrm{i}} - 1}{\prod\limits_{k=1}^{i-1} G_{\mathrm{k}}} \tag{5.77}$$

This is the famed *Friis equation* [11] which in effect postulates that in a receiver chain the overall noise figure is approximately the noise figure of the first stage, provided that that stage has sufficient gain – an important observation for the design of LNAs.

A consequence of the Friis formula is that in any amplifier, a low noise figure without sufficient gain is meaningless as the noise of the following stages will take over. It is therefore useful to combine noise figure and gain in a single figure of merit. This is the *noise measure* according to Haus and Adler [19]. When using noise figures, it is written as

$$M = \frac{F - 1}{1 - 1/G}, \tag{5.78}$$

where F is the noise figure and G the gain. The noise figure of an infinite chain of identical transistors is then

$$F_\infty = M + 1. \tag{5.79}$$

The noise measure can also be written in the form of a noise temperature:

$$M_\mathrm{T} = \frac{T_\mathrm{n}}{1 - 1/G}, \tag{5.80}$$

where T_n is the equivalent noise temperature of the two-port (see Equation (5.73)).

5.3.3 Noise figure with arbitrary generator admittance

We will now abandon the condition that all ports are power-matched by allowing arbitrary terminations for the input port. The source admittance $\underline{Y}_\mathrm{S}$ is expressed by its real and imaginary parts, $\underline{Y}_\mathrm{S} = G_\mathrm{S} + \jmath B_\mathrm{S}$. As before, we assume that all noise sources internal to the two-port can be combined at the input. Due to the arbitrary generator admittance, we now have to split the noise sources into a noise voltage source v and a noise current source i. This is shown in Figure 5.14.

The equivalent noise voltage and current sources, v and i, are partially correlated. This can be accounted for by introducing a *correlation admittance* $\underline{Y}_\mathrm{C}$ and splitting the current source into an uncorrelated part i_u and a fully correlated part $\underline{Y}_\mathrm{C} v$:

$$i = i_\mathrm{u} + \underline{Y}_\mathrm{C} v. \tag{5.81}$$

The real part of the generator admittance contributes a thermal noise current with the phasor:

$$\left\langle |i_\mathrm{T}|^2 \right\rangle = 8kT_0 B G_\mathrm{S}, \tag{5.82}$$

where $T_0 = 290\,\mathrm{K}$ is the standard temperature and B the measurement bandwidth.

Note that in Figure 5.14(b), all sources share the same source admittance Y_S, so that it is sufficient to calculate v_1 for the case $i_1 = 0$:

$$v_1 = v + \frac{i + i_\mathrm{T}}{Y_\mathrm{S}}. \tag{5.83}$$

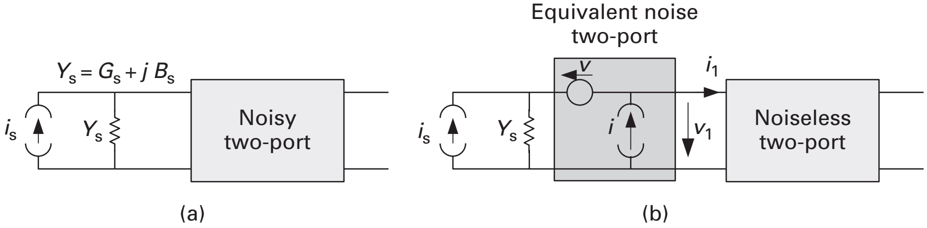


Fig. 5.14 (a) Noisy two-port with arbitrary generator admittance and (b) its equivalent circuit, introducing the equivalent noise voltage v and the equivalent noise current i.

We are really only interested in the powers, $N \propto \langle |v_1|^2 \rangle$:

$$\begin{aligned}\left\langle |v_1|^2 \right\rangle &= \langle v_1 v_1^\star \rangle \\ &= \left\langle |v|^2 \right\rangle + \frac{\langle |i|^2 \rangle + \langle |i_\mathrm{T}|^2 \rangle}{|\underline{Y}_\mathrm{S}|^2} + \frac{\langle i v^\star \rangle}{\underline{Y}_\mathrm{S}} + \frac{\langle i^\star v \rangle}{\underline{Y}_\mathrm{S}^\star} \\ &= \frac{\langle |i|^2 \rangle + \langle |i_\mathrm{T}|^2 \rangle}{|\underline{Y}_\mathrm{S}|^2} + \left\langle |v|^2 \right\rangle \left[1 + \frac{\underline{Y}_\mathrm{C}}{\underline{Y}_\mathrm{S}} + \left(\frac{\underline{Y}_\mathrm{C}}{\underline{Y}_\mathrm{S}} \right)^\star \right].\end{aligned} \tag{5.84}$$

The noise figure can be expressed as

$$F = \frac{N_{1,\mathrm{total}}}{N_{1,\mathrm{i_T}}} = \frac{|v_1|^2}{\langle |i_\mathrm{T}|^2 \rangle / |\underline{Y}_\mathrm{S}|^2}, \tag{5.85}$$

where $N_{1,\mathrm{total}}$ is the noise power due to all noise sources and $N_{1,\mathrm{i_T}}$ the noise power due to the generator thermal noise at $T = T_0$ alone.

Combining Equations (5.84) and (5.85) yields

$$F = 1 + \frac{\langle |i|^2 \rangle}{\langle |i_\mathrm{T}|^2 \rangle} + \frac{\langle |v|^2 \rangle}{\langle |i_\mathrm{T}|^2 \rangle} |\underline{Y}_\mathrm{S}|^2 \left[1 + \frac{\underline{Y}_\mathrm{C}}{\underline{Y}_\mathrm{S}} + \left(\frac{\underline{Y}_\mathrm{C}}{\underline{Y}_\mathrm{S}} \right)^\star \right]. \tag{5.86}$$

Equation (5.86) can be written in a simpler form by introducing

- an equivalent noise admittance

$$g_\mathrm{n} = \frac{\langle |i|^2 \rangle}{8kT_0 B},$$

- an equivalent noise resistance

$$R_\mathrm{n} = \frac{\langle |v|^2 \rangle}{8kT_0 B},$$

- and remembering that

$$G_\mathrm{S} = \frac{\langle |i_\mathrm{T}|^2 \rangle}{8kT_0 B},$$

from Equation (5.82).

This results in

$$F = 1 + \frac{g_\mathrm{n}}{G_\mathrm{S}} + \frac{R_\mathrm{n}}{G_\mathrm{S}} |\underline{Y}_\mathrm{S}|^2 \left[1 + \frac{\underline{Y}_\mathrm{C}}{\underline{Y}_\mathrm{S}} + \left(\frac{\underline{Y}_\mathrm{C}}{\underline{Y}_\mathrm{S}} \right)^\star \right]. \tag{5.87}$$

The noise figure depends therefore on the generator admittance $Y_\mathrm{S} = G_\mathrm{S} + \jmath Y_\mathrm{S}$. This was first pointed out by Rothe and Dahlke in 1955 [31, 32].

We can now search for the generator admittance where F will be minimal. This results in

$$\begin{aligned} G_{\mathrm{S,opt}} &= \sqrt{\frac{g_\mathrm{n}}{R_\mathrm{n}} - B_\mathrm{C}^2} \\ B_{\mathrm{S,opt}} &= -B_\mathrm{C}, \end{aligned} \tag{5.88}$$

where B_C is the imaginary part of the correlation admittance.

Introducing these terms into Equation (5.87), the noise figure can be written as follows:

$$F = 1 + 2R_\mathrm{n}(G_\mathrm{S,opt} + G_\mathrm{C}) + \frac{R_\mathrm{n}}{G_\mathrm{S}}\left[(G_\mathrm{S} - G_\mathrm{S,opt})^2 + (B_\mathrm{S} + B_\mathrm{S,opt})^2\right]. \tag{5.89}$$

For $Y_\mathrm{S} = Y_\mathrm{S,opt}$, the minimum noise figure is

$$F_\mathrm{min} = 1 + 2R_\mathrm{n}(G_\mathrm{S,opt} + G_\mathrm{C}) = 1 + 2R_\mathrm{n}\left(G_\mathrm{C} + \sqrt{\frac{g_\mathrm{n}}{R_\mathrm{n}} - B_\mathrm{C}^2}\right), \tag{5.90}$$

and therefore finally

$$\begin{aligned} F &= F_\mathrm{min} + \frac{R_\mathrm{N}}{G_\mathrm{S}}\left[(G_\mathrm{S} - G_\mathrm{S,opt})^2 + (B_\mathrm{S} - B_\mathrm{S,opt})^2\right] \\ &= F_\mathrm{min} + \frac{R_\mathrm{N}}{G_\mathrm{S}}\left|\underline{Y}_\mathrm{S} - \underline{Y}_\mathrm{S,opt}\right|^2. \end{aligned} \tag{5.91}$$

To describe the noise performance of a two-port, we need the following parameters:

- the minimum noise figure F_min,
- the equivalent noise resistance R_n,
- the noise-optimised generator admittance $\underline{Y}_\mathrm{S,opt}$.

For microwave applications, it is more customary to work with reflection coefficients instead of impedances or admittances. For this, we introduce a normalising admittance Y_0, typically 20 mS, and a normalised noise resistance $r_\mathrm{n} = R_\mathrm{n}Y_0$. Then Equation (5.91) turns into

$$F = F_\mathrm{min} + \frac{4r_\mathrm{n}\left|\Gamma_\mathrm{S} - \Gamma_\mathrm{S,opt}\right|^2}{\left(1 - |\Gamma_\mathrm{S}|^2\right)\left|1 + \Gamma_\mathrm{S,opt}\right|^2}, \tag{5.92}$$

where

$$\Gamma_\mathrm{S} = \frac{Y_0 - Y_\mathrm{S}}{Y_0 + Y_\mathrm{S}}.$$

Plotting $F = f(\Gamma_\mathrm{S})$ on a Smith chart, the contours of constant F are circles. An example is shown in Figure 5.15.

Associated gain

The condition for noise-optimised source reflection coefficient $\Gamma_\mathrm{S} = \Gamma_\mathrm{S,opt}$ is independent of the power matching conditions derived earlier, $\Gamma_\mathrm{S} = \Gamma_\mathrm{in}^\star$. The maximum gain under ‘noise matching’ conditions is therefore smaller than the MAG, which would be achieved when input and output of the two-port are conjugately matched. Here, $\Gamma_\mathrm{S} = \Gamma_\mathrm{S,opt}$ and the gain for the conjugately matched output becomes (see Equation (5.53))

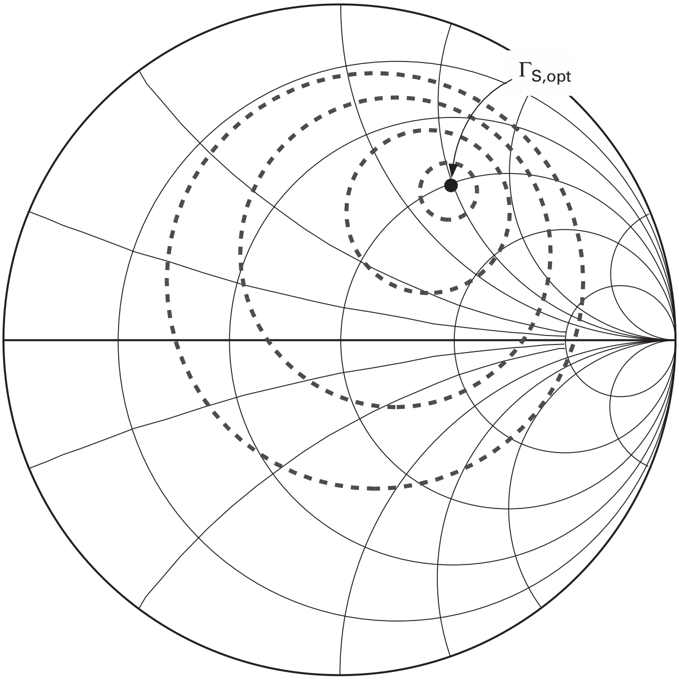


Fig. 5.15 Example for 'noise circles' – contours of constant noise figures on the Smith chart.

$$G_{\text{ass}} = \frac{1- \mid \Gamma_{\text{S,opt}} \mid^2}{\mid 1 - S_{11}\Gamma_{\text{S,opt}} \mid^2} \mid S_{21} \mid^2 \frac{1}{1- \mid \Gamma_{\text{out}} \mid^2} \tag{5.93}$$

$$= \frac{|S_{21}|^2 \left(1 - \left|\Gamma_{\text{S,opt}}\right|^2\right)}{\left|1 - S_{11}\Gamma_{\text{S,opt}}\right|^2 - \left|S_{22}(1 - S_{11}\Gamma_{\text{S,opt}}) + S_{12}S_{21}\Gamma_{\text{S,opt}}\right|^2},$$

since

$$\Gamma_{\text{out}} = S_{22} + \frac{S_{21}S_{12}\Gamma_{\text{S,opt}}}{1 - S_{11}\Gamma_{\text{S,opt}}}.$$

The available gain under noise matching conditions is called *associated gain*.

5.4 Transistor amplifiers

5.4.1 A brief historical discourse

Amplifiers are such an integral part of any wireless communication system that we have to explicitly recall that in the first decades of radio, they were not used at all. It was not before the invention of the electron tube triode that amplification of alternating current signals became possible. Lee De Forest's Audion tube (conceived in 1906 and patented in 1908 [10]) is probably the first example of an amplifying device. It was 24-year-old Edwin Armstrong, the prolific and finally tragic inventor of radio's early days, who

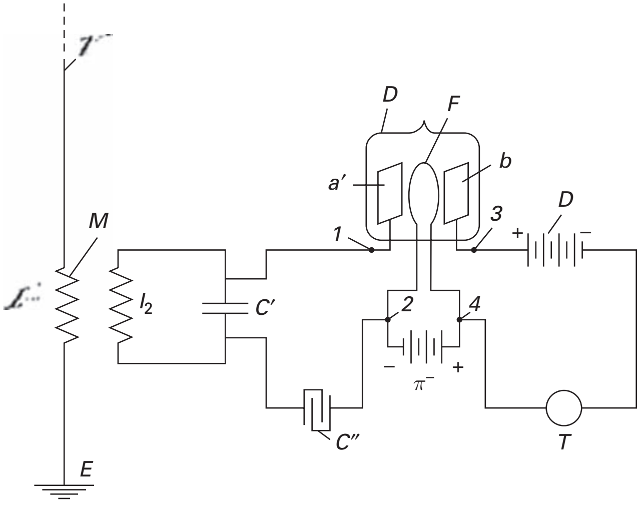


Fig. 5.16 De Forest’s three-terminal audion (from US Patent No. 879,582).

actually explained its operation in 1914. As Figure 5.16 shows, the De Forest Audion already has the common three-terminal arrangement which we also find in transistor amplifiers. De Forest’s claim to have invented the tube concept independently of the ‘thermionic valve’ patented by John Ambrose Fleming in 1905 [9] is doubtful, however.

Critical amplifier parameters vary with respect to where in a system it is being used, and what kinds of signals are being fed through the amplifiers. For example:

- When dealing with very low-level signals, for example immediately after the receiving antenna, or behind the optoelectronic converter in a fibre communication system, it is of paramount importance that the amplifier itself adds as little electronic noise as possible.
- At the output of a wireless transmitter, an amplifier should produce the required RF power with optimum efficiency and/or the required linearity, which in turn depends on the modulation format used.
- Other systems, such as in fibre-optic communications or ultra-wideband wireless systems, require extreme bandwidths, with little variation of gain and group velocity between the lower and upper cutoff frequencies.

These requirements can rarely be met simultaneously, so trade-offs have to be made which require a thorough study of the system specifications before the amplifier design is begun.

5.4.2 Fundamental amplifier configurations

Amplifiers will be discussed initially at a certain level of abstraction in order to make clear that the fundamental methods apply to FETs and bipolar transistors alike. The use of general methods will be emphasised rather than introducing a large number of circuit topologies. To this end, we will first introduce a generalised equivalent circuit for an

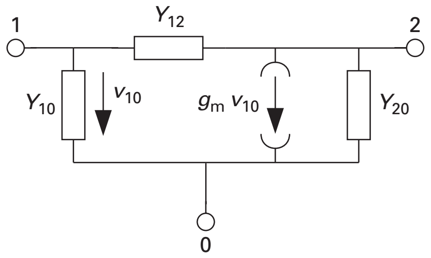


Fig. 5.17 Generic three-terminal amplifying device.

amplifying device, and then calculate its small-signal parameters – the y matrix in this case. Based on the y matrix, we develop general expressions for important amplifier parameters – the input and output admittances, and voltage and current gain. For different circuit topologies, we then calculate modified y matrices, which will immediately yield the amplifier parameters.

The move from abstract concept to actual circuit implementation will also be made occasionally. You will see that the elements of real devices can be readily matched to the general equivalent circuit.

The first important observation in amplifier design is that three-terminal devices can be used in three fundamentally different ways. Let us consider the generic three-terminal device depicted in Figure 5.17.

We restrict our discussion to linear behaviour for the time being – corresponding to the small-signal case. The amplifying action is due to the voltage-controlled current source between nodes 2 and 0. The controlling entity is the input voltage v_{10} and the parameter is the transconductance $\underline{g}_{\mathrm{m}}$, which is generally taken as a complex value. This allows us to include additional phase delays in $\underline{g}_{\mathrm{m}}$:

$$\underline{g}_{\mathrm{m}} = g_{\mathrm{m}0} e^{-\mathrm{j}\omega\tau} \tag{5.94}$$

Additionally, we included complex impedances $\underline{Y}_{10}$, $\underline{Y}_{20}$ and $\underline{Y}_{12}$. These can later be correlated with the parameters of FET or bipolar transistor small-signal equivalent circuits introduced in Chapter 2.

y matrix representation

For the 'hybrid π' equivalent circuit in Figure 5.17, the y matrix is easily calculated and will be used here. The advantage is that general amplifier properties, such as voltage and power gain or input and output admittance with arbitrary terminations, can be calculated once and can then be easily applied to the different topologies we will consider.

The y matrix expresses the following system of linear equations with respect to the two-port shown in Figure 5.18:

$$i_1 = y_{11}\, v_1 + y_{12}\, v_2 \tag{5.95}$$

$$i_2 = y_{21}\, v_1 + y_{22}\, v_2 \tag{5.96}$$

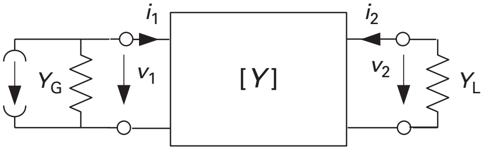


Fig. 5.18 A two-port in y matrix representation terminated with generator impedance $\underline{Y}_G$ and load impedance $\underline{Y}_L$.

The y matrix $[Y]$ can also be calculated from the scattering matrix [S] introduced earlier:

$$[Y] = Y_0 \cdot ([1] - [S])\,([1] + [S])^{-1} \tag{5.97}$$

where [1] is the identity matrix and Y_0 the normalising admittance used in the calculation of the scattering matrix – usually 20 mS.

First, consider the input admittance Y_1. The output port is terminated by an admittance $\underline{Y}_L$.

$$Y_1 = \frac{i_1}{v_1}$$
$$i_2 = -v_2\,\underline{Y}_L.$$

We find

$$Y_1 = y_{11} - \frac{y_{12}\,y_{21}}{Y_L + y_{22}}. \tag{5.98}$$

Likewise, the output admittance when the input is terminated with an arbitrary admittance $\underline{Y}_G$ is

$$Y_2 = y_{22} - \frac{y_{12}\,y_{21}}{Y_G + y_{11}}. \tag{5.99}$$

The forward voltage gain $A_V = v_2/v_1$ is

$$A_V = \frac{-y_{21}}{Y_L + y_{22}}. \tag{5.100}$$

And finally the current gain $A_I = i_2/i_1$:

$$A_I = \frac{y_{21}\,\underline{Y}_L}{y_{11}(y_{22} + \underline{Y}_L) - y_{21}\,y_{12}}. \tag{5.101}$$

The frequently used short-circuit current gain can be easily calculated from Equation (5.101) when $\underline{Y}_L \to \infty$:

$$h_{21} = \frac{i_2}{i_1}\,|(v_2 = 0) = \frac{y_{21}}{y_{11}}. \tag{5.102}$$

Common source/common emitter: node 0 as the common node

The most obvious connection is to ground node 0. In FETs, this configuration is called *common source*; in bipolar transistors, it is called *common-emitter* configuration.

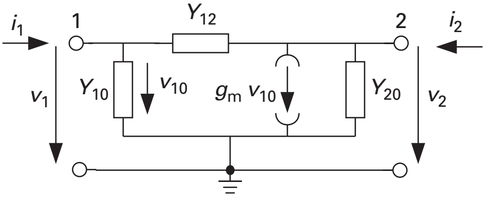


Fig. 5.19 Generic amplifier configuration with node 0 grounded.

The y matrix can be easily calculated (see Figure 5.19):

$$y_{11} = \underline{Y}_{10} + \underline{Y}_{12} \tag{5.103}$$

$$y_{12} = -\underline{Y}_{12} \tag{5.104}$$

$$y_{21} = \underline{g}_{\mathrm{m}} - \underline{Y}_{12} \tag{5.105}$$

$$y_{22} = \underline{Y}_{20} + \underline{Y}_{12}. \tag{5.106}$$

Let us first investigate the voltage gain using Equation (5.100):

$$A_{\mathrm{V}} = -\frac{y_{21}}{\underline{Y}_{\mathrm{L}} + y_{22}} = -\frac{\underline{g}_{\mathrm{m}} - \underline{Y}_{12}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20} + \underline{Y}_{12}}. \tag{5.107}$$

To facilitate interpretation, let us assume that $\underline{g}_{\mathrm{m}}$ is real, and that the feedback admittance is weak: $g_{\mathrm{m}} - \underline{Y}_{12} \approx g_{\mathrm{m}}$, $\underline{Y}_{\mathrm{L}} + \underline{Y}_{20} + \underline{Y}_{12} \approx \underline{Y}_{\mathrm{L}} + \underline{Y}_{20}$. Then,

$$A_{\mathrm{V}} \approx -\frac{g_{\mathrm{m}}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20}}.$$

For low frequencies, $\underline{Y}_{\mathrm{L}}$ and $\underline{Y}_{20}$ are real, and we find:

- The magnitude of voltage gain in common-source and common-emitter stages will be approximately equal to the product of transconductance and effective load resistance (the parallel connection of external load resistance and device output resistance).
- The output voltage lags the input voltage by 180°.

A useful figure of merit is the maximum voltage gain a three-terminal device can produce with node 0 grounded. We find it from Equation (5.107) by choosing $Y_{\mathrm{L}} = 0$ and assuming $\underline{Y}_{12} \ll \underline{Y}_{20}, \underline{g}_{\mathrm{m}}$:

$$A_{\mathrm{V,max}} \approx -\frac{\underline{g}_{\mathrm{m}}}{\underline{Y}_{20}}. \tag{5.108}$$

Now let us take a look at the input admittance, using Equation (5.98):

$$Y_1 = y_{11} - \frac{y_{12}\, y_{21}}{Y_{\mathrm{L}} + y_{22}} = \underline{Y}_{10} + \underline{Y}_{12}\left(1 + \frac{g_{\mathrm{m}} - \underline{Y}_{12}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20} + \underline{Y}_{12}}\right),$$

or recognising from Equation (5.107) that the last term in parentheses is A_{V}:

$$Y_1 = \underline{Y}_{10} + \underline{Y}_{12}(1 - A_{\mathrm{V}}). \tag{5.109}$$

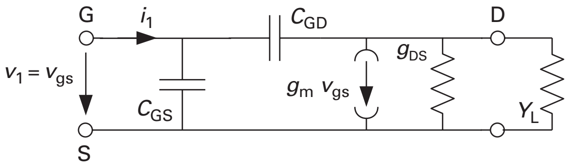


Fig. 5.20 Simplified hybrid π equivalent circuit of a FET with load admittance.

Miller effect

It is now time for a small practical example. In a FET, the small-signal equivalent circuit looks like Figure 5.20, if we neglect the series resistances. From comparison with Figure 5.17, we recognise that

$$\underline{Y}_{10} = \jmath\omega C_{GS}$$
$$\underline{Y}_{12} = \jmath\omega C_{GD}$$
$$\underline{Y}_{20} = g_{DS}.$$

The voltage gain is now

$$A_V = -\frac{g_m - \jmath\omega C_{GD}}{Y_L + g_{DS} + \jmath\omega C_{GD}} \approx -\frac{g_m}{Y_L + g_{DS}},$$

for low frequencies.

The input admittance is, using Equation (5.109),

$$Y_1 = \jmath\omega C_{GS} + \jmath\omega C_{GD}\,(1 - A_V)\,. \tag{5.110}$$

The feedback capacitance C_{GD} appears in parallel with C_{GS}, but is multiplied by the magnitude of the voltage gain, augmented by one – this is the dreaded *Miller Effect*, which we always have to be aware of in high-speed circuit design, because it may significantly increase the input capacitance. It was described as early as 1920 [29].

Going back to Figure 5.19, we calculate the output admittance to be

$$Y_2 = \underline{Y}_{20} + \underline{Y}_{12}\left(1 + \frac{\underline{g}_m - \underline{Y}_{12}}{\underline{Y}_G + \underline{Y}_{10} + \underline{Y}_{12}}\right). \tag{5.111}$$

The term

$$\frac{g_m - Y_{12}}{Y_G + Y_{10} + Y_{12}} = A_r$$

can be significantly larger than 1 and may act like a ‘reverse Miller Effect’. This has to be taken into account if a tuned circuit is connected to the output node, as is the case in typical tuned amplifiers. Figure 5.21 shows an example – the detuning effect of the feedback capacitance will be much larger than expected.

The current gain of the common-source/common-emitter amplifier, finally, becomes

$$A_I = \frac{(\underline{g}_m - \underline{Y}_{12})\,\underline{Y}_L}{\underline{Y}_{12}(\underline{Y}_L + \underline{g}_m + \underline{Y}_{10} + \underline{Y}_{20}) + \underline{Y}_{10}(\underline{Y}_L + \underline{Y}_{20})}. \tag{5.112}$$

We can safely assume here that the real parts of all admittances are larger than zero – then the current gain in the quasi-static limit ($f \to 0$) is positive.

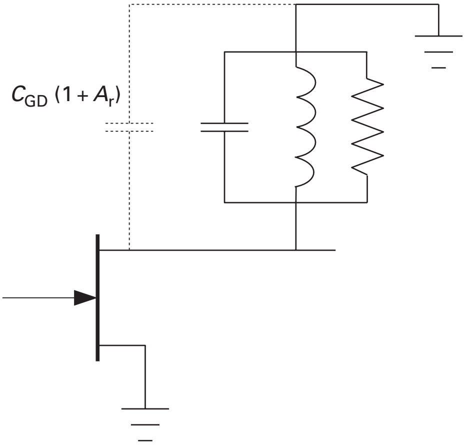


Fig. 5.21 Common-source amplifier stage with tuned load, and the output-referred Miller capacitance.

Let us check the latter equation again with our FET equivalent circuit by calculating the short-circuit current gain:

$$h_{21} = A_I|(\underline{Y}_L \to \infty) = \frac{g_m - \jmath\omega C_{GD}}{\jmath\omega(C_{GS} + C_{GD})} \approx \frac{g_m}{\jmath\omega C_{GS}}$$

assuming that $g_m \gg \omega C_{GD}$ and $C_{GS} \gg C_{GD}$.

This is the equation we had earlier used in Chapter 2 to estimate the transit frequency of a FET to be $f_T = g_m/(2\pi C_{GS})$.

Let us summarise our findings for the amplifier configuration with node 0 grounded – applicable to both the FET common-source and the bipolar common-emitter topologies:

- The configuration can provide substantial voltage and current gains.
- The voltage gain provides a phase shift of 180° in the quasi-static limit, whereas the current gain experiences no phase shift.
- Due to the presence of feedback, the input and output admittances always depend on the termination of the opposite port.
- For substantial voltage gains, the *Miller effect* has to be observed which can substantially increase the input capacitance.

Common gate/common base: node 1 as the common node

In the next set-up to be discussed, node 1 is grounded, and the input current is fed into node 0. Node 2 is still the output node. This applies to the FET common-gate and the bipolar common-base configurations. Again, we will first calculate the y matrix in a general form. For this, it is valuable to recognise in Figure 5.22 that $v_1 = -v_{10}$.

$$y_{11} = \underline{Y}_{20} + \underline{Y}_{10} + \underline{g}_m \tag{5.113}$$

$$y_{12} = -\underline{Y}_{20} \tag{5.114}$$

$$y_{21} = -\underline{Y}_{20} - \underline{g}_m \tag{5.115}$$

$$y_{22} = \underline{Y}_{20} + \underline{Y}_{12}. \tag{5.116}$$

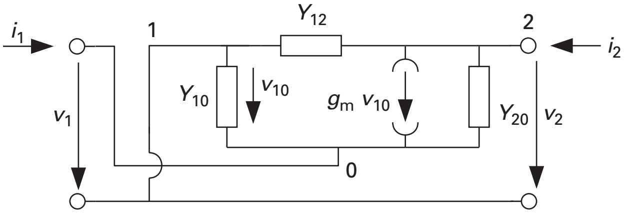


Fig. 5.22 Hybrid π equivalent circuit with node 1 grounded.

This input admittance is then

$$Y_1 = \underline{Y}_{10} + (\underline{g}_{\mathrm{m}} + \underline{Y}_{20})\frac{\underline{Y}_{\mathrm{L}} + \underline{Y}_{12}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20} + \underline{Y}_{12}}. \tag{5.117}$$

Let us simplify this expression somewhat. First, realise that in practical devices necessarily $\underline{g}_{\mathrm{m}} \gg \underline{Y}_{20}$, otherwise it would not have a reasonable voltage gain in common-source or common-emitter configuration (see above). Further, let us assume that $\underline{Y}_{12} \ll \underline{Y}_{\mathrm{L}}$. Then,

$$Y_1 \approx \underline{Y}_{10} + \underline{g}_{\mathrm{m}}\frac{\underline{Y}_{\mathrm{L}}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20}}.$$

An interesting observation is that now $\underline{Y}_{20}$ is the feedback admittance which determines the sensitivity of the input admittance on the load. If further, this admittance is very small, $\underline{Y}_{20} \ll \underline{Y}_{\mathrm{L}}$, then

$$Y_1 \approx \underline{Y}_{10} + \underline{g}_{\mathrm{m}}.$$

Using the example of Figure 5.20, we find

$$Y_1 \approx g_{\mathrm{m}}\left(1 + \frac{j\omega C_{\mathrm{GS}}}{g_{\mathrm{m}}}\right) = g_{\mathrm{m}}\left(1 + j\frac{\omega}{\omega_{\mathrm{T}}}\right).$$

The input admittance is hence approximately the transconductance, unless we are operating close to f_{T}. In practical transistors, this will be a quite large value – much larger than the input admittance in the topology with node 0 grounded.

The output admittance is

$$Y_2 = \underline{Y}_{12} + \underline{Y}_{20}\frac{\underline{Y}_{\mathrm{G}} + \underline{Y}_{10}}{\underline{Y}_{\mathrm{G}} + \underline{Y}_{10} + \underline{Y}_{20} + \underline{g}_{\mathrm{m}}}. \tag{5.118}$$

Assuming again that $\underline{g}_{\mathrm{m}} \gg \underline{Y}_{20}$,

$$Y_2 \approx \underline{Y}_{12} + \frac{\underline{Y}_{20}}{1 + \dfrac{\underline{g}_{\mathrm{m}}}{\underline{Y}_{\mathrm{G}} + \underline{Y}_{10}}}$$

Compared to the topology with node 0 grounded (common emitter or common source), the output admittance will be substantially smaller.

The voltage gain is calculated as

$$A_{\mathrm{V}} = \frac{\underline{g}_{\mathrm{m}} + \underline{Y}_{20}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20} + \underline{Y}_{12}}. \tag{5.119}$$

Assuming again that $\underline{g}_{\mathrm{m}} \gg \underline{Y}_{20}$, this simplifies to

$$A_{\mathrm{V}} = \frac{\underline{g}_{\mathrm{m}}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20} + \underline{Y}_{12}}.$$

Compare with Equation (5.107) – this is the same result, if $\underline{Y}_{12} \ll g_{\mathrm{m}}$, which is a safe assumption.

Finally, the current gain is

$$A_{\mathrm{I}} = -\frac{\underline{Y}_{\mathrm{L}}\underline{g}_{\mathrm{m}}}{(\underline{Y}_{\mathrm{L}} + \underline{Y}_{12})(\underline{Y}_{20} + \underline{Y}_{10} + \underline{g}_{\mathrm{m}}) + \underline{Y}_{20}\,\underline{Y}_{10}}. \tag{5.120}$$

The magnitude of A_{I} for this topology is hence less than 1. For greater simplicity, calculate the short-circuit current gain ($Y_{\mathrm{L}} \to \infty$):

$$\begin{aligned} h_{21} &= -\frac{\underline{Y}_{20} + \underline{g}_{\mathrm{m}}}{\underline{Y}_{20} + \underline{g}_{\mathrm{m}} + \underline{Y}_{10}} \\ &\approx -\frac{1}{1 + \dfrac{\underline{Y}_{10}}{\underline{g}_{\mathrm{m}}}}. \end{aligned}$$

Using again the simple FET equivalent circuit in Figure 5.20, this reduces to

$$h_{21} = -\frac{1}{1 + \jmath\dfrac{\omega}{\omega_{\mathrm{T}}}}.$$

In other words, the short-circuit current gain of the common-gate and common-base configurations will remain independent of frequency until quite close to f_{T}.

Common drain/common collector: node 2 as the common node

The last fundamental configuration of the generic amplifying three-terminal device (Figure 5.17) has node 2 as the common node (see Figure 5.23). In FETs, this will be called *common drain*; in bipolar transistors, this will be called *common-collector* configuration.

Again, first calculate the y matrix of this configuration.

$$y_{11} = \underline{Y}_{10} + \underline{Y}_{12} \tag{5.121}$$

$$y_{12} = -\underline{Y}_{10} \tag{5.122}$$

$$y_{21} = -(\underline{Y}_{10} + \underline{g}_{\mathrm{m}}) \tag{5.123}$$

$$y_{22} = \underline{Y}_{10} + \underline{Y}_{20} + \underline{g}_{\mathrm{m}}. \tag{5.124}$$

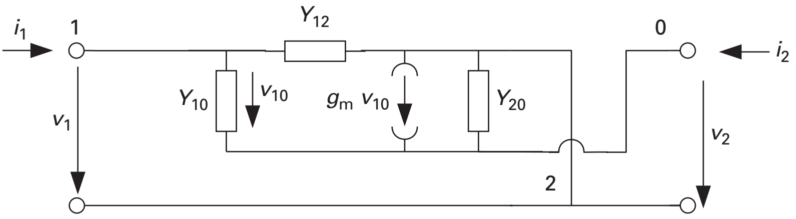


Fig. 5.23 Generic amplifier configuration with node 2 grounded.

Let us first consider the voltage gain A_V, when port 2 is terminated with a load admittance Y_L.

$$A_\mathrm{V} = -\frac{y_{21}}{Y_\mathrm{L} + y_{22}} = \frac{\underline{Y}_{10} + \underline{g}_\mathrm{m}}{\underline{Y}_\mathrm{L} + \underline{Y}_{10} + \underline{Y}_{20} + \underline{g}_\mathrm{m}} = \left(1 + \frac{\underline{Y}_\mathrm{L} + \underline{Y}_{20}}{\underline{Y}_{10} + \underline{g}_\mathrm{m}}\right)^{-1}. \qquad (5.125)$$

In practical devices, the second term in parentheses will be small compared to 1, at least at lower frequencies, so that $A_\mathrm{V} \approx 1$ (but always less than 1) – v_2 follows v_1, which is why this topology is also called a *source follower* or *emitter follower* for FETs and bipolar transistors, respectively.

The input admittance is calculated to be

$$\begin{aligned} Y_1 = y_{11} - \frac{y_{21}y_{12}}{Y_\mathrm{L} + y_{22}} &= \underline{Y}_{12} + \frac{(\underline{Y}_\mathrm{L} + \underline{Y}_{20})\underline{Y}_{10}}{\underline{Y}_\mathrm{L} + \underline{Y}_{20} + \underline{Y}_{10} + \underline{g}_\mathrm{m}} \\ &= \underline{Y}_{12} + \frac{\underline{Y}_{10}}{1 + \dfrac{\underline{Y}_{10} + \underline{g}_\mathrm{m}}{\underline{Y}_\mathrm{L} + \underline{Y}_{20}}} \qquad (5.126) \\ &\approx \underline{Y}_{12} + \frac{\underline{Y}_{10}}{1 + \dfrac{\underline{g}_\mathrm{m}}{\underline{Y}_\mathrm{L}}}, \qquad (5.127) \end{aligned}$$

because typically, at least at frequencies sufficiently below f_T, $\underline{g}_\mathrm{m} \gg \underline{Y}_{10}$, and assuming that $\underline{Y}_\mathrm{L} \gg \underline{Y}_{20}$.

Compare this to the input admittance of the topology where node 0 was grounded, Equation (5.109), and you will notice that the influence of $\underline{Y}_{10}$ is substantially reduced, while $\underline{Y}_{12}$ does not suffer from the augmentation due to the Miller effect. We can therefore state that the topology with node 2 grounded presents a much lower input admittance.

Going to our usual FET example where the amplifying device is represented by the equivalent circuit in Figure 5.20, we can show that Equation (5.126) may have an unexpected result. In this case,

$$Y_1 = j\omega C_\mathrm{GD} + \frac{\jmath\omega C_\mathrm{GS}}{1 + \dfrac{g_\mathrm{m} + \jmath\omega C_\mathrm{GS}}{Y_\mathrm{L} + g_\mathrm{DS}}}.$$

Now assume that $g_\mathrm{m} \gg \omega C_\mathrm{GS}$, i.e. $\omega \ll \omega_\mathrm{T}$, that Y_L is capacitive ($Y_\mathrm{L} = \jmath\omega C_\mathrm{L}$), and that $\omega C_\mathrm{L} \gg g_\mathrm{DS}$.

$$Y_1 = \jmath\omega C_\mathrm{GD} - \frac{\omega^2 C_\mathrm{GS} C_\mathrm{L}}{g_\mathrm{m} + \jmath\omega(C_\mathrm{L} + C_\mathrm{GS})},$$

or separating into real and imaginary parts,

$$Y_1 = -\frac{\omega^2 g_\mathrm{m} C_\mathrm{GS} C_\mathrm{L}}{g_\mathrm{m}^2 + \omega^2(C_\mathrm{L} + C_\mathrm{GS})^2} + \jmath\omega\left(\frac{\omega^2 C_\mathrm{GS} C_\mathrm{L}(C_\mathrm{L} + C_\mathrm{GS})}{g_\mathrm{m}^2 + \omega^2(C_\mathrm{L} + C_\mathrm{GS})^2} + C_\mathrm{GD}\right). \qquad (5.128)$$

We created an admittance with a negative real part! This can be useful, for example if we want to build an oscillator (see Section 5.5), but also very dangerous for amplifier stability.

Moving back to the more general case, let us calculate the current gain:

$$A_{\mathrm{I}} = -\frac{\underline{Y}_{\mathrm{L}}(\underline{Y}_{10} + \underline{g}_{\mathrm{m}})}{\underline{Y}_{12}(\underline{Y}_{10} + \underline{Y}_{20} + \underline{Y}_{\mathrm{L}} + \underline{g}_{\mathrm{m}}) + \underline{Y}_{10}(\underline{Y}_{\mathrm{L}} + \underline{Y}_{20})}. \tag{5.129}$$

Comparing this with the equation for the current gain with node 0 grounded (Equation (5.112)), we see that the denominators are equal. If further $\underline{g}_{\mathrm{m}} \gg \underline{Y}_{10}, \underline{Y}_{12}$, which is typically the case, the two current gains have equal magnitude.

Finally, the output admittance of the topology with node 2 grounded will be calculated. The input port is terminated with a generator admittance Y_{G}.

$$Y_2 = \underline{Y}_{20} + \frac{(\underline{Y}_{10} + \underline{g}_{\mathrm{m}})(\underline{Y}_{\mathrm{G}} + \underline{Y}_{12})}{\underline{Y}_{\mathrm{G}} + \underline{Y}_{10} + \underline{Y}_{12}}. \tag{5.130}$$

To interpret this equation, assume that $Y_{\mathrm{G}} \gg \underline{Y}_{10} + \underline{Y}_{12}$, $\underline{g}_{\mathrm{m}} \gg \underline{Y}_{10}$. Then,

$$Y_2 \approx \underline{Y}_{20} + \underline{g}_{\mathrm{m}} \approx \underline{g}_{\mathrm{m}},$$

in most cases. Compared to the output admittance of the original topology which had node 0 as the common node (Equation (5.111)), we see that now we have a substantially higher output conductance.

The combination of a very low input conductance (very high input impedance) and high output conductance (low output resistance) is the most important aspect of common-drain/common-collector topologies.

5.4.3 Feedback

Negative feedback is another important principle in amplifier design. In small-signal design, it is used for impedance matching purposes, to make an amplifier stable and to increase its bandwidth. The negative feedback amplifier was invented by Harold Black in 1927 [22].

The most important feedback implementations in high-speed amplifier design are *shunt–shunt* and *series–series* feedback, as shown in Figure 5.24.

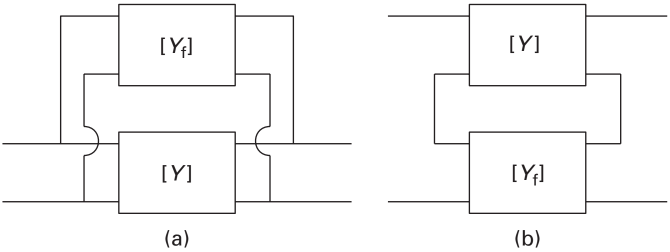


Fig. 5.24 Feedback configurations: (a) shunt–shunt feedback and (b) series–series feedback.

Shunt–shunt feedback

Case (a) is easily calculated using a y matrix representation, because the resulting y matrix is the sum of the individual matrices.

$$[Y_{\mathrm{T}}] = [Y] + [Y_{\mathrm{f}}] = \begin{bmatrix} y_{11} + y_{11,f} & y_{12} + y_{12,f} \\ y_{21} + y_{21,f} & y_{22} + y_{22,f}. \end{bmatrix} \tag{5.131}$$

Series–series feedback

Series–series feedback is better treated using a z matrix representation:

$$v_1 = z_{11} i_1 + z_{12} i_2 \tag{5.132}$$

$$v_2 = z_{21} i_1 + z_{22} i_2. \tag{5.133}$$

Conversion from y to z matrix is easy because the z matrix is simply the inverse of the y matrix:

$$[Z] = \frac{1}{\Delta(Y)} \begin{bmatrix} y_{22} & -y_{12} \\ -y_{21} & y_{11} \end{bmatrix}, \tag{5.134}$$

where $\Delta(Y)$ is the determinant of the y matrix and $\Delta(Y) = y_{11} y_{22} - y_{12} y_{21}$.

Once the z matrices have been obtained, the resulting z matrix of the circuit with series–series feedback is the sum of the individual z matrices:

$$[Z_T] = [Z] + [Z_{\mathrm{f}}] = \begin{bmatrix} z_{11} + z_{11,\mathrm{f}} & z_{12} + z_{12,\mathrm{f}} \\ z_{21} + z_{21,\mathrm{f}} & z_{22} + z_{22,\mathrm{f}} \end{bmatrix}. \tag{5.135}$$

The conversion back from z to y matrix is equally simple:

$$[Y] = \frac{1}{\Delta(Z)} \begin{bmatrix} z_{22} & -z_{12} \\ -z_{21} & z_{11} \end{bmatrix}. \tag{5.136}$$

Use of feedback in small-signal amplifiers

A very common feedback example is shown in Figure 5.25, where an admittance Y_{f} is connected between the output and the input of a common-source amplifier. The FET

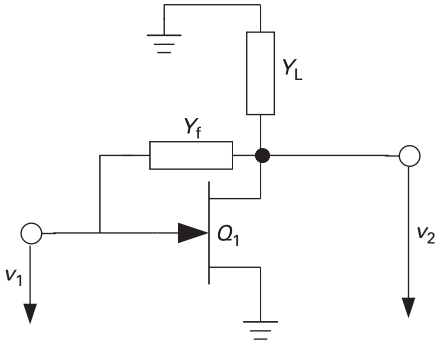


Fig. 5.25 Example of shunt–shunt feedback in a common-source amplifier.

shall be treated using the generic equivalent circuit in Figure 5.17. The feedback two-port contains only one element, Y_f. If $[Y_{\mathrm{Q1}}]$ is the y matrix of transistor Q_1, then the y matrix of the transistor with feedback is

$$[Y_\mathrm{T}] = \begin{bmatrix} y_{11,\mathrm{Q1}} + Y_\mathrm{f} & y_{12,\mathrm{Q1}} - Y_\mathrm{f} \\ y_{21,\mathrm{Q1}} - Y_\mathrm{f} & y_{22,\mathrm{Q1}} + Y_\mathrm{f} \end{bmatrix}. \tag{5.137}$$

Unilateralisation

An immediate application of this feedback technique is the elimination of the parameter y_{12}. Choosing

$$Y_\mathrm{f} = -y_{12,\mathrm{Q1}}$$

results in a *unilateralised* amplifier two-port where the input parameters no longer depend on the output load, and vice versa. This can be used to improve amplifier stability, and is referred to as *neutralisation*.

There are several ways of achieving this. In narrow-band amplifiers, the usually purely capacitive feedback may be tuned out using an inductor. The inductor is chosen to form a parallel resonance with the feedback capacitor at the frequency of operation.

A more elegant technique was invented by Harold A. Wheeler in the early 1920s for electron tubes. A current with equal magnitude – but opposite phase, as the current through the feedback admittance – is fed back from the output to the input node, where the two currents cancel out exactly. This is shown in Figure 5.26. In integrated circuits, the realisation of the autotransformer is hampered by the typically high losses of on-chip inductors. However, any kind of phase reversal will do; a particularly simple technique will be discussed further down in the context of the differential amplifier (p. 336).

Port matching

A very common task in amplifier design is matching, e.g. the input admittance Y_1 to the generator admittance: $Y_1 = Y_\mathrm{G}^*$, where Y_G^* is the complex conjugate of the generator

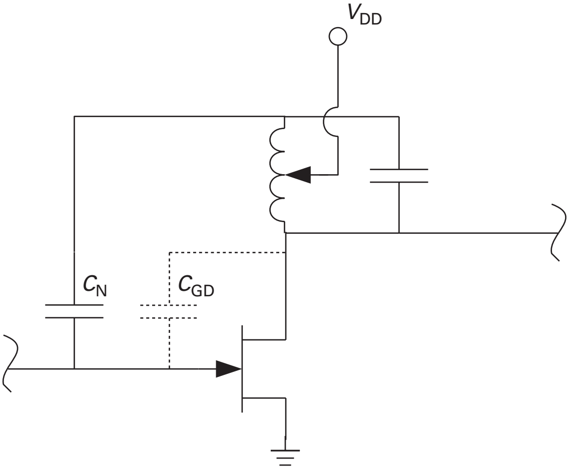


Fig. 5.26 Amplifier neutralisation using an autotransformer.

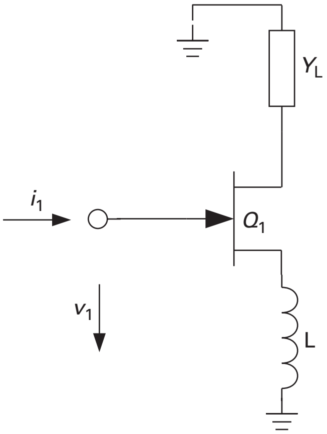


Fig. 5.27 Inductive emitter degeneration as an example for series–series feedback.

admittance. While this is commonly done by cascading a matching network and the amplifier two-port, the same result can frequently be achieved by using feedback.

With shunt–shunt feedback, the input admittance becomes

$$Y_1 = y_{11,\mathrm{Q1}} + Y_\mathrm{f} - \frac{(y_{12,\mathrm{Q1}} - Y_\mathrm{f})(y_{21,\mathrm{Q1}} - Y_\mathrm{f})}{Y_\mathrm{L} + y_{22,\mathrm{Q1}} + Y_\mathrm{f}}. \tag{5.138}$$

Setting $Y_1 = Y_\mathrm{G}^*$ and solving for Y_f, we obtain

$$Y_\mathrm{f} = \frac{(y_{11,\mathrm{Q1}} - Y_\mathrm{G}^*)(Y_\mathrm{L} + y_{22,\mathrm{Q1}}) - y_{21,\mathrm{Q1}}y_{12,\mathrm{Q1}}}{Y_\mathrm{G}^* - Y_\mathrm{L} - (y_{11,\mathrm{Q1}} + y_{12,\mathrm{Q1}} + y_{21,\mathrm{Q1}} + y_{22,\mathrm{Q1}})}. \tag{5.139}$$

Inductive source degeneration

Series–series feedback is also commonly used in matching problems. A practical example is shown in Figure 5.27. An inductor is inserted into the source lead of a FET. This is referred to as *inductive source degeneration*, and may equally be applied to bipolar transistors. We shall now investigate its effect on the input impedance. For simplicity's sake, we assume that for transistor Q_1, $\underline{Y}_{12}$ and $\underline{Y}_{20}$ can be neglected. For a general impedance Z_f in the source lead, the input impedance becomes

$$Z_1 = \frac{v_1}{i_1} = \frac{1}{\underline{Y}_{10}} + Z_\mathrm{f}\left(1 + \frac{g_\mathrm{m}}{\underline{Y}_{10}}\right). \tag{5.140}$$

In the specific case, $Z_\mathrm{f} = \jmath\omega L$. Using the simple FET equivalent circuit in Figure 5.20, further $\underline{Y}_{10} = \jmath\omega C_\mathrm{GS}$. Recall that $\omega_\mathrm{T} = g_\mathrm{m}/C_\mathrm{GS}$, and we find

$$Z_1 = \omega_\mathrm{T} L + \jmath\left(\omega L - \frac{1}{\omega C_\mathrm{GS}}\right). \tag{5.141}$$

The inductance hence creates a real part in the input impedance.

Bandwidth improvement

Both shunt–shunt and series–series feedback can be used to increase bandwidth, however at the expense of maximum gain.

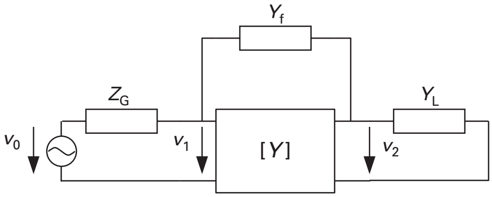


Fig. 5.28 Two-port in y matrix representation with shunt–shunt feedback.

Let us investigate the use of shunt–shunt feedback. First, realise that the influence of the generator impedance needs to be included. Figure 5.28 shows the corresponding schematic. We are interested in the voltage gain between generator and load. For $Y_{\mathrm{f}} = 0$, this is

$$G_{\mathrm{V}} = \frac{v_2}{v_0} = \frac{A_{\mathrm{V}}}{1 + Z_{\mathrm{G}}(y_{11} + A_{\mathrm{V}} y_{12})} \tag{5.142}$$

where A_{V} is given by Equation (5.100). With $Y_{\mathrm{f}} \neq 0$, we obtain

$$A'_{\mathrm{V}} = -A_{\mathrm{V}} \frac{1 - \dfrac{Y_{\mathrm{f}}}{y_{21}}}{1 + \dfrac{Y_{\mathrm{f}}}{y_{22} + Y_{\mathrm{L}}}} \tag{5.143}$$

$$G'_{\mathrm{V}} = \frac{A'_{\mathrm{V}}}{1 + Z_{\mathrm{G}}\left[y_{11} + A'_{\mathrm{V}} y_{12} + Y_{\mathrm{f}}(1 - A'_{\mathrm{V}})\right]}. \tag{5.144}$$

The noteworthy feature here is that Y_{f} appears magnified by $1 - A'_{\mathrm{V}}$ in the denominator. We found this already in a different context when discussing the Miller effect.

The bandwidth enhancement effect can be seen in a simple example. The amplifying device shall be the simple FET from Figure 5.20. Then the y matrix is, with some appropriate simplifications,

$$\begin{aligned}
y_{11} &= \jmath\omega(C_{\mathrm{GS}} + C_{\mathrm{GD}}) \approx \jmath\omega C_{\mathrm{GS}} \\
y_{12} &= -\jmath\omega C_{\mathrm{GD}} \\
y_{21} &= g_{\mathrm{m}} - \jmath\omega C_{\mathrm{GD}} \approx g_{\mathrm{m}} \\
y_{22} &= g_{\mathrm{DS}} + \jmath\omega C_{\mathrm{GD}} \approx g_{\mathrm{DS}}.
\end{aligned}$$

Let us further assume that

$$Y_{\mathrm{f}} \ll g_{\mathrm{m}}.$$

Then,

$$A_{\mathrm{V}} = -\frac{g_{\mathrm{m}}}{Y_{\mathrm{L}} + g_{\mathrm{DS}}},$$

$$A'_{\mathrm{V}} = A_{\mathrm{V}} \frac{Y_{\mathrm{L}} + g_{\mathrm{DS}}}{Y_{\mathrm{L}} + g_{\mathrm{DS}} + Y_{\mathrm{f}}},$$

and

$$G'_{\mathrm{V}} = \frac{A'_{\mathrm{V}}}{1 + Z_{\mathrm{G}}\left[(1 - A'_{\mathrm{V}})Y_{\mathrm{f}} + \jmath\omega(C_{\mathrm{GS}} - A'_{\mathrm{V}}C_{\mathrm{GD}})\right]}.$$

We calculate the 3 dB cutoff frequency as the frequency where the real and the imaginary parts in the denominator are equal. For $Y_{\mathrm{f}} = 0$, this is

$$\omega_{\mathrm{C}}(Y_{\mathrm{f}} = 0) = \frac{1}{Z_{\mathrm{G}}(C_{\mathrm{GS}} - A_{\mathrm{V}}C_{\mathrm{GD}})}.$$

Next, we apply a purely resistive feedback, $Y_{\mathrm{f}} = G_{\mathrm{f}}$:

$$\omega_{\mathrm{C}}(Y_{\mathrm{f}} = G_{\mathrm{f}}) = \frac{1 + Z_{\mathrm{G}}G_{\mathrm{f}}(1 - A'_{\mathrm{V}})}{Z_{\mathrm{G}}(C_{\mathrm{GS}} - A'_{\mathrm{V}}C_{\mathrm{GD}})}.$$

Frequently, $G_{\mathrm{f}} \ll (Y_{\mathrm{L}} + G_{\mathrm{DS}})$ and therefore $A'_{\mathrm{V}} \approx A_{\mathrm{V}}$. The resistive feedback hence results in a very substantial bandwidth enhancement by the factor $1 + Z_{\mathrm{G}}G_{\mathrm{f}}(1 - A_{\mathrm{V}})$. The low-frequency gain, however, decreases to

$$G_{\mathrm{V}}(\omega \to 0) = \frac{A_{\mathrm{V}}}{1 + Z_{\mathrm{G}}R_{\mathrm{f}}(1 - A_{\mathrm{V}})},$$

so that the product $G_{\mathrm{V}}(\omega \to 0)\omega_{\mathrm{C}} = \text{const.}$

Larger bandwidth enhancement is possible if we allow Y_{f} to have a negative imaginary part. This will be treated further down.

Bandwidth enhancement using series–series feedback will be treated for the specific example shown in Figure 5.29. This is a transadmittance stage which converts an input voltage into an output current. The series feedback using Z_{f} will first of all increase the input impedance to lower the load on the preceding stage. If it is made complex, it can be used for significant bandwidth enhancement as well.

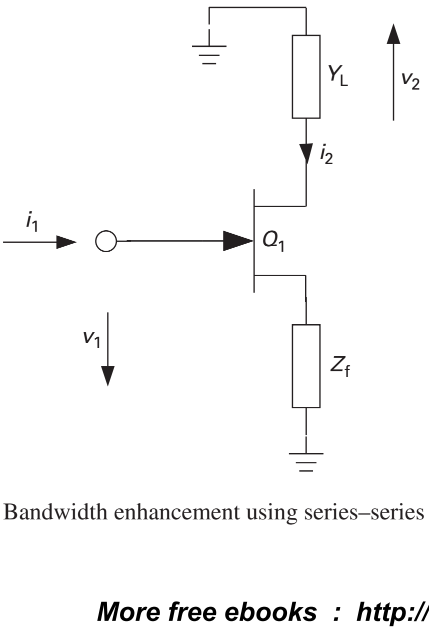


Fig. 5.29 Bandwidth enhancement using series–series feedback.

The transadmittance of this stage is

$$Y_{\mathrm{t}} = \frac{i_2}{v_1} = \frac{g_{\mathrm{m}}}{1 + g_{\mathrm{m}} Z_{\mathrm{f}} \left(1 + \dfrac{\jmath\omega}{\omega_{\mathrm{T}}}\right)}, \tag{5.145}$$

if the FET is described by the simple equivalent circuit in Figure 5.20, and using $\omega_{\mathrm{T}} = g_{\mathrm{m}}/C_{\mathrm{GS}}$.

If now Z_{f} is a parallel RC network,

$$Z_{\mathrm{f}} = \frac{R_{\mathrm{f}}}{1 + \jmath\omega R_{\mathrm{f}} C_{\mathrm{f}}},$$

and C_{f} is chosen,

$$C_{\mathrm{f}} = \frac{1}{\omega_{\mathrm{T}} R_{\mathrm{f}}},$$

the frequency dependence in Y_{T} will disappear:

$$Y_{\mathrm{T}} = \frac{g_{\mathrm{m}}}{1 + g_{\mathrm{m}} R_{\mathrm{f}}},$$

at least for this simple equivalent circuit!

5.4.4 Amplifier configurations with two transistors

In the first part of this chapter, we have seen how the fundamental topologies we can realise with a three-terminal amplifying device have very different properties in terms of input and output admittances, as well as voltage and current gains. Further flexibility in tailoring amplifier properties is achieved when we combine two of the fundamental topologies. We will use generic FETs in order to help visualise the circuits. However, the fundamental concepts apply equally to bipolar transistors – in fact, to any three-terminal amplifying device, as was outlined in the more abstract discussion above.

Common-drain/common-source configuration

Suppose that we want to construct a buffer amplifier, which shall impose a minimal load on a generator, yet also have a significant voltage gain. This can be achieved with the combination of

- a common-drain stage (node 2 as the common node), providing the high input impedance, and
- a common-source stage (node 0 as the common node), providing the voltage gain.

Figure 5.30 shows the schematic of the common-drain/common-source (CD/CS) topology. Q_1 is the common-drain transistor, Q_2 the common-source transistor, Y_{L} is the load admittance and Y_{G} the generator admittance.

The two important aspects of this configuration are input admittance and voltage gain, as discussed.

The voltage gain of the second stage is

$$A_{\mathrm{V,Q2}} = \frac{v_2}{v_{\mathrm{A}}} \approx -\frac{\underline{g}_{\mathrm{m}}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20}} = -\frac{g_{\mathrm{m,Q2}}}{g_{\mathrm{DS,Q2}} + Y_{\mathrm{L}}},$$

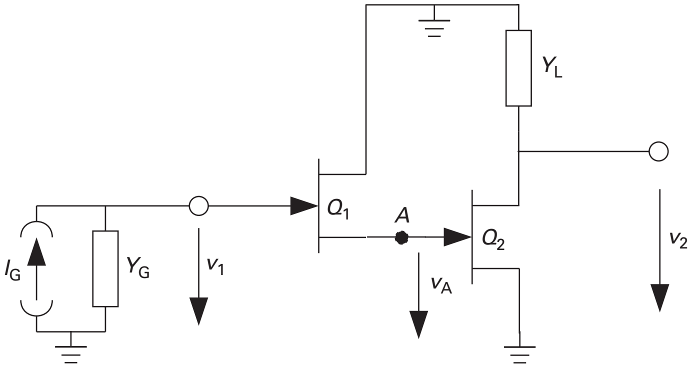


Fig. 5.30 Schematic of a CD/CS amplifier cell. Bias arrangement has been omitted for clarity's sake.

using Equation (5.107) with the simplifying assumptions indicated there, and the FET equivalent circuit in Figure 5.20.

We can now calculate the input admittance of stage two:

$$Y_{1,\mathrm{Q2}} = \underline{Y}_{10,\mathrm{Q2}} + \underline{Y}_{12,\mathrm{Q2}}(1 - A_{\mathrm{V,Q2}}) = \jmath\omega\left[C_{\mathrm{GS,Q2}} + C_{\mathrm{GD,Q2}}(1 - A_{\mathrm{V,Q2}})\right].$$

Because Q_2 is in common-source configuration, we observe the Miller effect, which may significantly increase the capacitance seen from node A into Q_2. To judge the importance of this, consider the output admittance of transistor Q_1, which appears in parallel to $Y_{1,\mathrm{Q2}}$ at node A. We can use Equation (5.130) with the appropriate simplifications:

$$Y_{2,\mathrm{Q1}} \approx \underline{Y}_{20,\mathrm{Q1}} + \underline{g}_{\mathrm{m,Q1}} = g_{\mathrm{DS,Q1}} + g_{\mathrm{m,Q2}} \approx g_{\mathrm{m,Q1}}.$$

The admittance from node A to ground can then be written as

$$Y_{\mathrm{A}} = g_{\mathrm{m,Q1}}\left(1 + \jmath\omega\tau_{\mathrm{A}}\right),$$

where τ_{A} is the characteristic time constant of node A:

$$\tau_{\mathrm{A}} = \frac{C_{\mathrm{GS,Q2}}}{g_{\mathrm{m,Q1}}} + \frac{C_{\mathrm{GD,Q2}}}{g_{\mathrm{m,Q1}}}(1 - A_{\mathrm{V,Q2}}).$$

As long as $(2\pi\tau_{\mathrm{A}})^{-1}$ is significantly outside of the intended frequency range of operation, its effect can be neglected.

The concept of the *characteristic time constant* of internal nodes is very helpful in high-speed design, especially when tracking down reasons for unexpected limitations in performance.

Here, the situation may not be so bad, because the high capacitance seen into Q_2 is compensated for by the high conductance seen into the output of Q_1 in its common-drain configuration.

The voltage gain of stage Q_1 is Equation (5.125):

$$A_{\mathrm{V,Q1}} = \frac{v_{\mathrm{A}}}{v_1} = \left(1 + \frac{Y_{\mathrm{L}} + \underline{Y}_{20,\mathrm{Q1}}}{\underline{Y}_{10,\mathrm{Q1}} + \underline{g}_{\mathrm{m,Q1}}}\right)^{-1} \approx \frac{g_{\mathrm{m,Q1}}}{g_{\mathrm{m,Q1}} + Y_{1,\mathrm{Q1}}} = \frac{1}{1 + \jmath\omega\tau_{\mathrm{A}}}.$$

The total voltage gain is finally

$$A_{\mathrm{V}} = A_{\mathrm{V,Q1}} A_{\mathrm{V,Q2}} \approx \frac{A_{\mathrm{V,Q2}}}{1 + j\omega\tau_{\mathrm{A}}} \approx A_{\mathrm{V,Q2}},$$

if $\omega \ll \tau_{\mathrm{A}}^{-1}$. Again we notice the importance of the characteristic impedance of node A.

The input admittance can be calculated from Equation (5.126):

$$Y_1 \approx \underline{Y}_{12,\mathrm{Q1}} + \frac{\underline{Y}_{10,\mathrm{Q1}}}{1 + \dfrac{g_{\mathrm{m,Q1}}}{Y_{1,\mathrm{Q2}}}} = j\omega C_{\mathrm{GD,Q1}} + \frac{j\omega C_{\mathrm{GS,Q1}}}{1 + \dfrac{g_{\mathrm{m,Q1}}}{Y_{1,\mathrm{Q2}}}}.$$

Recall that in our case, $Y_{1,\mathrm{Q2}}$ is purely capacitive. As shown earlier, a capacitive load to a common-drain stage leads to a negative real part in the input admittance (see Equation (5.128)). Whether this represents a problem for amplifier stability depends on the generator admittance value Y_{G}. This should be kept in mind when investigating stability problems.

Finally, we take a look at the overall current gain. Because the output current of the first stage feeds the input of the second, we expect the overall current gain to be the product of the two individual current gains. However, we have to observe that we always counted currents positive when they flow *into* the device (see Figure 5.18). Then,

$$A_{\mathrm{I}} = -A_{\mathrm{I,Q1}} \cdot A_{\mathrm{I,Q2}}. \tag{5.146}$$

The current gain of the first (common-drain) stage is given by Equation (5.129), observing that now the load admittance is the input admittance of the second stage:

$$A_{\mathrm{I,Q1}} = -\frac{Y_{1,\mathrm{Q2}}(\underline{Y}_{10,\mathrm{Q1}} + \underline{g}_{\mathrm{m,Q1}})}{\underline{Y}_{12,\mathrm{Q1}}(\underline{Y}_{10,\mathrm{Q1}} + \underline{Y}_{20,\mathrm{Q1}} + Y_{1,\mathrm{Q2}} + \underline{g}_{\mathrm{m,Q1}}) + \underline{Y}_{10,\mathrm{Q1}}(Y_{1,\mathrm{Q2}} + \underline{Y}_{20,\mathrm{Q1}})}, \tag{5.147}$$

while the current gain of the common-source stage is given by Equation (5.112):

$$A_{\mathrm{I,Q2}} = \frac{(\underline{g}_{\mathrm{m,Q2}} - \underline{Y}_{12,\mathrm{Q2}})\,\underline{Y}_{\mathrm{L}}}{\underline{Y}_{12,\mathrm{Q2}}(\underline{Y}_{\mathrm{L}} + \underline{g}_{\mathrm{m,Q2}} + \underline{Y}_{10,\mathrm{Q2}} + \underline{Y}_{20,\mathrm{Q2}}) + \underline{Y}_{10,\mathrm{Q2}}(\underline{Y}_{\mathrm{L}} + \underline{Y}_{20,\mathrm{Q2}})}, \tag{5.148}$$

where Y_{L} is the load admittance connected to the drain of Q_2.

Since the explicit calculation of A_{I} presents considerable difficulty, let us make a number of simplifying assumptions. First, we make the two-ports *unilateral*, i.e. we assume $\underline{Y}_{12} = 0$. Then, we assume that the load admittances are always much larger than the elements $\underline{Y}_{20}$ for both transistors: $Y_{1,\mathrm{Q2}} \gg \underline{Y}_{20,\mathrm{Q1}}$, $Y_{\mathrm{L}} \gg \underline{Y}_{20,\mathrm{Q2}}$. Equation (5.146) then has a very simple solution:

$$A_{\mathrm{I}} = \frac{g_{\mathrm{m,Q2}}}{\underline{Y}_{10,\mathrm{Q2}}}\left(1 + \frac{g_{\mathrm{m,Q1}}}{\underline{Y}_{10,\mathrm{Q1}}}\right). \tag{5.149}$$

For further interpretation, turn again to our simple FET equivalent circuit (Figure 5.20), and recall the transit (cutoff) frequency $\omega_{\mathrm{T}} \approx g_{\mathrm{m}}/C_{\mathrm{GS}}$. Equation (5.149) can then be written as

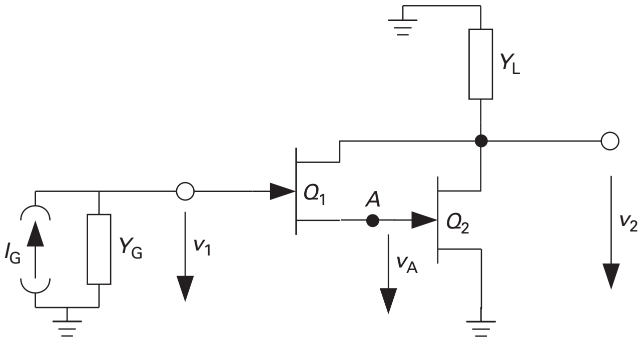


Fig. 5.31 Darlington amplifier configuration.

$$A_I = -\left(\frac{\omega_{T,Q1}\,\omega_{T,Q2}}{\omega^2} + \jmath\frac{\omega_{T,Q2}}{\omega}\right). \tag{5.150}$$

Assume now that $\omega_{T,Q1} = \omega_{T,Q2} = \omega_T$. For $\omega \ll \omega_T$, the current gain is now approximately the product of the current gains of the individual devices, but rolls off at −40 dB/decade, instead of −20 dB/decade. The frequency where $|A_I| = 1$ is $\sqrt{2/(\sqrt{5}-1)}\omega_T = 1.272\,\omega_T$.

Darlington amplifier

The CD/CS configuration is not quite the same as the popular Darlington [7] topology, shown in Figure 5.31. The difference is that in the Darlington amplifier, the drain of Q_1 is connected to the drain of Q_2. While the goal is similar, there are two noteworthy differences:

- The feedback admittance of device Q_1, $\underline{Y}_{12,Q1}$, is now in the path between the output and the input nodes, and not connected directly to ground as in the CD/CS configuration. Therefore, the Miller effect will be present at the input.
- The output current of Q_1 now also flows through the load. This changes the current gain equation. Using the same strong simplifications as in deriving Equation (5.150), we now find

$$A_I = -\left[\frac{\omega_{T,Q1}\,\omega_{T,Q2}}{\omega^2} + \jmath\left(\frac{\omega_{T,Q1}}{\omega} + \frac{\omega_{T,Q2}}{\omega}\right)\right]. \tag{5.151}$$

Compared to the CD/CS amplifier, the Darlington has slightly more short-circuit current gain close to ω_T. If again both transistors are equal and equally biased, the frequency where $|A_I| = 1$ is $2\omega_T$. This is why this configuration is sometimes also called f_T *doubler*. The expression should be taken with a grain of salt. Remember that f_T is derived here from current gain, and that we neglected the feedback admittances in calculating Equation (5.151). The Miller effect disadvantage of the Darlington stage therefore does not show up in the simplification, but can significantly affect circuit performance for small values of Y_L. Further, the current gain rolls off with −40 dB/decade, which may lead to stability problems when feedback is applied around the stage. So your mileage may vary.

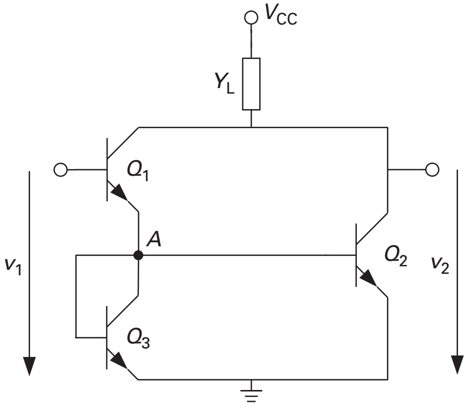


Fig. 5.32 Battjes f_T doubler circuit.

Battjes f_T Doubler

The well-known circuit shown in Figure 5.32, patented by C. R. Battjes [2], is essentially a Darlington amplifier (Q_1, Q_2) combined with a current mirror (Q_2, Q_3), which makes sure that both transistors in the signal path are operated with the same current. If they are also of equal size, they will have the same transit frequency. The circuit shown uses bipolar transistors (as in the patent), but the concept equally works with FETs. Note that the input capacitance of Q_3 needs to be accounted for – if Q_1/Q_3 and Q_2 are supposed to have the same current, then Q_2 and Q_3 need to have the same size, and the effective capacitance attached to node A approximately doubles (neglecting the Miller capacitance seen into Q_2).

Cascode amplifier

The cascode amplifier is a combination of a common-source (or common-emitter) with a common-gate (or common-base) topology. It was conceived as a way to overcome the Miller effect and first described in 1939 using two triode tubes [21], where the cathode of tube 2 was series-connected ('cascaded') to the anode of tube 1. The term *cascode* hence refers to *casc*aded an*ode*. Figure 5.33 shows a cascode realised using FETs.

Let us assess the input admittance first. Because Q_1 is in common-source configuration (node 0 grounded), we use Equation (5.109):

$$Y_1 = \underline{Y}_{10,\mathrm{Q1}} + \underline{Y}_{12,\mathrm{Q1}}\,(1 - A_{\mathrm{V,Q1}}).$$

When calculating $A_{\mathrm{V,Q1}}$, we recognise that the load admittance is the input admittance of Q_2 at node A:

$$A_{\mathrm{V,Q1}} = -\frac{\underline{g}_{\mathrm{m,Q1}} - \underline{Y}_{12,\mathrm{Q1}}}{Y_{1,\mathrm{Q2}} + \underline{Y}_{20,\mathrm{Q1}} + \underline{Y}_{12,\mathrm{Q1}}}.$$

The input admittance $Y_{1,\mathrm{Q2}}$ is calculated using Equation (5.117), because Q_2 is in common-gate configuration:

$$Y_{1,\mathrm{Q2}} = \underline{Y}_{10,\mathrm{Q2}} + (\underline{g}_{\mathrm{m,Q2}} + \underline{Y}_{20,\mathrm{Q2}})\frac{Y_{\mathrm{L}} + \underline{Y}_{12,\mathrm{Q2}}}{\underline{Y}_{\mathrm{L}} + \underline{Y}_{20,\mathrm{Q2}} + \underline{Y}_{12,\mathrm{Q2}}}.$$

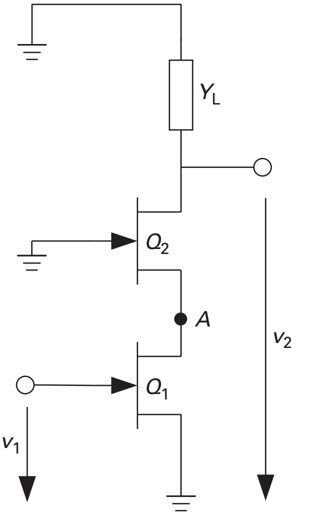


Fig. 5.33 Schematic of a cascode stage built with FETs (bias arrangement omitted).

Assume that $\underline{g}_{m,Q2} \gg \underline{Y}_{20,Q2}$, further $Y_L \gg \underline{Y}_{12,Q2}, \underline{Y}_{20,Q2}$. Then, the input admittance of Q_2 simplifies to

$$Y_{1,Q2} \approx \underline{Y}_{10,Q2} + \underline{g}_{m,Q2}.$$

With this simplification, the voltage gain of Q_1 is then

$$\begin{aligned} A_{V,Q1} &= -\frac{\underline{g}_{m,Q1} - \underline{Y}_{12,Q1}}{\underline{Y}_{10,Q2} + \underline{g}_{m,Q2} + \underline{Y}_{20,Q1} + \underline{Y}_{12,Q1}} \\ &\approx -\frac{\underline{g}_{m,Q1}}{\underline{Y}_{10,Q2} + \underline{g}_{m,Q2}}, \end{aligned}$$

further assuming that $\underline{g}_{m,Q2} \gg (\underline{Y}_{12,Q1} + \underline{Y}_{20,Q1})$.

The input admittance of the cascode finally becomes

$$Y_1 \approx \underline{Y}_{10,Q1} + \underline{Y}_{12,Q1} \left(1 + \frac{\underline{g}_{m,Q1}}{\underline{g}_{m,Q2}} \frac{1}{1 + \dfrac{\underline{Y}_{10,Q2}}{\underline{g}_{m,Q2}}} \right). \tag{5.152}$$

Using our simple FET equivalent circuit (Figure 5.20) this finally becomes

$$Y_1 = j\omega \left[C_{GS,Q1} + C_{GD,Q1} \left(1 + \frac{g_{m,Q1}}{g_{m,Q2}} \frac{1}{1 + j\frac{\omega}{\omega_T}} \right) \right].$$

The suppression of the Miller effect is simply explained by the low voltage gain of the common-source stage $-g_{m,Q1}/g_{m,Q2}$ for low frequencies. Frequently, transistors Q_1 and Q_2 are chosen the same size, and since they share the same drain (or collector) current, it follows that $g_{m,Q1} = g_{m,Q2}$ and $A_{V,Q1} = -1$.

The calculation of the output admittance starts with the output admittance of transistor Q_2, which is in common-gate configuration (node 1 grounded), using Equation (5.118):

$$Y_2 = Y_{2,\mathrm{Q2}} = \underline{Y}_{12,\mathrm{Q2}} + \underline{Y}_{20,\mathrm{Q2}} \frac{Y_{2,\mathrm{Q1}} + \underline{Y}_{10,\mathrm{Q2}}}{Y_{2,\mathrm{Q1}} + \underline{Y}_{10,\mathrm{Q2}} + \underline{Y}_{20,\mathrm{Q2}} + \underline{g}_{\mathrm{m,Q2}}},$$

where $Y_{2,\mathrm{Q1}}$ is the output admittance of transistor Q_1 in common-source configuration (see Equation (5.111)):

$$Y_{2,\mathrm{Q1}} = \underline{Y}_{20,\mathrm{Q1}} + \underline{Y}_{12,\mathrm{Q1}} \left(1 + \frac{\underline{g}_{\mathrm{m,Q1}} - \underline{Y}_{12,\mathrm{Q1}}}{Y_{\mathrm{G}} + \underline{Y}_{10,\mathrm{Q1}} + \underline{Y}_{12,\mathrm{Q1}}}\right),$$

where Y_{G} is the admittance terminating the input port. We simplify the expressions by assuming that the feedback admittances are small and the corresponding terms can be neglected. Then $Y_{2,\mathrm{Q1}} \approx \underline{Y}_{20,\mathrm{Q1}}$ and the overall output conductance becomes

$$Y_2 \approx \underline{Y}_{20,\mathrm{Q2}} \left(1 + \frac{\underline{g}_{\mathrm{m,Q2}} + \underline{Y}_{20,\mathrm{Q2}}}{\underline{Y}_{10,\mathrm{Q2}} + \underline{Y}_{20,\mathrm{Q1}}}\right)^{-1}. \tag{5.153}$$

Because $\underline{g}_{\mathrm{m}}/\underline{Y}_{20}$ is the magnitude of the maximum voltage gain in common-source configuration (see Equation (5.108)), an additional sensible assumption is that $\underline{g}_{\mathrm{m,Q2}} \gg \underline{Y}_{20,\mathrm{Q2}}$. Then,

$$Y_2 \approx \underline{Y}_{20,\mathrm{Q2}} \left(1 + \frac{\underline{g}_{\mathrm{m,Q2}}}{\underline{Y}_{10,\mathrm{Q2}} + \underline{Y}_{20,\mathrm{Q1}}}\right)^{-1}. \tag{5.154}$$

In our simple FET example, we finally find

$$Y_2 \approx g_{\mathrm{DS,Q2}} \left(1 + \frac{g_{\mathrm{m,Q2}}}{g_{\mathrm{DS,Q1}} + \jmath\omega C_{\mathrm{GS,Q2}}}\right)^{-1},$$

and for the quasi-static case, $\omega \to 0$:

$$Y_2 \approx \frac{g_{\mathrm{DS,Q2}}}{1 + \dfrac{g_{\mathrm{m,Q2}}}{g_{\mathrm{DS,Q1}}}}.$$

The output admittance is therefore significantly reduced compared to the common-gate or common-source configurations.

The voltage gain of the cascode stage, $A_{\mathrm{V}} = A_{\mathrm{V,Q1}} \cdot A_{\mathrm{V,Q2}}$, is

$$A_{\mathrm{V}} \approx -\frac{\underline{g}_{\mathrm{m,Q1}}}{Y_{\mathrm{L}} + \underline{Y}_{20,\mathrm{Q2}}} \cdot \frac{\underline{g}_{\mathrm{m,Q2}} + \underline{Y}_{20,\mathrm{Q2}}}{\underline{g}_{\mathrm{m,Q2}} + \underline{Y}_{10,\mathrm{Q2}}}, \tag{5.155}$$

neglecting the feedback admittances. In the FET example,

$$A \approx \frac{g_{\mathrm{m,Q1}}}{Y_{\mathrm{L}} + g_{\mathrm{DS,Q2}}} \cdot \frac{1 + g_{\mathrm{DS,Q2}}/g_{\mathrm{m,Q2}}}{1 + \jmath\omega/\omega_{\mathrm{T}}},$$

or provided that $g_{\mathrm{m,Q2}} \gg g_{\mathrm{DS,Q2}}$ and $\omega \ll \omega_{\mathrm{T}}$:

$$A_{\mathrm{V}} = -\frac{g_{\mathrm{m,Q1}}}{Y_{\mathrm{L}} + g_{\mathrm{DS,Q2}}}.$$

In summary, the cascode configuration provides a comparable voltage gain to the common-source topology, but its input admittance is significantly lower due to the reduction of the Miller capacitance and its output admittance is significantly higher.

Finally, an important side effect of the cascode shall be pointed out here: the real part of the output admittance may become negative. In practical devices, the assumption that parameter $\underline{Y}_{20}$ is purely real is not correct; a better approximation is $\underline{Y}_{20} = g_{\mathrm{DS}} + \jmath\omega C_{\mathrm{DS}}$. If we insert this into Equation (5.153) and separate real and imaginary parts, we find that the real part becomes negative if

$$g_{\mathrm{DS,Q1}} \cdot g_{\mathrm{DS,Q2}} < \omega^2 C_{\mathrm{DS,Q2}} \cdot (C_{\mathrm{DS,Q1}} + C_{\mathrm{GS,Q2}}),$$

assuming that $g_{\mathrm{m,Q2}} \gg g_{\mathrm{DS,Q1}}$ along the way. Frequently, this can lead to amplifier instabilities, but it may also be used to compensate losses in travelling-wave amplifiers, as will be discussed later.

5.4.5 Differential amplifiers

An important component in many high-speed electronic circuits is the differential amplifier. One of the most influential pioneers of biomedical engineering, Otto Schmitt, is frequently held to be the father of the differential amplifier topology [34] – the ability of the differential amplifier to reject common-mode signals at its input is crucial for the measurement of weak bio-electric signals. Incidentally, he also invented the Schmitt trigger circuit.

A generic differential amplifier topology realised with FETs is shown in Figure 5.34. A first noteworthy difference between the amplifiers discussed so far is that the input and output voltages are not referenced to ground, but to the other input and output electrodes, respectively.

Figure 5.35 shows the small-signal representation of the differential amplifier, where the transistors are represented using the generic small-signal equivalent circuit from Figure 5.17. The transistors are identical.

Differential mode

Any combination of nodal input voltages $v_{\mathrm{A}} = v'_{\mathrm{A}} + v''_{\mathrm{A}}$, $v_{\mathrm{B}} = v'_{\mathrm{B}} + v''_{\mathrm{B}}$ can be split into a differential mode ($v'_{\mathrm{A}} = -v'_{\mathrm{B}},\ v''_{\mathrm{A}} = v''_{\mathrm{B}} = 0$) and a common mode ($v''_{\mathrm{A}} = v''_{\mathrm{B}},\ v'_{\mathrm{A}} = v'_{\mathrm{B}} = 0$). First, we concentrate on the differential mode. To calculate the voltage of the common mode v_0, we first loop through $v_{\mathrm{A}},\ v_{1,1},\ v_{1,2}$ and v_{B}:

$$-v'_{\mathrm{A}} + v_{1,1} - v_{1,2} - v'_{\mathrm{A}} = 0 \rightarrow v'_{\mathrm{A}} = \frac{v_{1,1} - v_{1,2}}{2}.$$

On the other hand, $v_0 = v'_{\mathrm{A}} - v_{1,1}$ and therefore

$$v_0 = -\frac{v_{1,1} + v_{1,2}}{2}.$$

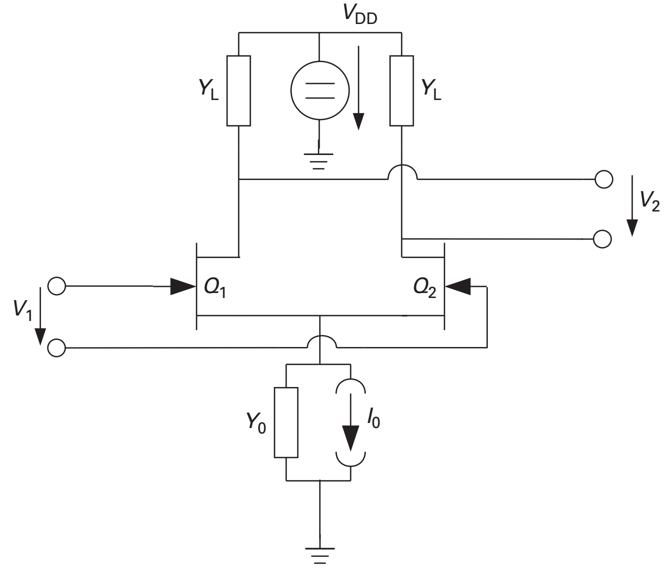


Fig. 5.34 Generic differential amplifier topology.

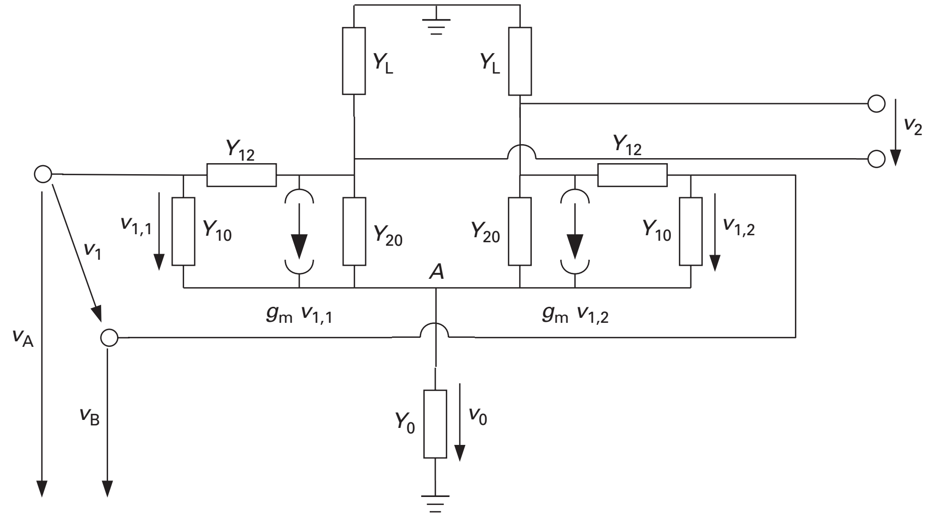


Fig. 5.35 Small-signal representation of the differential amplifier.

For symmetry reasons, $v'_A = -v'_B$ also implies $v_{1,1} = -v_{1,2}$ and hence

$$v_0 = 0$$

in differential mode! The common node *A* constitutes a *virtual ground*. This is a very important concept in high-speed circuit design, as it dramatically reduces problems with common-node impedances, such as in bond wires to ground, which otherwise may lead to a variety of feedback problems.

Now that A is grounded, the two halves of the differential amplifier reduce to standard common-source (or common-emitter) circuits which we have already analysed. The input voltage to the left half is $v_{1,1} = v_1/2$, while the right half receives $v_{1,2} = -v_1/2$.

Using Equation (5.109) to calculate the common-source input admittance $Y_{1,\mathrm{CS}}$ for the individual transistors, the differential mode input admittance is

$$Y_{1,\mathrm{d}} = \frac{i_1}{v_1} = \frac{Y_{1,\mathrm{CS}}}{2}. \tag{5.156}$$

Likewise, the output admittance is half the output admittance $Y_{2,\mathrm{CS}}$ for the common-source stage given by Equation (5.111):

$$Y_{2,\mathrm{d}} = \frac{i_2}{v_2} = \frac{Y_{2,\mathrm{CS}}}{2}. \tag{5.157}$$

Equation (5.107) is used to calculate the common-source voltage gain $A_{\mathrm{V,CS}}$. The differential voltage gain is then

$$A_{\mathrm{V,d}} = \frac{v_2}{v_1} = A_{\mathrm{V,CS}}. \tag{5.158}$$

Common mode

In common mode, both input terminals have the same potential to ground: $v''_{\mathrm{A}} = v''_{\mathrm{B}}$. The individual transistors are connected in parallel then at input and output, resulting in the equivalent circuit shown in Figure 5.36, and their y matrices can simply be added. We arrive at an equivalent transistor Q_{e} with the following y matrix:

$$[y_{\mathrm{Qe}}] = 2\begin{bmatrix} \underline{Y}_{10,1} + \underline{Y}_{12,1} & -\underline{Y}_{12,1} \\ \underline{g}_{\mathrm{m},1} - \underline{Y}_{12,1} & \underline{Y}_{20,1} + \underline{Y}_{12,1} \end{bmatrix}.$$

The parameters are those of the individual transistor. This problem can be treated using the results of the feedback discussion (see p. 323), converting the y matrix first to a z matrix, and adding the z matrix corresponding to Y_0, which is

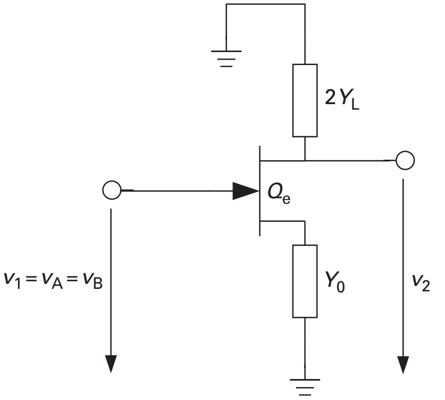


Fig. 5.36 Equivalent circuit of the differential amplifier under common mode excitation.

$$[Z_\mathrm{f}] = \begin{bmatrix} \dfrac{1}{Y_0} & \dfrac{1}{Y_0} \\ \dfrac{1}{Y_0} & \dfrac{1}{Y_0} \end{bmatrix},$$

and converting back to a y matrix.

Here, we will consider a quick solution using a simplified equivalent circuit where we neglect both $\underline{Y}_{12,\mathrm{e}}$ and $\underline{Y}_{20,\mathrm{e}}$ for the transistor Q_e. Let $\underline{g}_{\mathrm{m,e}} = 2g_{\mathrm{m},1}$ and $\underline{Y}_{10,\mathrm{e}} = 2\underline{Y}_{10,1}$, as discussed. The voltage gain $A_{\mathrm{V,cm}}$ for common-mode excitation is then

$$\begin{aligned} A_{\mathrm{V,cm}} = \frac{v_2}{v_1} &= \frac{\dfrac{-2g_{\mathrm{m},1}}{(2Y_\mathrm{L})}}{1 + \dfrac{2}{Y_0 g_\mathrm{m}}} \\ &= -\frac{g_{\mathrm{m},1}}{Y_\mathrm{L}\left(1 + \dfrac{2g_\mathrm{m}}{Y_0}\right)} \\ &\approx -\frac{Y_0}{2Y_\mathrm{L}}, \end{aligned} \tag{5.159}$$

assuming $g_{\mathrm{m},1}/Y_0 \gg 1$. The output voltage here is taken between one of the output terminals and ground – the differential output voltage for common-mode excitation is 0, provided that the circuit is perfectly symmetrical. The voltage gain under differential excitation, but with the output voltage taken between one of the output terminals and ground, is

$$A_{\mathrm{V,d,gnd}} = -\frac{g_{\mathrm{m},1}}{2Y_\mathrm{L}}.$$

The ratio

$$\left|\frac{A_{\mathrm{V,d,gnd}}}{A_{\mathrm{V,cm}}}\right| = \left|\frac{g_{\mathrm{m},1}}{Y_0}\right| \tag{5.160}$$

is the *common-mode rejection ratio*, a measure for the suppression of common-mode input signals. We see that Y_0 should be as small as possible.

Neutralisation of differential amplifiers

The fact that the output voltages of a differential amplifier are exactly 180° out of phase can be used to elegantly eliminate the effect of the feedback capacitances in the transistors. This is shown in Figure 5.37. The capacitances C_n, which have to be exactly equal to C_{GD}, feed a current into the gate nodes of the two transistors, which is equal in magnitude, but of opposite sign, compared to the currents flowing through C_{GD}, cancelling these capacitances.

A more complex differential amplifier example

Differential amplifiers for high-speed applications are frequently more complex and exploit the special properties discussed in the section on basic amplifier topologies using two transistors. Figure 5.38 shows a cell common to many high-speed differential amplifiers. Transistors Q_1 and Q_2 are in common-collector configuration, connected to the transistor pairs Q_3, Q_5 and Q_4, Q_6, respectively, which form a differential cascode.

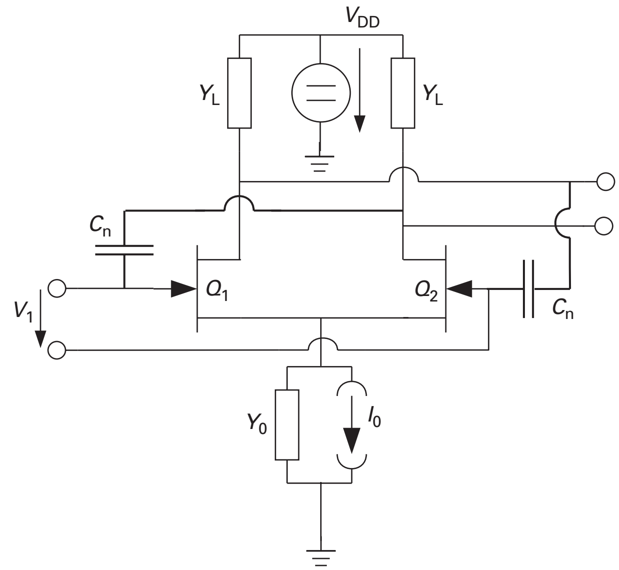


Fig. 5.37 Differential amplifier with neutralisation.

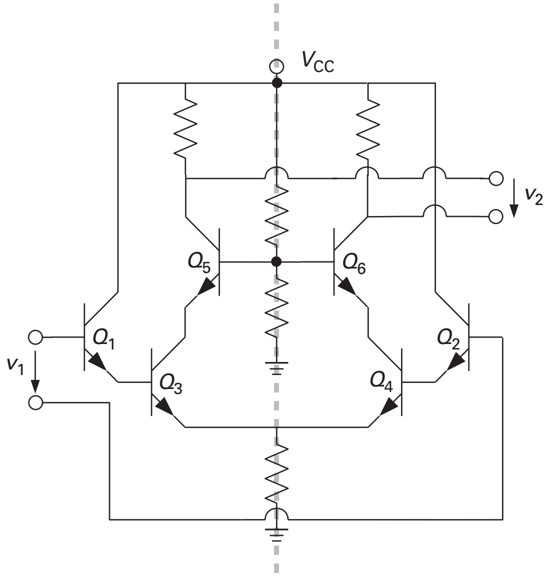


Fig. 5.38 A more complex differential amplifier example. The dashed line indicates the symmetry plane; all nodes along this plane are virtual grounds.

All nodes along the median, which is indicated as a dashed line, are virtual grounds, provided that the circuit is driven fully differentially. This is particularly interesting for the bases of Q_5 and Q_6, because proper grounding of the base terminal can be a problem in cascode stages – here, it is easy due to the virtual ground property. Equally, the emitters of Q_3 and Q_4 are properly grounded. The DC bias voltage terminal, V_{CC}, is also an RF ground, facilitating RF/DC decoupling.

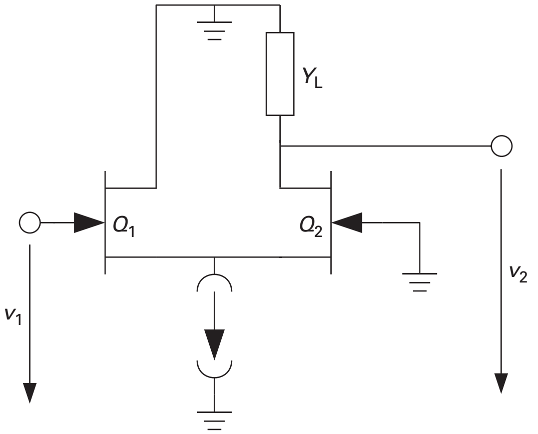


Fig. 5.39 Source-coupled amplifier schematic (bias elements omitted).

These advantages lead to an increasing use of differential topologies in micro- and millimetre-wave circuits. Drawbacks are the increased power consumption due to the doubled component count and the increased area consumption. Another problem may be on-wafer testing, due to the necessity for differential probes.

5.4.6 Source-coupled amplifier

The amplifier topology shown in Figure 5.39 has, at first glance, a configuration very similar to the differential amplifier. We immediately recognise the source-coupled pair and the common-current source. However, the amplifier is driven single-endedly and also has only a single output. Upon closer investigation, Q_1 is in common-drain and Q_2 in common-gate topology.

The idea is therefore very similar to the CD/CS amplifier discussed earlier. The common-drain input transistor creates a low input admittance, while the common-gate stage delivers the voltage gain. The Miller effect is eliminated, and the input is well isolated from the output.

Compared to the CD/CS amplifier, the input admittance is higher, because the input admittance of the common-gate transistor Q_2 is much higher than that of a comparably biased common-source transistor: $Y_{1,\mathrm{Q2}} \approx \underline{Y}_{10,\mathrm{Q2}} + \underline{g}_{\mathrm{m}}$; see Equation (5.117) with $Y_{\mathrm{L}} \gg \underline{Y}_{20,\mathrm{Q2}}$. The input admittance of the source-coupled amplifier is then

$$Y_1 = \underline{Y}_{12,\mathrm{Q1}} + \frac{\underline{Y}_{10,\mathrm{Q1}}}{1 + \dfrac{\underline{g}_{\mathrm{m,Q2}} + \underline{Y}_{10,\mathrm{Q2}}}{\underline{g}_{\mathrm{m,Q1}} + \underline{Y}_{10,\mathrm{Q1}}}}. \tag{5.161}$$

Using our simple FET equivalent circuit, $\underline{Y}_{10} = \jmath\omega C_{\mathrm{GS}}$ and

$$\underline{g}_{\mathrm{m}} + \underline{Y}_{10} = g_{\mathrm{m}}\left(1 + \jmath\frac{\omega}{\omega_{\mathrm{T}}}\right).$$

If the transistors therefore have the same ω_{T}, the input admittance is

$$Y_1 = \jmath\omega\left(C_{\mathrm{GD,Q1}} + \frac{C_{\mathrm{GS,Q1}}}{1 + \dfrac{g_{\mathrm{m,Q2}}}{g_{\mathrm{m,Q1}}}}\right).$$

It is purely capacitive and does not show the risk of a negative real part, which the CD/CS amplifier had posed.

The circuit can also be compared to the cascode – the source-coupled amplifier has a lower input admittance, is non-inverting and requires a lower supply voltage than the cascode, but the cascode requires less current, because the current through the common-gate stage is recycled in the common-source transistor.

5.4.7 Tuned amplifiers

Tuned amplifiers are commonly used at micro- and millimetre-wave frequencies when the fractional bandwidth is small. The fractional bandwidth is the required operational bandwidth divided by the centre frequency. For example, the 24 GHz license-free ISM band has a total allowed spectral width of 250 MHz, so any amplifier will need sensibly only a fractional bandwidth of 10^{-2}. Other applications, such as emerging ultra-wideband sensor and communications standards, will have fractional bandwidths which are orders of magnitude larger – the design of amplifiers for such systems will be treated in the next section (p. 350).

A typical tuned amplifier will use three fundamental circuit techniques:

(i) A resonant load – the load admittance goes through a minimum at the frequency of operation, maximising the voltage gain for a given transconductance.
(ii) Complex conjugate match at the input, ensuring that the available power from the source is delivered to the amplifier.
(iii) Complex conjugate match at the output, ensuring that the available power from the amplifier is delivered to the load.

For LNAs and power amplifiers, other matching strategies may apply for the input and output ports, respectively. These will be treated in the sections on LNA design (p. 365) and power amplifier design (p. 376). For now, we assume that achieving the maximum gain is our objective.

Resonant loads

Let us first investigate the resonant load, using the simple example of Figure 5.40. The schematic also indicates the generator and the equivalent input admittance of the following stage – it is essential to include at least the next-stage input admittance in the calculations, and due to feedback, the generator admittance will also have an effect, albeit more weakly.

The admittances $Y_{2,\mathrm{Q1}}$ and $Y_{1,\mathrm{Q2}}$ can typically be represented by a conductance in parallel with a capacitive reactance (exception – if the following stage is a common-gate

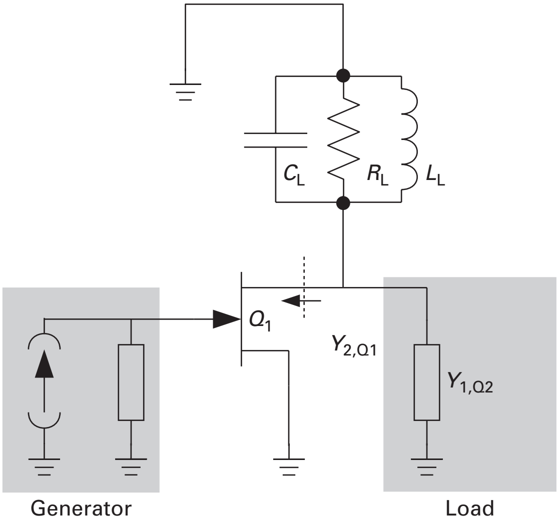


Fig. 5.40 Tuned amplifier stage with resonant load.

or common-base stage, the reactance may be inductive). These elements are absorbed into the load. The resulting reactances of the tank circuit are then

$$G_T = R_L^{-1} + \mathrm{Re}(Y_{2,Q1} + Y_{1,Q2})$$
$$C_T = C_L + \frac{1}{\omega}\mathrm{Im}(Y_{2,Q1} + Y_{1,Q2})$$
$$L_T = L_L.$$

The transfer function of the voltage gain is

$$A_V(\omega) = -\frac{g_m}{G_T}\frac{1}{1 + \jmath\left(\omega\frac{C_T}{G_T} - \frac{1}{\omega L_T G_T}\right)}, \tag{5.162}$$

which has its maximum at

$$\omega_0 = \frac{1}{\sqrt{L_T C_T}},$$

and its −3 dB corner points at

$$\omega_{\frac{1}{2}} = \sqrt{\frac{G_T^2}{4C_T^2} + \frac{1}{L_T C_T}} \pm \frac{G_T}{2C_T}.$$

The bandwidth between the −3 dB points is therefore

$$\Delta\omega = \omega_1 - \omega_2 = \frac{G_T}{C_T}. \tag{5.163}$$

This equation can be used to choose the proper G_T for the required bandwidth of the amplifier.

Using Equations (5.162) and (5.163), we find the product of the voltage gain at $\omega = \omega_0$ and the -3 dB bandwidth:

$$-\,A_{\mathrm{V}}(\omega_0)\cdot\Delta\omega = \frac{g_{\mathrm{m}}}{C_{\mathrm{T}}}, \tag{5.164}$$

which is interestingly independent of frequency. This is due to simplifying assumptions, of course. In the *ansatz* for Equation (5.162) we used Equation (5.107) with the assumption that $\underline{g}_{\mathrm{m}} \gg \underline{Y}_{12}$, hence that feedback is negligible, which is no longer true at very high frequencies.

A tuned tank circuit always bears the risk of amplifier instability. For the common-source amplifier in the example, we use Equation (5.109) to calculate the input admittance of the circuit, using the expression in Equation (5.162) for $A_{\mathrm{V}}(\omega)$:

$$\begin{aligned} Y_1 &= \underline{Y}_{10} + \underline{Y}_{12}\left\{1 + \frac{\underline{g}_{\mathrm{m}}}{G_{\mathrm{T}}}\,\frac{1}{1 + \jmath\left[\omega\frac{C_{\mathrm{T}}}{G_{\mathrm{T}}} - \frac{1}{L_{\mathrm{T}}G_{\mathrm{T}}}\right]}\right\} \\ &= \underline{Y}_{10} + \underline{Y}_{12} + \underline{Y}_{12}\frac{\underline{g}_{\mathrm{m}}}{G_{\mathrm{T}}}\,\frac{1 - \jmath\left[\omega\frac{C_{\mathrm{T}}}{G_{\mathrm{T}}} - \frac{1}{\omega L_{\mathrm{T}}G_{\mathrm{T}}}\right]}{1 + \left[\omega\frac{C_{\mathrm{T}}}{G_{\mathrm{T}}} - \frac{1}{\omega L_{\mathrm{T}}G_{\mathrm{T}}}\right]^2}. \end{aligned}$$

If, as is usually the case, $\underline{Y}_{12} \approx \jmath\omega C_{\mathrm{GD}}$, the third term in the sum has a negative real part for $\omega < \omega_0$. The risk of parasitic oscillations increases with increasing peak gain. Neutralisation measures as discussed already (p. 325) may become necessary in such cases.

Input and output matching networks

A common requirement in microwave amplifiers is that input and output admittances need to have a predefined value. There are two major reasons for this:

(i) If the input and output admittances are the complex conjugates of the source and load admittances, the source's available power is transferred to the amplifier, and the amplifier's available power is transferred to the load, resulting in the maximum power gain – this value is called the *maximum available gain* and will be discussed shortly.

(ii) To avoid standing waves on interconnecting transmission lines, the lines need to be terminated by their characteristic impedances at least at one end.

The characteristic impedance of a lossless transmission line is real; hence, input and output admittances are normally tuned to a purely real value where the circuit will interface with a transmission line. For internal nodes, however, this is not necessary – in fact, as we will see in our discussion of broadband amplifier techniques, at internal nodes impedance matching is often abandoned altogether, in favour of increased bandwidth, but with penalties in power gain.

Because the Smith chart (see p. 295) is the most important tool in solving matching problems, we will conduct the matching discussions using scattering parameters.

Most importantly, we need to translate source and load admittances as well as two-port input and output admittances into reflection coefficients. This is easily done:

$$\Gamma = \frac{Y_0 - Y}{Y_0 + Y} = \frac{Z - Z_0}{Z + Z_0}, \tag{5.165}$$

where $Y_0 = 1/Z_0$ is the normalising admittance, which is frequently 20 mS (correspondingly $Z_0 = 50\,\Omega$), but can be chosen arbitrarily.

We have seen that in two-ports which are not unilateral ($y_{12} \neq 0$, correspondingly $S_{12} \neq 0$), the input admittance depends on the load admittance, and the output admittance depends on the source admittance. In general terms and with the two-port expressed as a scattering matrix, the input (Γ_1) and output (Γ_2) reflection coefficients are (p. 300)

$$\Gamma_1 = S_{11} + \frac{S_{12}S_{21}\Gamma_{\mathrm{L}}}{1 - S_{22}\Gamma_{\mathrm{L}}} \tag{5.166}$$

$$\Gamma_2 = S_{22} + \frac{S_{12}s_{21}\Gamma_{\mathrm{G}}}{1 - S_{11}\Gamma_{\mathrm{G}}}, \tag{5.167}$$

where Γ_{G} and Γ_{L} are the generator and load reflection coefficients, respectively.

For simultaneous power match at input and output ports, we need these coupled equations to be satisfied:

$$\Gamma_{\mathrm{G}}^* = S_{11} + \frac{S_{12}S_{21}\Gamma_{\mathrm{L}}}{1 - S_{22}\Gamma_{\mathrm{L}}}$$

$$\Gamma_{\mathrm{L}}^* = S_{22} + \frac{S_{12}S_{21}\Gamma_{\mathrm{G}}}{1 - S_{11}\Gamma_{\mathrm{G}}},$$

where Γ^* is the complex conjugate of Γ. Solving these equations for the necessary generator and load reflection coefficients $\Gamma_{\mathrm{G,m}}$ and $\Gamma_{\mathrm{L,m}}$, we find

$$\Gamma_{\mathrm{G,m}} = \frac{C_1^*}{|C_1|}\left(\frac{B_1}{2|C_1|} - \sqrt{\frac{B_1^2}{|C_1|^2} - 1}\right), \tag{5.168}$$

with

$$B_1 = 1 - |S_{22}|^2 + |S_{11}|^2 - |\Delta(S)|^2$$

$$C_1 = S_{11} - \Delta(S)\,S_{22}^*,$$

where $\Delta(S)$ is the determinant of the scattering matrix. For the load reflection coefficients, we find likewise:

$$\Gamma_{\mathrm{L,m}} = \frac{C_2^*}{|C_2|}\left(\frac{B_2}{2|C_2|} - \sqrt{\frac{B_2^2}{|C_2|^2} - 1}\right), \tag{5.169}$$

with

$$B_2 = 1 - |S_{11}|^2 + |S_{22}|^2 - |\Delta(S)|^2$$

$$C_2 = S_{22} - \Delta(S)\,S_{11}^*.$$

Simultaneous input and output power match is not always possible, but requires a two-port to be *unconditionally stable* (see p. 303).

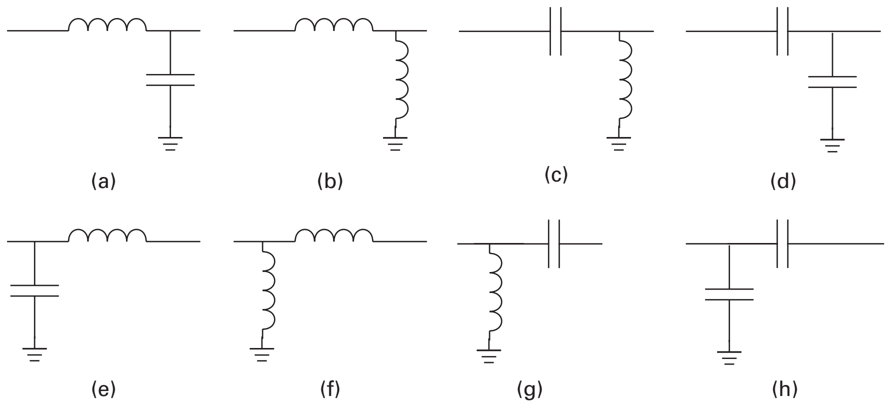


Fig. 5.41 Generic L network topologies.

Generally speaking, impedance transformation can be achieved using

- reactances L, C,
- transformers,
- transmission line impedance transformation.

At micro- and millimetre-wave frequencies, 'true' transformers based on coils are rarely used, because when realised on-chip using planar inductors, they tend to be very lossy, and additionally have high parasitic capacitances. So only impedance transformations using reactive networks and transmission line impedance transformation will be discussed here.

The most fundamental impedance transforming network is the *L network*, which can have any of the shapes shown in Figure 5.41.

There is always more than one topology which achieves the desired impedance transformation. This is an important observation, because other considerations need to be taken into account also. For example, the input port may have to be DC-blocked, in which case a topology with a series C may be suitable(cases c, d, g, or h in Figure 5.41), or DC bias may have to be supplied through the port, in which case a series L and no shunt L are needed (cases a or e). Likewise, it may be advantageous to ground the input port at low frequencies, favouring a topology with a shunt L (cases b, c, or f).

Figure 5.42 shows an example of an impedance matching problem, solved using several topologies. In all cases, the impedance in the lower left quadrant is the starting point and the centre of the Smith chart is the target.

- Path 1 uses an L in series with the start impedance and then a shunt L.
- Path 2 also starts with a series L, but a larger one, and then uses a shunt C.
- Path 3 starts with a shunt L, and then continues with a series L.
- Path 4 starts equally with a shunt L, but a smaller one, and then uses a series C to reach the required impedance.

The several options are best visualised using the Smith chart.

Table 5.1 Matching a complex load (100 Ω parallel to 2.5 pF) to 50 Ω using different L network topologies, $f = 1$ GHz

Path	Components	Bandwidth
1	$L_{S,1}$=3.1 nH, $L_{P,2}$=8.8 nH	508 MHz
2	$L_{S,1}$=11 nH, $C_{P,2}$=2.75 pF	547 MHz
3	$L_{P,1}$=26 nH, $L_{S,2}$=7.9 nH	561 MHz
4	$L_{P,1}$=6.1 nH, $C_{S,2}$=3.15 nH	324 MHz

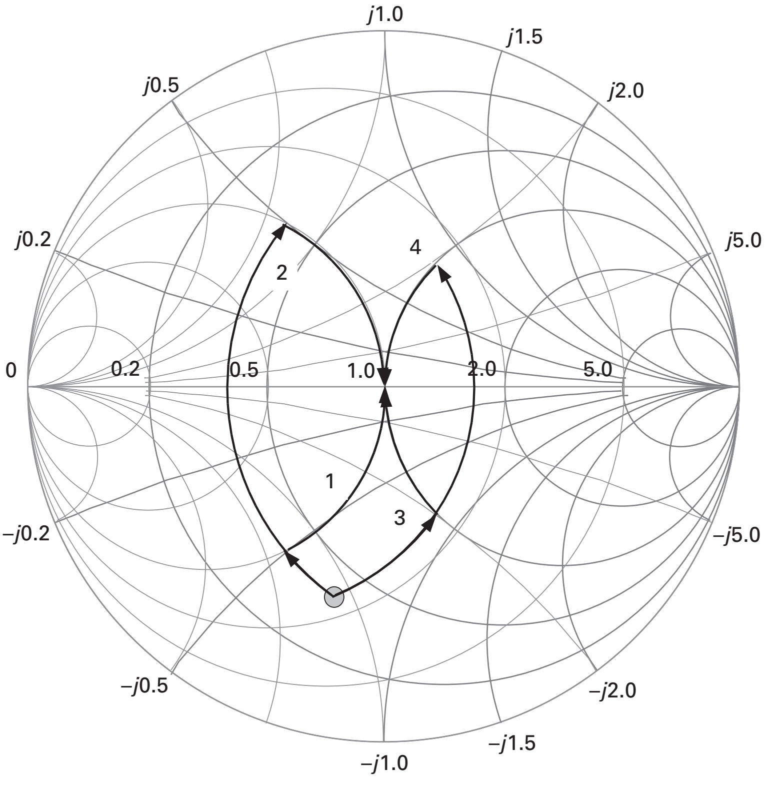


Fig. 5.42 Example for multiple impedance transformation paths using L networks.

Other aspects of matching networks need to be considered as well. This shall be done in an example, where a parallel RC load ($R = 100\,\Omega$, $C = 2.5$ pF) is matched to $Z_0 = 50\,\Omega$ using the different topologies in Figure 5.42. The results are shown in Table 5.1.

First of all, we note that the matching bandwidth, defined as the difference between the frequencies where the reflection coefficient becomes $|\Gamma| > 0.32$ (return loss less than 10 dB), is vastly different – path 4 has less than 60% of the bandwidth of the others. Also, component values may become impractically larger for on-chip implementation – for example, $L_{P,1}$ for path 3.

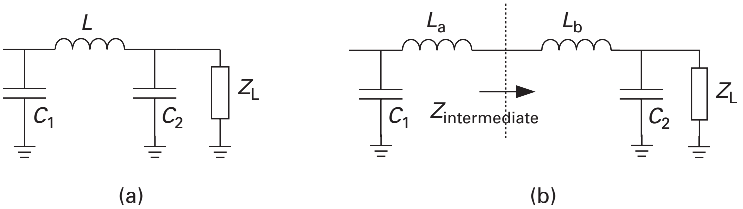


Fig. 5.43 Example of (a) a π-type matching network and (b) its decomposition into two cascaded L networks.

These calculations have been performed using ideal components. In practice, large-value spiral inductors also come with considerable series resistances, which is another aspect to consider.

π networks are an extension of L networks – they are best thought of as being separated into two L networks, as shown in Figure 5.43. The first L network transforms to an intermediate impedance $Z_{\text{intermediate}}$, which is then transformed by the second L network to the desired value. π networks offer an additional degree of freedom, so we can additionally design for different matching bandwidths. They are additionally attractive, because they allow the absorption of interconnect parasitics into the matching network – e.g. bond pad parasitics on chip and in the package (or on the PCB board) can form part of C_1 and C_2, while the bond wire inductance can be absorbed into L.

Other combinations of L-type networks exist and can be useful for specific matching problems, but this is beyond the scope of this book.

Figure 5.44 shows three examples of compact tuned amplifiers in an 80 GHz f_T Si/SiGe HBT technology [6]. The amplifiers share the same basic topology – three cascaded cascode stages with resonant loads and LC interstage matching using spiral inductors. Additionally, inductive emitter degeneration (Equation (5.141)) is used to assist the match by increasing the real part of the input impedance. The use of concentrated reactances, even at millimetre-wave frequencies, leads to an extremely compact layout.

Transmission line segments can also be used to transform impedances. Assuming lossless transmission lines, the input impedance looking into a transmission line of length l and characteristic impedance Z_0, terminated by an impedance Z_L, is

$$Z_1 = Z_0 \frac{Z_L + \jmath Z_0 \tan\left(2\pi \frac{l}{\lambda'}\right)}{Z_0 + \jmath Z_L \tan\left(2\pi \frac{l}{\lambda'}\right)}, \tag{5.170}$$

where λ' is the wavelength on the transmission line,

$$\lambda' = \frac{c_0}{f\sqrt{\epsilon_{\text{r,eff}}}}.$$

A very popular example is the *quarter-wavelength transformer*. In case $l = \lambda'/4$, the input impedance becomes

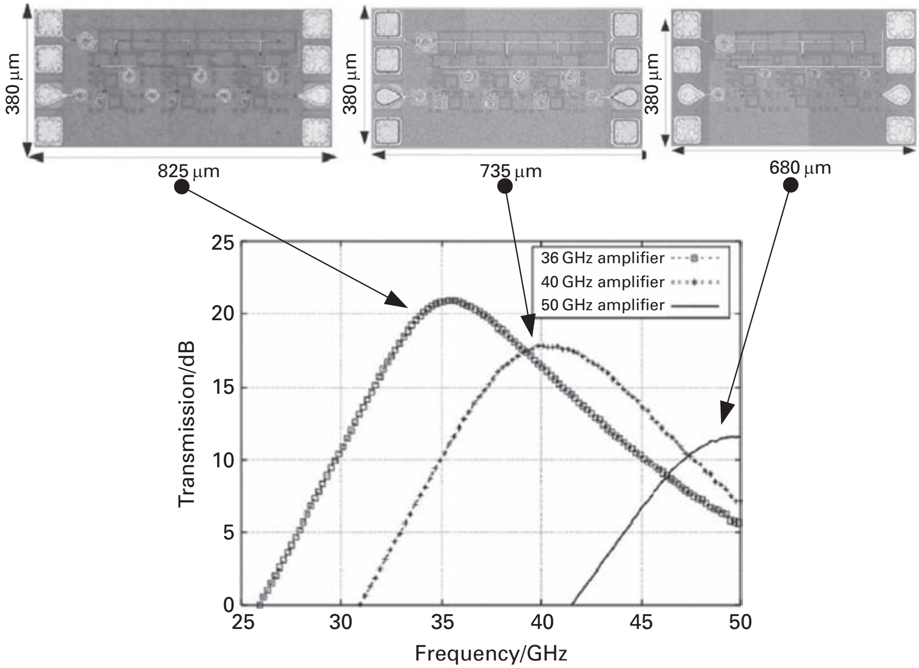


Fig. 5.44 Tuned millimetre-wave amplifiers in a Si/SiGe HBT technology, using LC loads and matching networks. (After [6])

$$Z_1 = \frac{Z_0^2}{Z_L}. \tag{5.171}$$

In other words, to match two impedances Z_A, Z_B, they need to be connected with a transmission line which is $\lambda'/4$ long and has a characteristic impedance of $Z_0 = \sqrt{Z_A Z_B}$. Quarter-wave transmission line sections are also called *impedance inverters* – the reason is obvious from Equation (5.171).

Transmission lines open up additional possibilities in matching. This is shown in Figure 5.45, again using the same start impedance as above:

- In path 1, a transmission line section of impedance $Z_0 = 50\,\Omega$ is used first to make the impedance real. The intermediate impedance is 14.5 Ω; hence the quarter-wave section must have an impedance of $\sqrt{50 \cdot 14.5} = 26.9\,\Omega$.
- Path 2 first uses a series inductance to make the impedance real, the intermediate impedance is 28.7 Ω. The quarter-wave section then needs to have a characteristic impedance of 37.8 Ω.
- Path 3, finally, uses a shunt inductance to make the impedance real ($Z_{\text{intermediate}} = 100\,\Omega$) and a quarter-wave section with $Z_0 = 70.7\,\Omega$.

Option 2 has the widest matching bandwidth, but the transmission line in option 3 is likely the easiest to realise.

With increasing frequency, tuned amplifiers using transmission line segments become increasingly interesting, because spiral inductors are especially difficult to realise and

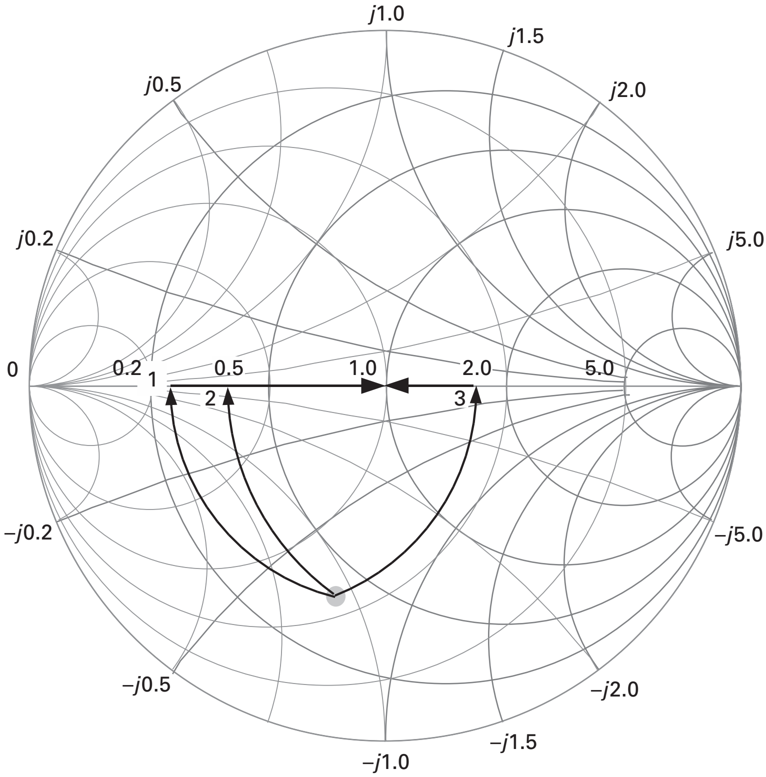


Fig. 5.45 Matching examples using quarter-wave transmission line transformers.

model, and the main objection against the use of transmission lines – their physical size in layout – becomes irrelevant as the wavelength shrinks. Figure 5.46 shows an example. The IC represents a three-stage fully differential amplifier for 77 GHz automotive RADAR systems, realised in a 190 GHz f_T Si/SiGe BiCMOS technology [5]. The amplifier provides 16 dB gain while consuming 90 mW from a 3 V supply. Thin-film microstrip[3] lines (TFMSLs) are used here for impedance matching purposes. Due to the high frequency, the resulting IC is still very compact (740 × 540 μm^2 chip size).

5.4.8 Broadband amplifier techniques

Tuned loads and reactive impedance matching networks are not suitable for amplifiers with large fractional bandwidths, such as those used in high-speed fibre-optic systems, micro/millimetre-wave instruments, many military systems with high frequency ability, or impulse-radio ultra-wideband systems. All of these applications need amplifiers where the gain must be flat over a wide frequency range (often the ratio of upper to

[3] In TFMSLs, the ground plane is realised on top of the substrate. This shields the signal line from the lossy Si substrate, but leads to very narrow signal lines.

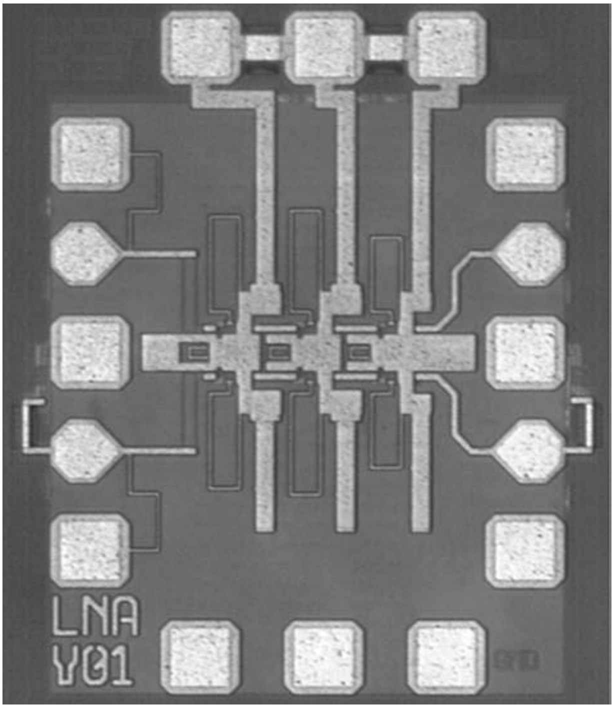


Fig. 5.46 Fully differential Si/SiGe HBT amplifier for 77 GHz, using tuned transmission lines. (After [5])

lower cutoff frequency exceeds the factor of two – *multi-octave* bandwidths), and almost always the input and output return loss also needs to stay below a specified value over the full frequency range.

In the following section, we will discuss some common techniques which prove useful in the realisation of amplifiers with very large bandwidths using concentrated circuit components. Discussion of distributed amplification, which is also a very important concept for wideband amplifiers, will start on p. 354.

Shunt peaking

We have already emphasised the importance of the characteristic time constant in the discussion of multi-stage amplifier topologies (p. 330). We will see that broadband amplifier design always comes down to modifying these internal characteristic time constants.

Consider the simple cascading of common-source amplifiers, shown in Figure 5.47 together with a strongly simplified equivalent circuit. The load resistance and the input capacitance of the following stage are combined into an equivalent impedance to ground Z_eq. Using Z_eq, the transadmittance of the cascaded stage can be expressed as

$$Y_\mathrm{T} = \frac{i_2}{v_1} = -g_{\mathrm{m},1}g_{\mathrm{m},2}Z_\mathrm{eq} = -\frac{R_\mathrm{L}g_{\mathrm{m},1}g_{\mathrm{m},2}}{1+\jmath\omega R_\mathrm{L}C_{\mathrm{GS},2}}. \tag{5.172}$$

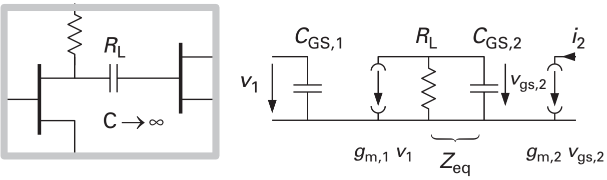


Fig. 5.47 Intermediate node of two cascaded common-source amplifiers, with small-signal equivalent circuit.

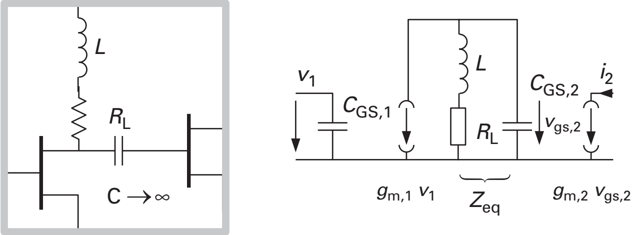


Fig. 5.48 Cascade connection of two common-source amplifiers with shunt peaking inductor.

Obviously, $R_L C_{GS,2}$ is the characteristic time constant of the intermediate node, which limits the bandwidth to

$$\omega_1 = \frac{1}{R_L C_{GS,2}} = \frac{\omega_{T,2}}{g_{m,2} R_L}, \tag{5.173}$$

using $\omega_T = g_m / C_{GS}$.

We will now partially compensate the capacitive reactance by connecting an inductor in series with the load resistor (see Figure 5.48). The transadmittance now becomes

$$\begin{aligned} Y_T &= -g_{m,1} g_{m,2} Z_{eq} \\ &= -g_{m,1} g_{m,2} R_L \frac{1 + \dfrac{\jmath\omega L}{R_L}}{1 - \omega^2 L C_{gs,2} + \jmath\omega R_L C_{gs,2}}. \end{aligned} \tag{5.174}$$

Introducing

$$\tau = \frac{L}{R_L};\; m = \frac{R_L^2 C_{gs,2}}{L} = \frac{1}{\omega_1 \tau},$$

we rewrite Equation (5.174) [25]:

$$Y_T = -g_{m,1} g_{m,2} R_L \frac{1 + \jmath\left(\dfrac{\omega}{\omega_1}\right) m^{-1}}{1 - \left(\dfrac{\omega}{\omega_1}\right)^2 m^{-1} + \dfrac{\jmath\omega}{\omega_1}}. \tag{5.175}$$

The new −3 dB cutoff frequency is

$$\omega_2 = \omega_1 \sqrt{\left(\frac{-m^2}{2} + m + 1\right) + \sqrt{\left(\frac{-m^2}{2} + m + 1\right)^2 + m^2}}. \tag{5.176}$$

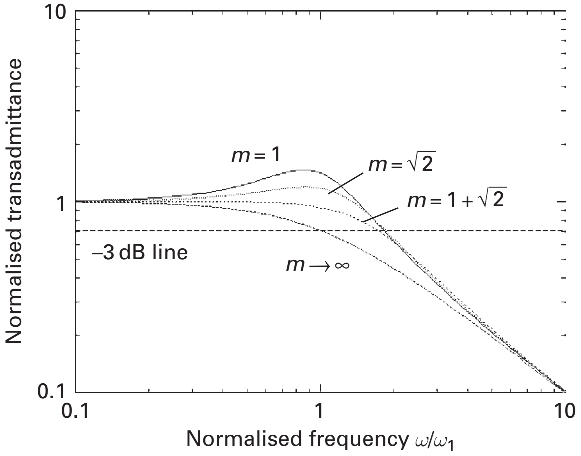


Fig. 5.49 Normalised transadmittance of an amplifier cascade with shunt peaking versus frequency, for different values of parameter m.

Equation (5.176) is maximum for

$$m = \sqrt{2},$$

or finally

$$\tau = \frac{1}{\sqrt{2}\omega_1}. \tag{5.177}$$

Figure 5.49 plots the normalised transadmittance $Y_T/\left(g_{m,1}g_{m,2}R_L\right)$ versus the normalised frequency, and for several values of m. We note that

- we can achieve 1.8-fold increase in bandwidth;
- the increase in bandwidth comes at the expense of gain flatness;
- however, for $m = 1 + \sqrt{2}$, the response becomes *maximally flat* with only a marginal decrease in bandwidth.

Feedback techniques

We had already seen (Figure 5.29) that a parallel RC combination in series–series feedback can be used to completely eliminate the dominant pole in the frequency response of the transadmittance. Let us consider a somewhat more complicated example now where the amplifier is loaded by a complex load formed by a resistor and a capacitor in parallel – the typical equivalent circuit of a following amplification stage. The small-signal equivalent circuit is shown in Figure 5.50. The voltage gain is

$$A_V = -g_m R_L \frac{1 + \jmath\omega\tau_S}{(1 + \jmath\omega\tau_L)\left[1 + \frac{g_m}{G_S} + \jmath\omega\left(\tau_S + \frac{g_m}{G_S}\frac{\omega}{\omega_T}\right)\right]}, \tag{5.178}$$

where

$$\tau_S = R_S C_S;\ \tau_L = R_L C_L;\ \omega_T = \frac{g_m}{C_{GS}}.$$

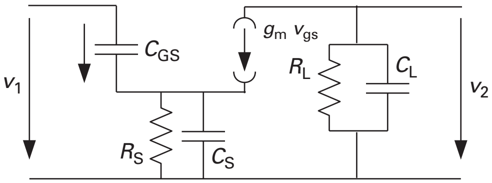


Fig. 5.50 Bandwidth enhancement using series–series feedback.

The enumerator term can now be used to cancel one of the denominator poles:

- If $1 + \jmath\omega\tau_L$ dominates, then $\tau_S = \tau_L$ is the proper choice.
- If the second term dominates, then choose $\tau_S = \omega_T^{-1}$. This corresponds to the solution already discussed in Equation (5.145).

5.4.9 Distributed amplification

The amplifier topologies discussed so far employed concentrated circuit elements and are as such not very different from topologies employed at lower frequencies. The distributed nature of components, especially interconnect lines, only comes in at the layout stage. In the wideband amplifier technique we will discuss now, the transmission line nature is consciously used to establish *distributed amplification*.

A common problem in achieving high gain at microwave frequencies is that the necessary large transconductance of the amplifying device requires a large device size (source width or emitter area), which in turn invariably increases the input capacitance. In FETs, in a first-order approximation, the ratio of transconductance to input capacitance is the transit frequency: $g_m/C_{GS} = \omega_T$. In a common-source amplifier, the dominant time constant at the input is therefore

$$\tau_1 = Z_G C_{GS} = Z_G \frac{g_m}{\omega_T} \approx -Z_G\, Y_L\, \frac{A_V}{\omega_T},$$

where Z_G is the generator admittance Y_L the load admittance and A_V the quasi-static voltage gain in common-source configuration. The input time constant is therefore directly linked to the voltage gain of the cell, for a given load admittance.

In narrowband amplifiers, we may be able to compensate for the input capacitance using a matching network, as we have seen. Very wideband amplifiers, however, preclude the use of tuned networks.

To find a way around the input capacitance limitation, we follow two fundamental steps:

(i) Instead of using one large device, we will use several smaller ones to deliver the needed overall transconductance.
(ii) The input (and output) capacitances will then be absorbed into an artificial transmission line.

The second step is the most crucial one. To understand this concept, remember that any transmission line can be modelled using a ladder-type network of concentrated

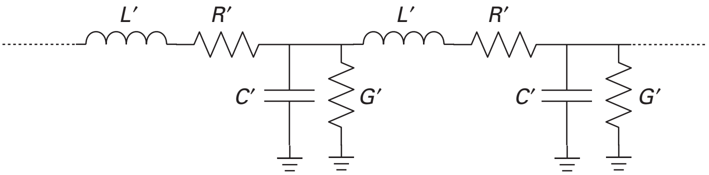


Fig. 5.51 Lumped-element equivalent circuit of a transmission line.

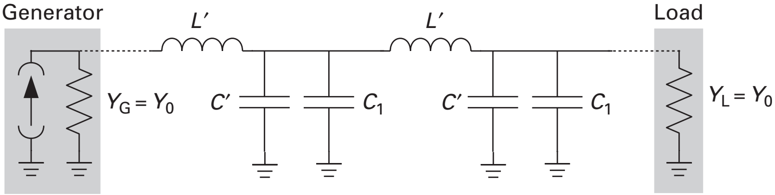


Fig. 5.52 Lossless transmission line loaded with additional shunt capacitances.

elements, such as shown in Figure 5.51. The line is characterised by its distributed inductance L', capacitance C', and the distributed series (R') and shunt (G') losses. The characteristic impedance Z_0 and the propagation constant γ of the line are then:

$$Z_0 = \sqrt{\frac{R' + \jmath\omega L'}{G' + \jmath\omega C'}} \tag{5.179}$$

$$\gamma = \sqrt{(R' + \jmath\omega L')(G' + \jmath\omega C')}. \tag{5.180}$$

Note that the propagation constant $\gamma = \alpha + \jmath\beta$, where α is the attenuation constant and β is the phase constant. In many cases, the losses can be neglected ($R' \ll \omega L'$, $G' \ll \omega C'$) and we obtain the simple relationships:

$$Z_0 \approx \sqrt{\frac{L'}{C'}} \tag{5.181}$$

$$\beta = \omega\sqrt{L'C'}. \tag{5.182}$$

This opens up a fundamental idea: any capacitance to ground can be made to disappear if it is absorbed into a transmission line – it will simply lower the characteristic impedance, and increase the phase constant.

Consider Figure 5.52. The lossless transmission line is loaded by additional shunt capacitances C_1. The transmission line parameters are now

$$Z_0 = \sqrt{\frac{L'}{C' + \frac{C_1}{l}}} \tag{5.183}$$

$$\beta = \omega\sqrt{L'\left(C' + \frac{C_1}{l}\right)}. \tag{5.184}$$

The parameter l is the length of the transmission line segment between each shunt capacitance.

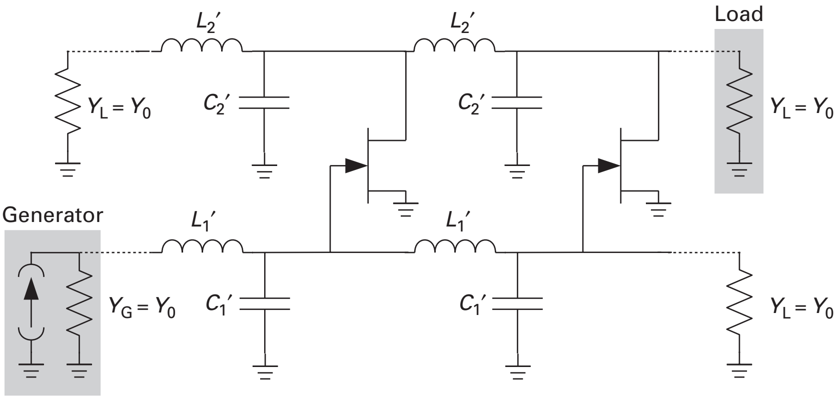


Fig. 5.53 Distributed amplifier concept using FETs in common-source configuration.

Provided that L' and C' are chosen in such a way that $Y_G = Y_L = Z_0^{-1}$, the transmission between generator and load is unaltered by the presence of the additional shunt capacitances!

This observation is not new at all. Its earliest implementation is in the *Pupin coils*, periodically inserted series loading coils (increasing L' in our example) which compensate for the capacitance to ground of telegraph and telephony lines. They were invented in 1894 by Serbian physicist Mihajlo Idvorski Pupin, following earlier suggestions by Oliver Heaviside in 1893.

Of course, the LC combination also acts as a low-pass filter. The frequency

$$\omega_{\text{Bragg}} = \frac{1}{l\sqrt{L'(C' + C_1/l)}} \tag{5.185}$$

is called the *Bragg frequency* of the transmission line structure. The length l must be chosen such that the Bragg frequency is significantly above the intended frequency of operation.

Distributed amplifier structures using electron tubes were first described by W. S. Percival in his 1937 patent [30].

General design procedure

We will now apply the concept to an arrangement of FETs in common-source configuration along two transmission lines, connecting the inputs and outputs, as shown in Figure 5.53. Note that the transmission lines at input and output have different inductance and capacitance per unit area. The loading capacitances are now the imaginary parts of the input and output admittances of the common-source gain cells. Using Equations (5.109) and (5.111) and a simplified FET equivalent circuit, we write for the shunt capacitance loading the input line:

$$C_1 = C_{\text{GS}} + C_{\text{GD}}\left(1 + \frac{g_{\text{m}}}{2Y_0}\right), \tag{5.186}$$

provided that the output transmission line is terminated in its characteristic admittance Y_0.

The shunt capacitance loading the output transmission line is

$$C_2 = C_{\mathrm{DS}} + C_{\mathrm{GD}} \left(1 + \frac{g_{\mathrm{m}}}{2Y_0}\right), \tag{5.187}$$

where C_{DS} is the parasitic drain–source capacitance.

The unloaded input and output transmission lines must be chosen such that

- the loaded characteristic impedances correspond to generator and load impedances and
- the phase delays between corresponding nodes on the (loaded) input and output lines are equal.

Assuming identical generator and load impedances, $Z_{\mathrm{G}} = Z_{\mathrm{L}} = Z_0$, we find

$$Z_1 = \sqrt{\frac{L_1'}{C_1' + \frac{C_1}{l_1}}} \overset{!}{=} Z_0 \tag{5.188}$$

$$Z_2 = \sqrt{\frac{L_2'}{C_2' + \frac{C_2}{l_2}}} \overset{!}{=} Z_0. \tag{5.189}$$

The phase synchronism requirement translates into

$$\beta_1 l_1 = \beta_2 l_2$$

$$l_1 \cdot \sqrt{L_1' \left(C_1' + \frac{C_1}{l_1}\right)} = l_2 \cdot \sqrt{L_2' \left(C_2' + \frac{C_2}{l_2}\right)}. \tag{5.190}$$

The difference in the unit amplifier cell input and output capacitances may result in very different design parameters for the input and output transmission lines. Figure 5.54 shows this in a practical example. The distributed amplifier shown was fabricated in an experimental Si/SiGe HFET technology [1]. The transmission lines are realised in coplanar waveguide form. The difference in geometry for the input (gate) and output (drain) lines is clearly visible.

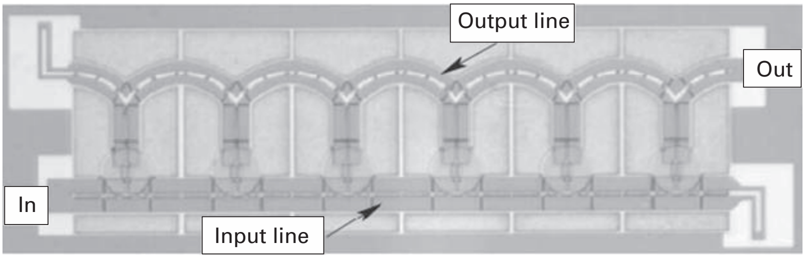


Fig. 5.54 Chip micrograph of a distributed amplifier with 32 GHz bandwidth, realised in a Si/SiGe HFET technology (P. Abele, I. Kallfass, M. Zeuner, J. Müller, Th. Hackbarth, D. Chrastina, H.v.Känel, U. König, and H. Schumacher, *Electronics Letters*, Vol. 39, pp. 1448–1449, 2003. © 2003 IEEE).

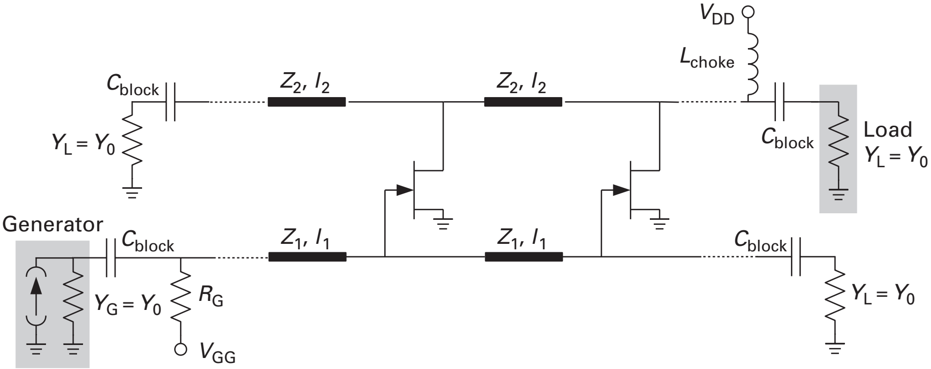


Fig. 5.55 Distributed amplifier with bias arrangement.

The terminating impedances for the input and output lines are placed off-chip in this example – which brings us to a general problem we did not address so far. The distributed amplifier concept in Figure 5.53 did not include the bias arrangement. If we apply a gate voltage to the input and a drain voltage to the output line, a constant current would flow through the terminating impedances attached to the ends of the transmission lines opposite to the input and output ports – resulting in generally unacceptable power dissipation there. The terminating impedances therefore need to be galvanically isolated from the transmission lines. A more practical schematic for a distributed amplifier would therefore look like Figure 5.55. The bias-related elements C_{block} and L_{choke} set the lower cutoff frequency. If a very low lower cutoff frequency is desired, then the on-chip realisation especially of the blocking capacitors may be a significant challenge. L_{choke} is generally placed off-chip.

Gain and loss in distributed amplifiers

Without any losses, the theoretical voltage gain of a distributed amplifier with n stages should be

$$A_{\text{V}} = n\, g_{\text{m}} \frac{Z_0}{2}, \tag{5.191}$$

where g_{m} is the transconductance of the individual cell and Z_0 the characteristic impedance of the output line.

So far, we assumed that the transmission lines were lossless, and that the input and output admittances of the unit amplifier cells were purely capacitive. The latter assumptions particularly are too bold, of course, and we need to assess how the resistive parts of the input and output admittances impact distributed amplifier performance.

In most calculations so far, the gate (or base) series resistance was neglected. This we will abandon here. For the case of a FET, the input line is then loaded with an complex admittance:

$$Y_1 = \frac{\jmath\omega C_1}{1 + \jmath\omega C_1 R_G} \tag{5.192}$$

$$= \omega^2 \frac{R_G C_1^2}{1 + \omega^2 R_G^2 C_1^2} + \jmath\omega \frac{C_1}{1 + \omega^2 R_G^2 C_1^2}, \tag{5.193}$$

where R_G is the gate resistance and C_1 the input capacitance as before. As long as $\omega \ll (R_G C_1)^{-1}$, losses due to R_G need not be accounted for, but they will increase strongly for higher frequencies.

For the output line, some attenuation is always present due to the real part of $\underline{Y}_{20}$ in Equation (5.111), which is g_{DS} in FETs:

$$Y_2 = g_{DS} + \jmath\omega C_2, \tag{5.194}$$

where C_2 is the output capacitance as before. The loss introduced to the drain line is hence frequency-independent.[4]

When the number of stages, n, is increased, the power consumption scales linearly. However, with increasing n, the losses introduced by the amplifier cells become more important and lead to a situation where the gain scales sub-linearly. This introduces a practical limitation to the number of stages. For a detailed analysis, refer to Beyer *et al.* (1984) [4].

Distributed amplifier variations

Matching input and output capacitances

A common problem in distributed amplifiers is that the amplifier cell input capacitance C_1 is much larger than the output capacitance C_2. In turn, the unloaded characteristic impedance of the output line will be significantly smaller than that of the input line. This is significant because the dispersion characteristics of the lines depend on their geometries – different geometries lead to different dispersions, and phase synchronism between input and output lines is increasingly lost with rising frequency.

A simple technique is to increase the output capacitance. This can be done easily using a transmission line stub between the amplifier cell output and the output transmission line. As the amplifier output shows a reasonably high impedance, the transmission line stub acts capacitively when seen from the output line. Figure 5.56 shows this simple concept, which is used in many practical amplifier examples.

The input capacitance can also be lowered by introducing a series capacitance in the unit cell input port. This leads to a capacitive voltage division between the series capacitor and the input impedance of the amplifier, and hence a reduction in gain, but depending on the application, this may be tolerated for the benefit of an increased bandwidth. To allow proper biasing, the capacitor must be bridged with a high-value resistor which has no influence on the RF performance. The measure is shown in Figure 5.57.

[4] The loss due to additional drain (or collector) resistances can be neglected unless they are excessive.

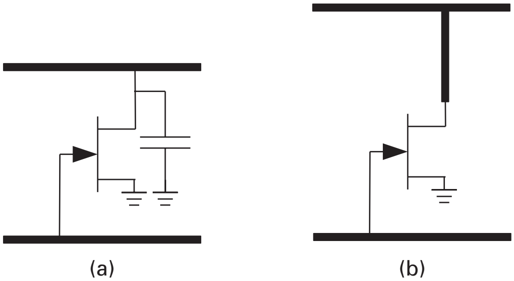


Fig. 5.56 Distributed amplifier unit cell with increased output capacitance: (a) concept and (b) implementation using a transmission line stub.

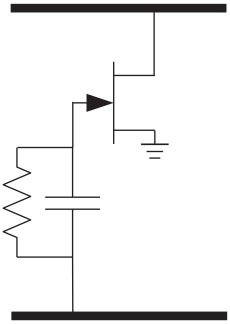


Fig. 5.57 Input capacitance reduction using a series capacitor.

By changing the series capacitance value along the input transmission line (lower towards the generator and higher towards the termination), the input voltage across the amplifying device can be made equal despite the decreasing signal on the transmission line.

Distributed amplifiers with a cascode cell

Despite the potential of the distributed amplifier concept to eliminate input and output capacitances by embedding them into an artificial transmission line, there are good reasons to keep input and output capacitances low. One reason is that high input and output capacitances force the unloaded characteristic impedances of the lines to be very high – the signal-carrying lines then have to be very narrow and will exhibit high ohmic loss. Further, a high input capacitance means that the loss due to the gate resistance will start to matter at much lower frequencies (see Equation (5.192)).

Choosing a cascode as the amplifier unit cell is therefore a logical choice. A simplified configuration is shown in Figure 5.58.

We had seen that the cascode gain cell is prone to producing a negative real part of the output admittance (see p. 336). Here, this effect may be used with benefit to compensate for losses on the output line, but amplifier stability has to be carefully checked, especially at higher frequencies.

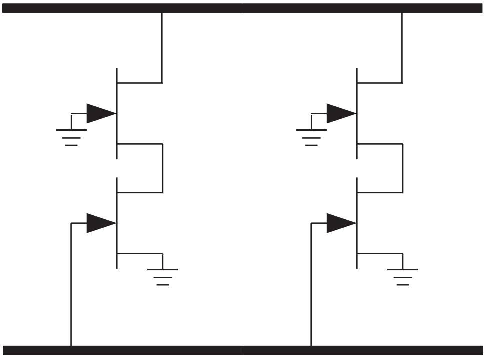


Fig. 5.58 Cascode gain cells in a distributed amplifier structure (bias elements not shown).

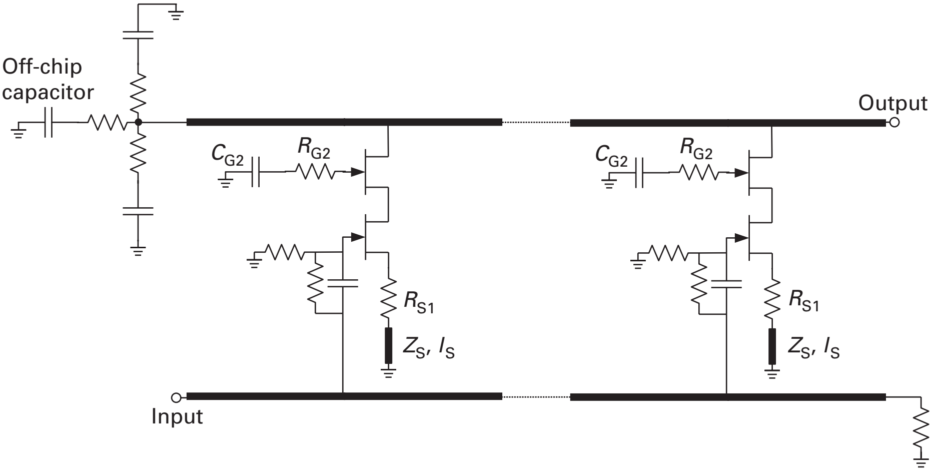


Fig. 5.59 Practical distributed amplifier design using (Al,Ga)As/InGaAs pHEMTs (bias circuitry omitted).

Practical distributed amplifier examples

40 GHz bandwidth distributed amplifier using GaAs pHEMTs

Figure 5.59 shows the schematic diagram of a practical distributed amplifier using a pseudomorphic HEMT process [16]. Several of the measures discussed above have been taken here. The unit cell has a cascode topology, but additionally the input capacitance was reduced using a series capacitor in the input line. The series capacitor is bridged using a high-value resistor; the additional resistor to ground at the gate node improves gain flatness at low frequencies.

The gate termination does not have a DC blocking capacitor here, because the gate line is held at 0 V – the source resistor R_{S1} provides the slightly negative gate–source voltage. Note the elaborate drain termination. This is rather typical of distributed amplifiers for fibre-optic systems where a lower cutoff frequency in the kHz range is required: a broadband termination is created using several RC networks with staggered time constants. The largest capacitor (100 nF in this case) is necessarily placed off-chip.

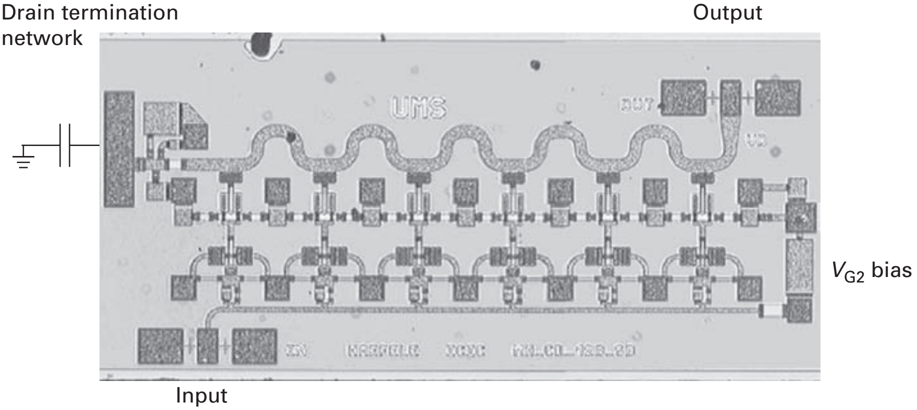


Fig. 5.60 Chip photo of the amplifier shown in Figure 5.59.

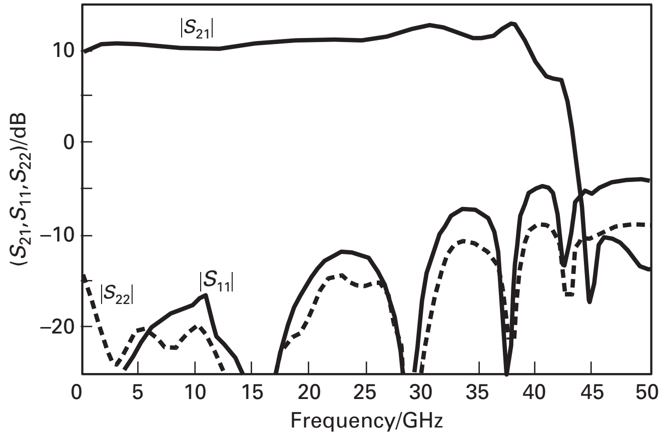


Fig. 5.61 Frequency response of gain ($|S_{21}|$), and input and output reflection coefficients ($|S_{11}|$, $|S_{22}|$) of the distributed amplifier in Figure 5.59.

The design deliberately uses the negative real part of the cascode cell output admittance to compensate for drain–line losses. R_{G2} and R_{S1} improve stability together with the transmission line in the source lead of the cascode, which acts as a small inductor and reduces the cell's gain with increasing frequency, avoiding instability at higher frequencies.

Figure 5.60 shows the chip micrograph of the distributed amplifier. It has six gain stages and is implemented using standard microstrip line technology (the back of the chip is metallised). Two adjacent stages share via the connections to ground – this requires careful assessment of interstage cross-talk issues, but is very efficient in reducing the necessary chip area.

The experimental frequency response (Figure 5.61), shows a very flat gain up to about 40 GHz, where the gain drops sharply. This is a very typical feature of distributed amplification. Another noteworthy feature is the low reflection coefficient for both input and output over a very wide frequency range, which is due to the distributed nature of the input and output impedances.

The midband gain is 11 dB, the output power at 1 dB gain compression (for a definition, see Figure 5.72 on p. 372) is 22.6 dBm measured at 20 GHz.

A distributed amplifier on Si using Si/SiGe HBTs

The distributed amplifier concept is not restricted to FETs. They can also be realised using bipolar transitors or HBTs. In the example used here, the goal is to realise a distributed amplifier in a production Si/SiGe HBT process on lossy substrates.

The latter issue, the lossy substrate (20 Ωcm specific resistivity), introduces an additional complication because neither standard microstrip transmission lines (which use the substrate as the dielectric) nor coplanar waveguides (which would equally introduce large substrate losses) can be used. Instead, a thin-film microstrip transmission line technique (Figure 5.62) was chosen, which creates the microstrip line entirely above the substrate. Here, the signal line was placed in metal 3, while metal 1 acts as the ground plane, shielding the signal completely from the lossy substrate. The thin dielectric, however, leads to very narrow signal lines for the characteristic impedances in question (50–100 Ω) and strongly increases series resistance losses.

Furthermore, the input admittance of a bipolar transistor is not purely capacitive, as we could safely assume for FETs. Using the hybrid π equivalent circuit of Figure 5.17, we can estimate the admittance $\underline{Y}_{10}$ for a bipolar transistor:

$$\underline{Y}_{10,\mathrm{bipolar}} \approx \frac{I_\mathrm{C}}{\beta_\mathrm{f}\, V_\mathrm{T}} + \jmath\omega \left(C_\mathrm{JBE} + \tau_\mathrm{B} \frac{I_\mathrm{C}}{V_\mathrm{T}} \right), \tag{5.195}$$

where β_f is the small-signal current gain in common-emitter configuration, τ_B is the base transit time, I_C is the collector current in this bias point and $V_\mathrm{T} = kT/q$ is the thermal voltage. The real part of $\underline{Y}_{10}$ would strongly attenuate the signal travelling on the input line and has to be eliminated.

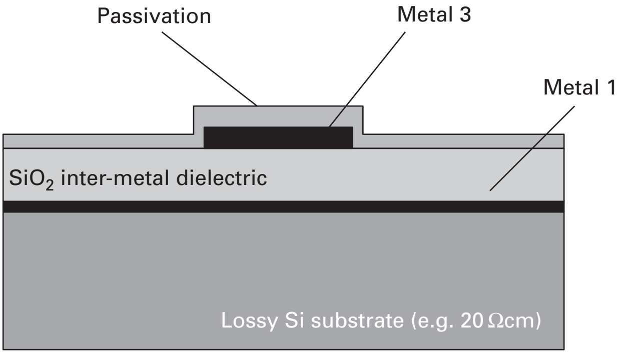


Fig. 5.62 Example of a TFMSL on a silicon substrate.

The latter problem can be solved using a common-collector (emitter follower) input stage (Equation (5.126)):

$$Y_1 \approx \underline{Y}_{12} + \frac{\underline{Y}_{10}}{1 + \frac{g_\mathrm{m}}{Y_\mathrm{L}}}.$$

It is evident that the input admittance is much smaller. Furthermore, we had seen in Equation (5.128) that given a capacitive component of Y_L, the real part of the input admittance becomes negative. This can be used to compensate for ohmic losses on the input line, but always bears the risk of instability.

If a cascode gain cell is chosen, the negative real part of its output admittance can equally be used to compensate for ohmic losses on the output line, with the same stability caveat.

Figure 5.63 shows an example of a differential amplifier where all of these measures have been taken [33]. It was realised in Si/SiGe HBT technology, with transistors of f_T, $f_\mathrm{max} = 80\,\mathrm{GHz}$, on a $20\,\Omega\mathrm{cm}$ substrate.

Three cascaded emitter followers are used in the input to achieve the appropriate low input capacitance and negative input conductance. The differential cascode gain cell has open collector outputs which connect directly to the output transmission lines. The capacitively shunted emitter degeneration resistors in the common-source pair improve the bandwidth through a positive gain slope of this stage.

Note the extensive use of level shifting diodes (transistors with their base–collector contacts tied together). This is necessary due to the low collector–emitter breakdown voltages typical of high-f_T Si/SiGe HBTs.

The unusual differential topology solves an additional problem of silicon-based MMICs: The absence of through-the-substrate via holes makes low-inductance grounding highly critical. The differential topology eases packaging by creating an on-chip

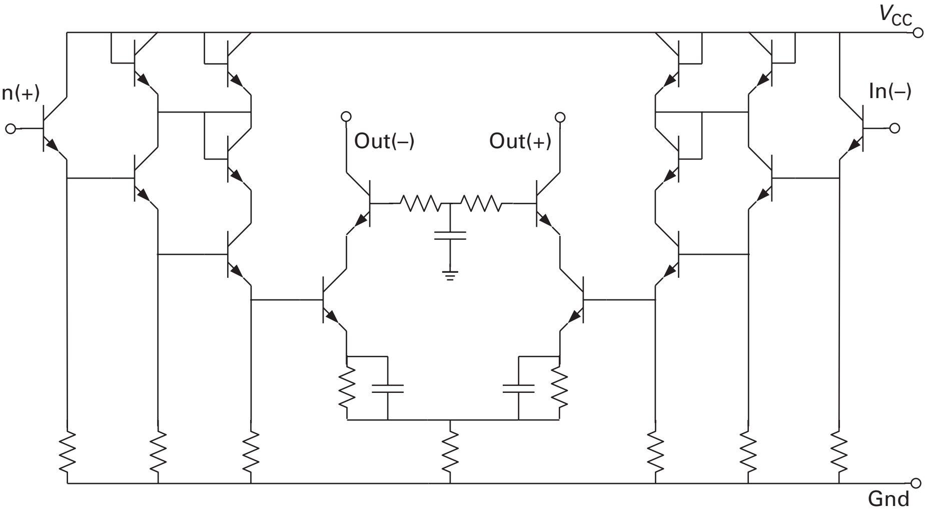


Fig. 5.63 Schematic of a differential distributed amplifier gain cell using Si/SiGe HBTs.

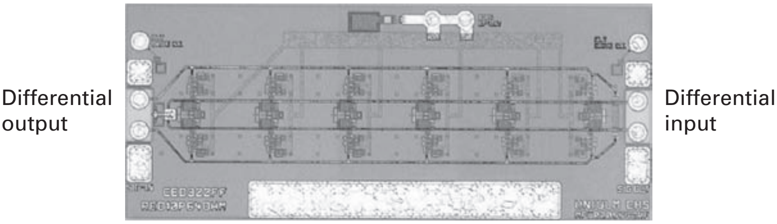


Fig. 5.64 Chip photo of the differential distributed amplifier.

ground, as already discussed. In wideband amplifiers, it is not suitable for all system architectures, however, due to the need for ultra-wideband baluns.

Figure 5.64 shows the chip micrograph of the structure. The chip size is 1.7 × 0.7 mm^2. The narrow width of the thin film Microstrip line is very apparent. The differential gain is 13.6 dB and the −3 dB bandwidth is 32.2 GHz.

5.4.10 Low-noise amplifier

A very frequent requirement is the design of an amplifier with minimum noise figure – an LNA. This is especially important in weak signal reception environments such as in satellite receivers.

We have seen earlier that the noise figure of any two-port depends on the source reflection coefficient presented to it (see p. 310). The parameters needed for noise-optimum design are

(i) the noise-optimised source reflection coefficient for which the two-port noise figure is minimal: $\Gamma_{\mathrm{S,opt}}$;
(ii) the minimum noise figure $F_{\min}$ which provides the two-port noise figure under the condition that the source reflection coefficient is the noise-optimised one: $\Gamma_{\mathrm{S}} = \Gamma_{\mathrm{S,opt}}$;
(iii) the normalised equivalent noise resistance r_{n}, which describes the sensitivity of the noise figure F on deviations from the noise-optimised source reflection coefficient $\Gamma_{\mathrm{S,opt}}$.

Using these parameters, the noise figure is given by

$$F = F_{\min} + \frac{4r_{\mathrm{n}} \left|\Gamma_{\mathrm{S}} - \Gamma_{\mathrm{S,opt}}\right|^2}{\left(1 - |\Gamma_{\mathrm{S}}|^2\right) \left|1 + \Gamma_{\mathrm{S,opt}}\right|^2}. \tag{5.196}$$

In practical two-ports using active devices, the noise parameters are also bias-dependent. Of particular interest is the dependence of $F_{\min}$ on the drain or collector current. Qualitatively, it is shown in Figure 5.65.

An additional aspect needs to be considered – while in principle any reflection coefficient $|\Gamma| \leq 1$ can be transformed into any other using reactive matching networks, practical limitations need to be considered. If the end points of the transformation are located too far apart, the resulting matching network will either be very narrow band (if the reacting matching elements are sufficiently low loss) or introduce significant

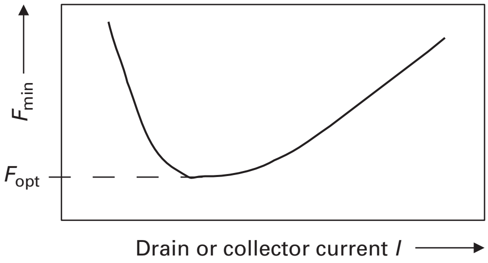


Fig. 5.65 Qualitative dependence of the minimum noise figure on the source or collector current.

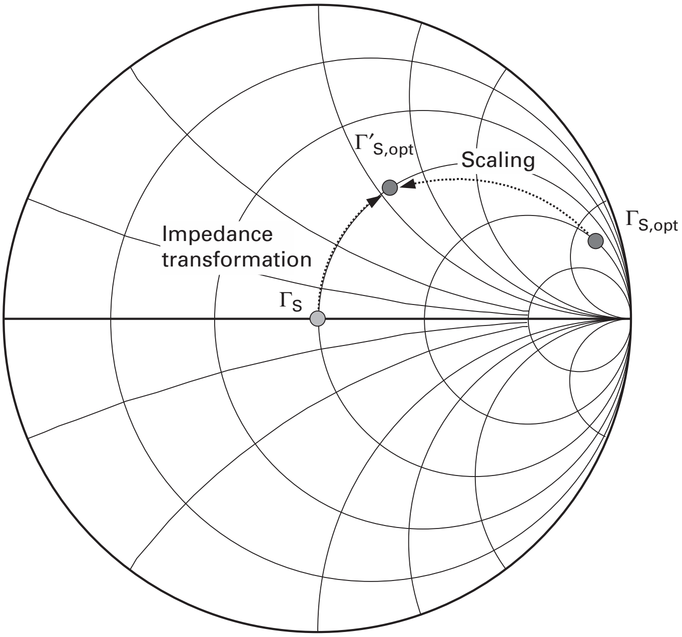


Fig. 5.66 Noise matching example using device scaling and impedance transformation.

additional losses, which deteriorate the noise figure according to Friis' formula. For LNA design, this means that $\Gamma_{S,opt}$ should be suitably located. $\Gamma_{S,opt}$ can be changed by changing the device width ('scaling') – a larger device width results in larger values of $Y_{S,opt}$.

For a better understanding, refer to Figure 5.66. We assume that the original source reflection coefficient Γ_S, e.g. the feed point impedance of an antenna at resonance, is real, and the corresponding impedance is equal to the normalising impedance of the Smith chart, hence $\Gamma_S = 0$. The original noise-optimised reflection coefficient $\Gamma_{S,opt}$ is located too far towards the outside of the Smith chart. By choosing a larger device, $\Gamma_{S,opt}$ is achieved in a location which is much closer to Γ_S. In fact, this location is ideal because the transformation from Γ_S to $\Gamma'_{S,opt}$ can be achieved conveniently using only a series inductance.

The fundamental design steps of the LNA's input stage are hence the following:

(i) Pick a suitable device size which puts $\Gamma_{\mathrm{S,opt}}$ into a convenient location with respect to the original source reflection coefficient Γ_{S}.
(ii) Adjust the bias point such that the optimum F_{min} is achieved.
(iii) Design the input matching network.

Because the bias point affects $\Gamma_{\mathrm{S,opt}}$, a few iterations may be necessary.

In principle, matching for optimum noise performance ($\Gamma_{\mathrm{S}} = \Gamma_{\mathrm{S,opt}}$) and matching for optimum power transfer at the input ($\Gamma_{\mathrm{S}} = \Gamma_{\mathrm{in}}^{\star}$) are unrelated. A frequent requirement, however, is the combination of optimum noise performance and a minimum return loss, hence $\Gamma_{\mathrm{in}}^{\star} \approx \Gamma_{\mathrm{S,opt}}$. This cannot be achieved using impedance transformation networks between the source and the LNA input, because that would modify $\Gamma_{\mathrm{S,opt}}$ for the resulting two-port and Γ_{in} in the same way. Instead, Γ_{in} can be modified in two ways which leave $\Gamma_{\mathrm{S,opt}}$ invariant:

(i) through lossless feedback;
(ii) by mismatching the output for non-unilateral two-ports, utilising the fact that the input reflection coefficient depends also on the load reflection coefficient:

$$\Gamma_{\mathrm{in}} = S_{11} \frac{S_{21} S_{12} \Gamma_{\mathrm{L}}}{1 - S_{22} \Gamma_{\mathrm{L}}}.$$

Figure 5.67 summarises the individual reactive networks surrounding the LNA core, which can be used in the design to fulfil noise and return loss specifications.

The feedback elements Z_{A} (series or current–voltage feedback) and Z_{B} (parallel or voltage–current feedback) are used to set Γ_{in} while leaving $\Gamma_{\mathrm{S,opt}}$ invariant, as discussed. $M1$ provides noise match or, after suitable modification of Γ_{in} using feedback techniques, simultaneous noise and power match (minimum noise figure and minimum input return loss). $M2$ can be chosen either to present the needed Γ_{L} to the LNA core for adjustment of Γ_{in} (see above), or to achieve power match at the output (minimum output return loss).

Very commonly, Z_{A} is an inductor. As already shown in Equation (5.141) this provides an increased real part of the input impedance. Consider the case depicted in Figure 5.68.

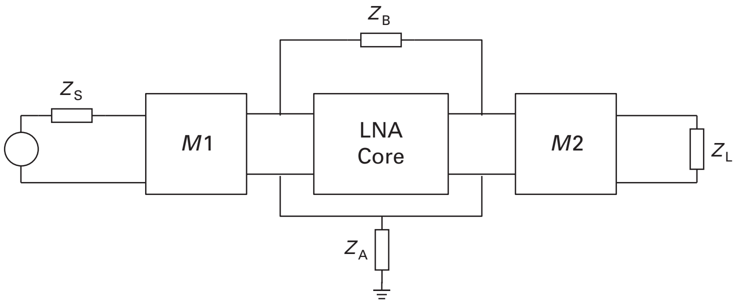


Fig. 5.67 Matching and feedback networks in LNA design.

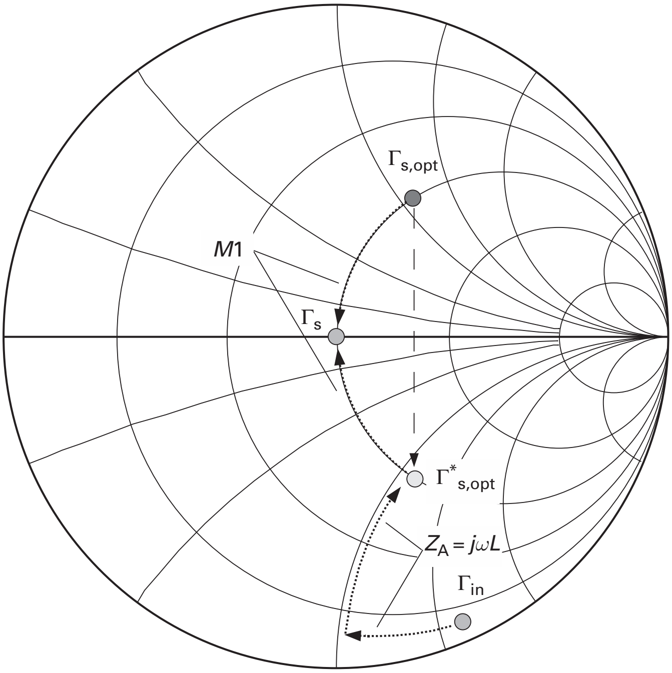


Fig. 5.68 Simultaneous noise and power match example using inductive series feedback and an input matching network $M1$.

Without any feedback or matching network, the input reflection coefficient is Γ_{in}, corresponding to an input impedance of $R_{\text{in}} - \jmath X_{\text{in}}$. The location indicated in the example would be typical for a FET. The goal is now to transform Γ_{in} to a new location $\Gamma'_{\text{in}} \approx \Gamma^{\star}_{\text{S,opt}}$. We connect an inductor L in series to the LNA core. Applying Equation (5.141), we find for the input impedance of the LNA core with feedback:

$$Z'_{\text{in}} = R_{\text{in}} + \omega_{\text{T}} L + \jmath\,(\omega_0 L - X_{\text{in}})\,. \tag{5.197}$$

On the Smith chart, the transformation path corresponding to the effect of L can be interpreted as first increasing the imaginary part, starting from Γ_{in} and then increasing the real part, as shown in the lower part of Figure 5.68.[5]

In a second step, matching network $M1$ (which in the example is simply a series inductor) transforms both Γ'_{in} and $\Gamma_{\text{S,opt}}$ towards Γ_{S}, achieving the required simultaneous optimisation of noise and input return loss.

For the LNA core, cascode stages (see p. 333) are very frequently being used at microwave frequencies. This is because the aforementioned scaling, placing $\Gamma_{\text{S,opt}}$ in an ‘easily matchable’ location results frequently in relatively large transistors, where the Miller effect (discussed on p. 318) can be significant – adoption of a cascode topology is a proven way to reduce the increased input capacitance associated with the Miller effect.

[5] As an aside, you may notice that with increasing L, F_{min} decreases – the associated gain G_{ass}, however, also decreases. The entity invariant to reactive feedback is the *noise measure* $M = F_{\text{min}}/\left(1 - G_{\text{ass}}^{-1}\right)$.

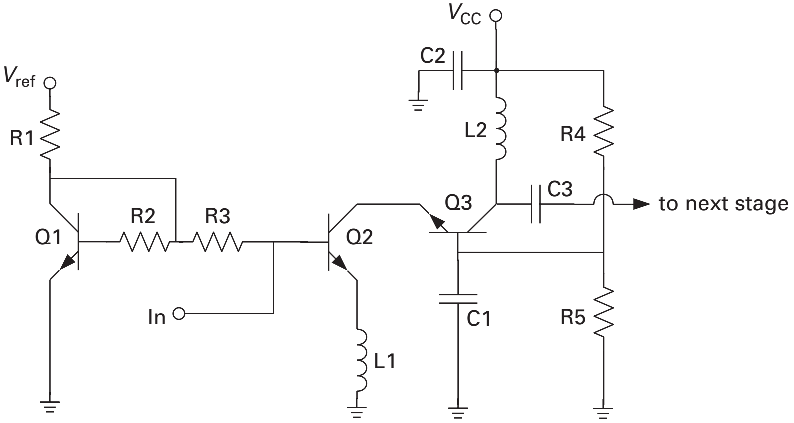


Fig. 5.69 First stage of a three-stage LNA for 24 GHz using Si/SiGe HBTs.

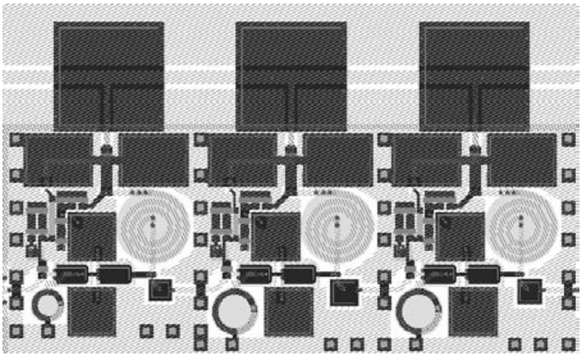


Fig. 5.70 Layout of the three-stage 24 GHz LNA.

As a practical example, we will discuss a three-stage LNA for 24 GHz using Si/SiGe HBTs [35]. The schematic of the first stage is shown in Figure 5.69.

Transistors Q2 and Q3 form the cascode LNA core; Q1 forms a current mirror with Q2 to set the latter's collector current. Q3's base voltage is then set using the voltage divider R4/R5. All capacitors are large-value bypass capacitances.

Inductor L1 is used to allow simultaneous noise and power match along with the proper sizing of Q2. There is no on-chip inductance in series with the *In* port, because the bond wire is used instead, efficiently including this parasitic into the design. L2 forms, together with the capacitance between the collector Q3 and ground and the input capacitance of the following stage, a parallel resonance which provides the LNA with a bandpass characteristic.

The other two stages are identical in topology, but due to the different source impedances of the preceding stages, the inductive source degeneration of the common-emitter transistor (Q2) is adjusted.

Figure 5.70 shows the layout of the three-stage amplifier. Note that the source degeneration inductors are constructed as two-layer stacked inductors, while the tank circuit inductors (L2) are conventional spiral inductors.

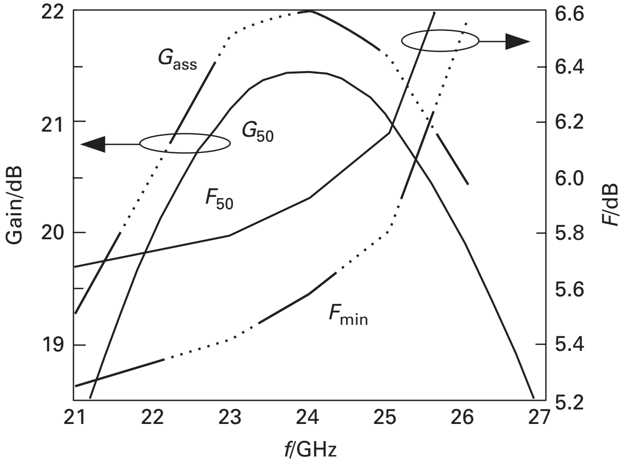


Fig. 5.71 Gain and noise performance of the 24 GHz LNA using Si/SiGe HBT technology.

The circuit was realised in a Si/SiGe HBT technology with f_T, $f_{max} = 80$ GHz and characterised on wafer. Results of a small-signal characterisation are shown in Figure 5.71. The circuit shows the targeted bandpass performance with the gain peak at 24 GHz (the intended application is in the 24 GHz ISM band). The minimum noise figure at 24 GHz is 5.6 dB, while the 50 Ω noise figure is slightly below 6 dB. This deviation is not surprising, as the circuit was designed to provide optimum noise figure with the bond wire parasitic included. The gain with a 50 Ω source impedance is $G_{50} = 21.4$ dB, while the associated gain under noise match conditions is $G_{ass} = 22$ dB – in this circuit, the noise-optimised source impedance is actually slightly closer to 50 Ω than the input impedance.

5.4.11 Amplifier linearity

So far, we treated amplifiers as perfectly linear systems – the output signal can always be described as a linear combination of the input signals. In reality, however, any circuit including active devices will show a non-linear behaviour and the assumption of linearity holds only for small deviations around a given operating point.

In practice, the non-linear behaviour of amplifiers will generate *nonlinear distortions*, which create non-linear deviations in time-domain signal shape, and additional spectral components in the frequency domain which have to be reckoned with.

Single-tone excitation

A common way to treat general non-linear functions is the Taylor series expansion. A non-linear function $f(x)$ is expanded around $x = x_0$ as

$$f(x) = \sum_{\nu=0}^{\infty} \frac{f^{(\nu)}(x_0)}{\nu!}(x - x_0)^{\nu}, \tag{5.198}$$

where $f^{(\nu)}(x_0)$ is the ν^{th} derivative of f with respect to x in $x = x_0$.

Now assume that we apply a single sinusoidal signal to our non-linear system: $a(\omega t) = A_0 + A\sin(\omega t)$.

The output signal $f(\omega t)$ can now be described by the following Taylor series expansion ($x_0 = 0$):

$$\begin{aligned} f(\omega t) = k_0 A_0 & \\ &+ k_1 \sin(\omega t) \\ &+ k_2 \sin^2(\omega t) \\ &+ k_3 \sin^3(\omega t) \\ &+ \ldots . \end{aligned} \tag{5.199}$$

The first two lines in Equation (5.199) provide the linear response, while the following terms are non-linear distortions. Consider that

$$\begin{aligned} \sin^2(\omega t) &= \frac{1}{2}\left[1 - \cos 2\,\omega t\right] \\ \sin^3(\omega t) &= \frac{1}{4}\left[3\sin(\omega t) - \sin(3\omega t)\right], \end{aligned}$$

and we find that Equation (5.199) turns into

$$\begin{aligned} f(\omega t) = k_0 A_0 + \frac{k_2}{2} & \\ &+ \left(k_1 + \frac{3k_3}{4}\right)\sin(\omega t) \\ &- \frac{k_2}{2}\cos(2\omega t) \\ &- \frac{k_3}{4}\sin(3\omega t) \\ &+ \ldots . \end{aligned} \tag{5.200}$$

We easily see that the non-linear distortion results in new spectral components (harmonics) being generated, which are related to the fundamental components as integer multiples.

A simple procedure to assess an amplifier's linearity is the single-tone excitation test. A test generator with high spectral purity and adjustable power is connected to the input of the amplifier, and a spectrum analyser to the output. Increasing the input power (P_{in}), the power of individual spectral components at the output is recorded. Plotting the output power levels as a function of the input power on a double-logarithmic scale, we obtain a graph similar to the one shown in Figure 5.72.

For low power levels, the output power of the fundamental spectral line will increase linearly with the input power. Gradually, it will, however, rise more slowly – *gain saturation* sets in. When the power ratio between the extrapolated linear increase and the actual curve is 1 dB, the 1 dB compression point ($P_{-1\text{dB}}$) has been reached. It is a measure of the maximum power the amplifier can deliver in linear operation. Depending on the application, it is referred to the input (e.g. LNAs) or the output (e.g. power amplifiers).

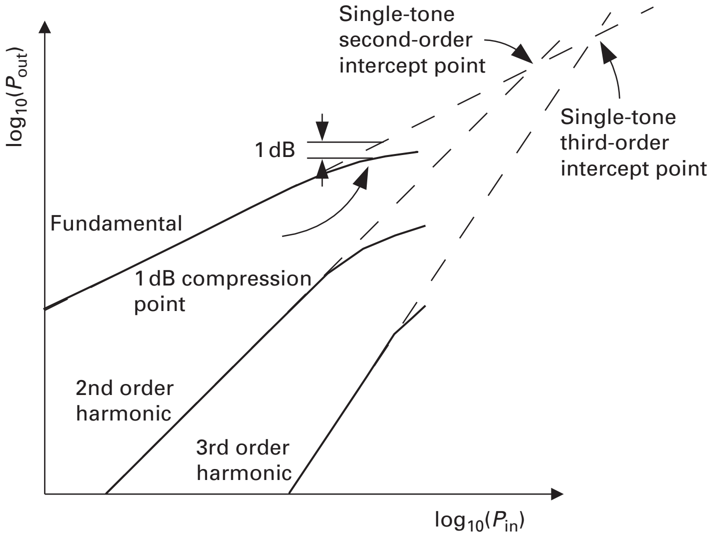


Fig. 5.72 Single-tone excitation test of an amplifier, showing definitions for the 1 dB compression, second-order intercept and third-order intercept points.

The spectral power at the second harmonic increases twice as fast as the fundamental power, before it also shows saturation. Extrapolating the curve at low input powers, we find the *single-tone second-order intercept point* at the point where the extrapolation intersects the extrapolated fundamental power.

The spectral power at the third harmonic increases three times as fast as the fundamental power. Its extrapolation intersects the extrapolated fundamental power at the *single-tone third-order intercept point*.

The intercept points can also be referred to the input or the output, depending on the application.

In many applications where the operational bandwidth is only a small fraction of the carrier frequency, the generation of harmonics is not necessarily a problem, because they can easily be removed by filtering. For example, frequency modulated (FM) transmitters are operated under strongly non-linear conditions (class C, see p. 377), and the resulting harmonics in the output signal are simply removed by low-pass filtering.

Two-tone excitation

An FM signal is a particularly simple example of modulation, because the resulting signal has only a single spectral component (which varies in frequency, but that is irrelevant here). Most modulated signals, however, consist of many spectral components which are present at the amplifier input simultaneously.

To understand what amplifier non-linearity will do to these signals, let us construct a simple experiment, where the input signal is formed by two spectral components (at ω_1 and ω_2) of equal amplitude, applied to the input of the amplifier. The output is connected to a spectrum analyser again. The corresponding block diagram is shown in Figure 5.73.

Again, the test generators need to have very high spectral purity.

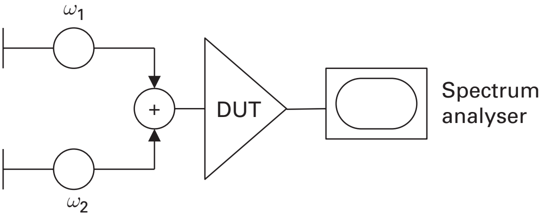


Fig. 5.73 Schematic representation of a two-tone excitation test of an amplifier (DUT = device under test).

Mathematically, the description of the distorted output signal becomes much more complex. We obtain:

- Fundamental components at ω_1 and ω_2.
- Harmonics of the input signals (at $2\omega_{1,2}$, $3\omega_{1,2}$, ...).
- Components due to the product of the two input signals – consider that

$$\sin(\omega_1 t)\sin(\omega_2 t) = \frac{1}{2}\left\{\cos\left[(\omega_1 - \omega_2)t\right] - \cos\left[(\omega_1 + \omega_2)t\right]\right\}. \tag{5.201}$$

  The multiplication term therefore produces spectral components at the sum and difference of the two input spectral lines. These components are called *two-tone second-order intermodulation products*.
- Components due to the product of a fundamental component and a second-order harmonic

$$\sin^2(\omega_1 t)\sin(\omega_2 t) = \frac{1}{2}\left[\sin(\omega_2 t) - \sin(2\omega_1 t + \omega_2 t) + \sin(2\omega_1 t - \omega_2 t)\right] \tag{5.202}$$

$$\sin^2(\omega_2 t)\sin(\omega_1 t) = \frac{1}{2}\left[\sin(\omega_1 t) - \sin(2\omega_2 t + \omega_1 t) + \sin(2\omega_2 t - \omega_1 t)\right]. \tag{5.203}$$

  These terms hence generate spectral components at $2\omega_{1,2} \pm \omega_{2,1}$, which are called *two-tone third-order intermodulation products.*
- Higher-order components which are neglected here.

Figure 5.74 schematically shows the spectral components generated by non-linear distortion of a two-tone signal, up to the third order. Note that second- and third-order harmonics as well as the second-order intermodulation products are significantly far away from the original signal and can most often be removed by filtering. Of particular concern are two third-order intermodulation products at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, because they are close to the original spectral components and cannot be removed by filtering.

Just as in case of the single-tone excitation, we can plot the output powers at the fundamental tones and the close-in third-order intermodulation components as a function of the input power. Figure 5.75 shows an example of such a measurement. The two-tone third-order intercept point is found by extrapolating the low-power portions of the curves, where the relationship between input and output powers has a linear shape on a double-log plot. It can be referred to the input or the output.

Often the system requirement will be formulated in terms of the *intermodulation distance (IMD)* or the *dynamic range*, not in terms of the intercept points.

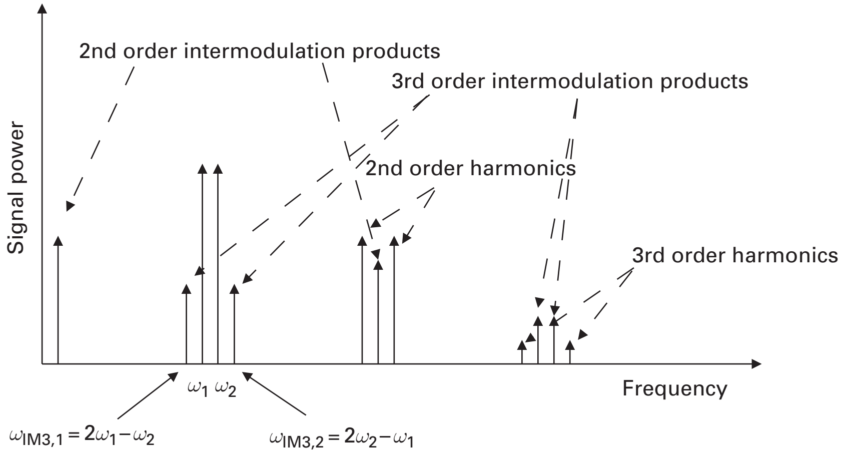

Signal power
2nd order intermodulation products
3rd order intermodulation products
2nd order harmonics
3rd order harmonics
ω_1 ω_2
Frequency
$\omega_{IM3,1} = 2\omega_1 - \omega_2$
$\omega_{IM3,2} = 2\omega_2 - \omega_1$

Fig. 5.74 Schematic representation of spectral components generated from a two-tone excitation through second- and third-order non-linearities.

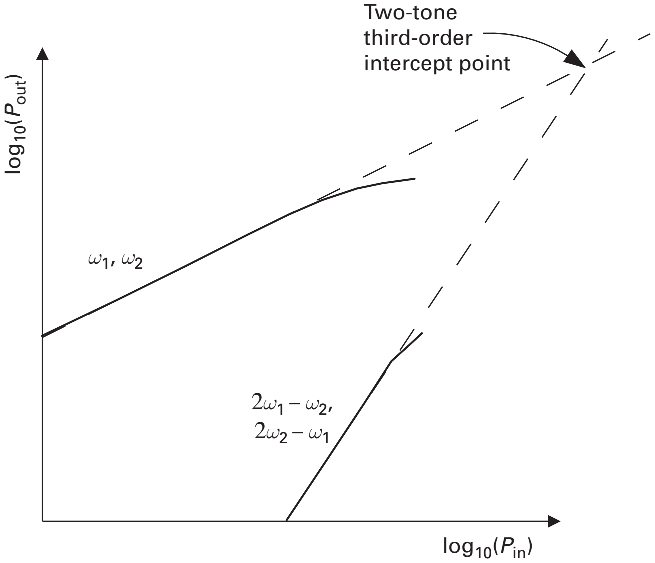


Fig. 5.75 Determination of the two-tone third-order intercept point.

The IMD, measured in a two-tone excitation test, is simply the power ratio between the power level of the two carriers at the amplifier output and the highest intermodulation spectral lines. The most prominent ones will typically be third-order intermodulation products. Then, the IMD can be calculated from the third-order intercept point. Consider again Figure 5.75 and remember that on the double-log scale, the $P_{\text{out}} = f(P_{\text{in}})$ transfer curve for the fundamental component has a slope of 1, while it is 3 for the third-order intermodulation products. The distance between the curves for the fundamental and the intermodulation products is the IMD on the log–log scale. The

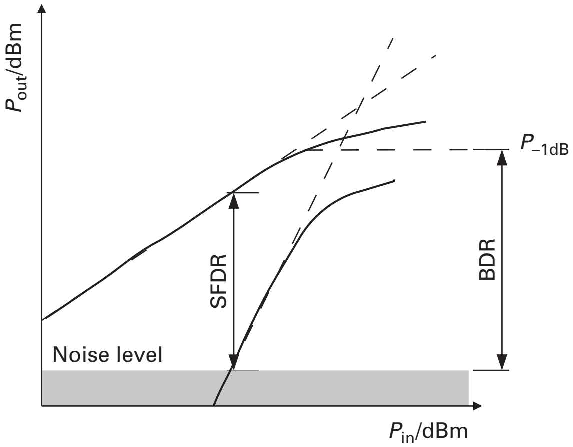


Fig. 5.76 Determination of the SFDR and the BDR in a two-tone excitation measurement.

powers are expressed most often in dBm[6] on a logarithmic scale. Therefore,

$$\frac{\mathrm{IMD}}{\mathrm{dB}} = 2\left(\frac{\mathrm{IIP3}}{\mathrm{dBm}} - \frac{P_{\mathrm{in}}}{\mathrm{dBm}}\right), \tag{5.204}$$

where IIP3 is the third-order intercept point referred to the input.

On a linear scale (powers in W), the IMD can be expressed as

$$\mathrm{IMD} = \left(\frac{\mathrm{IIP3}}{P_{\mathrm{in}}}\right)^2. \tag{5.205}$$

Specification of an amplifier in terms of dynamic range combines linearity and noise. There are two definitions, which are compared in Figure 5.76.

The *spurious-free dynamic range (SFDR)* is the IMD at the point where the power of the third-order intermodulation products is equal to the noise floor. The *blocking dynamic range (BDR)* is the distance between the 1 dB compression point $P_{-1\mathrm{dB}}$ and the noise floor.

Adjacent channel power ratio

Modern communication systems have frequently very complex modulation schemes, with many spectral components present. They are, therefore, very sensitive to intermodulation effects in non-linear amplifiers. A two-tone measurement can only give an indication of linearity, but is no solid proof of the amplifier's suitability.

A very realistic test is the ACPR (adjacent channel power ratio) test, which is always specific to a certain modulation technique. Figure 5.77 shows an example for a UMTS signal. In a first step, the integral powers within the channel bandwidth need to be calculated from the spectral analysis. Then, the ACPR is calculated as the ratio of the power in the band of operation to either the lower or higher adjacent channel. It is a

[6] dBm means decibels relative to 1 mW, i.e. 0 dBm = 1 mW, 20 dBm = 100 mW, etc.

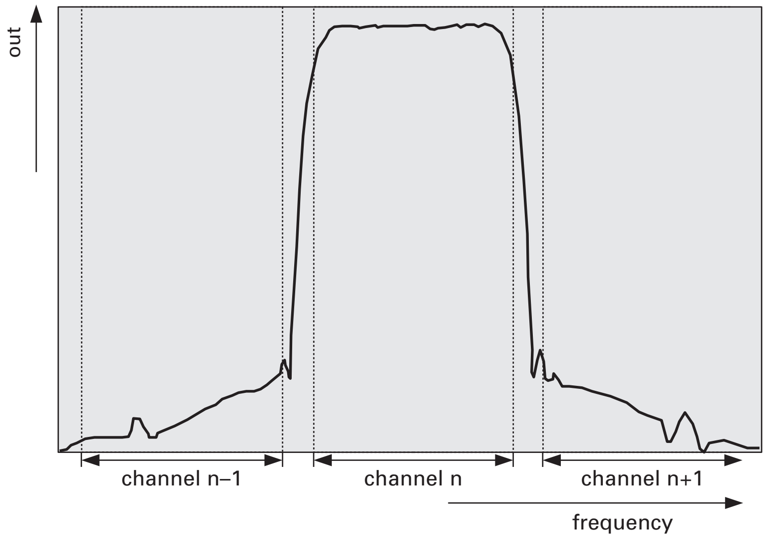


Fig. 5.77 Example of an Adjacent Channel Power ratio (ACPR) measurement: Power spectral density (PSD) versus frequency, with channel limits indicated.

direct measure of the interference generated by transmitter non-linearities in adjacent channels.

5.4.12 Power amplifiers

Power amplifiers have the task of amplifying signals before they are delivered to loads, such as antennas or cables. Critical criteria are

- maximum output power, for example measured in terms of output power at the 1 dB compression point $P_{-1\,\mathrm{dB}}$ (see Figure 5.72);
- gain (either small-signal gain or large-signal gain at a given output power);
- gain and potentially phase deviation across the operational bandwidth;
- linearity, defined by parameters such as the output-referred two-tone third-order intercept point, the IMD at a given output power, or the ACPR at a given output power;
- efficiency – at microwave frequencies, it is customary to use the *power added efficiency (PAE)*, the ratio of the power difference between output and input to the DC power:

$$PAE = \frac{P_{\mathrm{out}} - P_{\mathrm{in}}}{P_{\mathrm{DC}}} = \eta\left(1 - \frac{1}{G}\right), \tag{5.206}$$

where η is the collector or drain efficiency ($\eta = P_{\mathrm{out}}/P_{\mathrm{DC}}$) and G the amplifier gain.

Classes of operation

Since the days of vacuum tubes, amplifier operation has been described by *classes*, which describe where the amplifying devices are biased in a quiescent state.

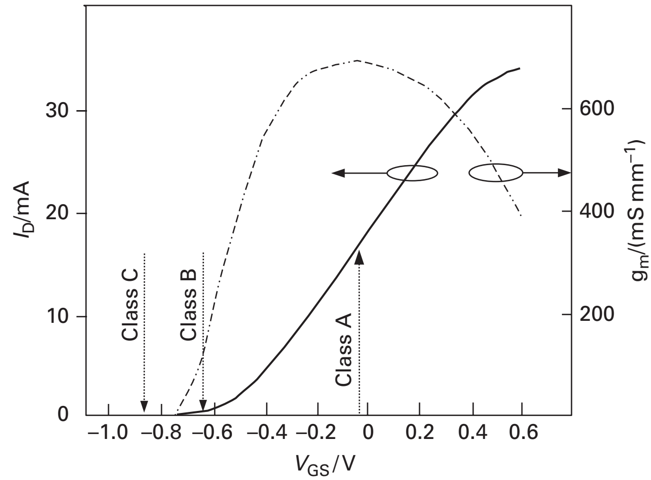


Fig. 5.78 HEMT drain current I_D and transconductance g_m as a function of the gate-source voltage V_{GS} with bias points for power amplifier classes A, B and C indicated.

For an understanding of the ‘classical’ classes A, B and C, refer to Figure 5.78. The example shows the drain current and transconductance of a HEMT. For the classification, we observe the drain current curve.

In a class A amplifier, the gate-source voltage V_{GS} is set in the region where the output current I_D is a linear function of the input voltage V_{GS} – the transconductance is approximately constant. For both positive and negative half-waves of the input signal, current will flow – the *conduction angle* is 360°. In this bias point, the amplifier will exhibit a very high linearity, but low efficiency. The theoretical maximum is 50%, but at microwave frequencies, values of 30% would already be very satisfactory.

For class B, the device is biased at pinch-off. Only the positive half-wave of the input signal will then generate an output current flow – the conduction angle is 180°. The efficiency will increase theoretically to 78.5% ($\pi/4$), and at microwave frequencies it can still reach 60% or higher, but the deviation from a sine wave in the output current creates non-linear distortions.

A class C amplifier has a quiescent bias point where V_{GS} is significantly below the threshold voltage V_{th}. Output current will flow only if the momentary $V_{GS}(t) > V_{th}$, therefore the conduction angle is $<180°$. The efficiency can still be higher; however, due to the lower conduction angle, the non-linear distortions are also increased.

Switched amplifiers

There is another interpretation of class C operation, which is helpful in understanding the way that amplifiers in class D, E and F operate. For this, look at Figure 5.79. It shows the output I–V characteristics of a HEMT (but this could be any FET). Provided that the driving voltage is large enough, the transistor simply changes between two saturated states with very different differential resistances. In the quiescent point, the transistor is in cut-off and the differential resistance between drain and source is very

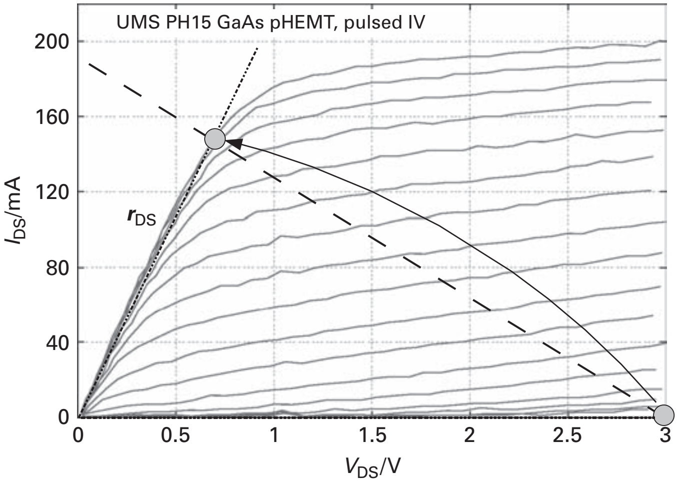


Fig. 5.79 Saturated class C operation in the output I–V characteristics of a HEMT.

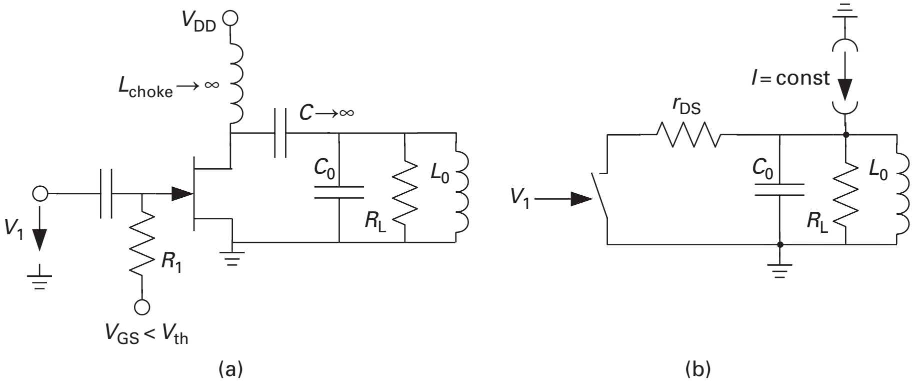
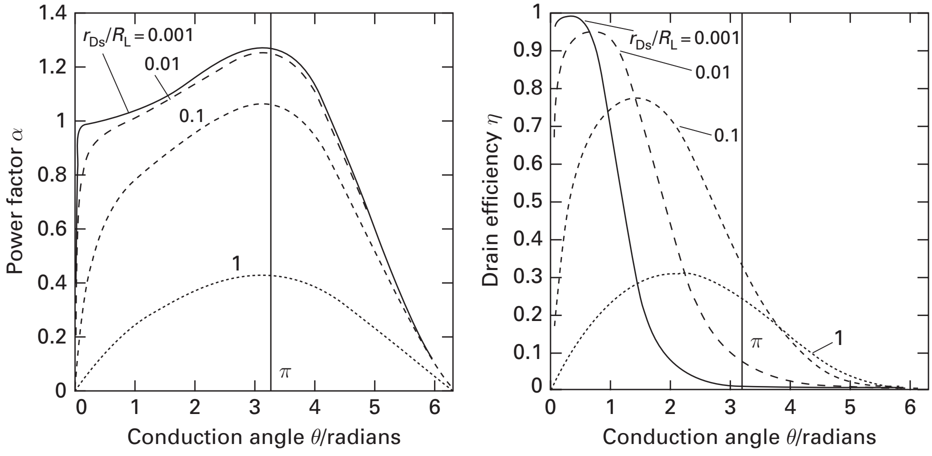


Fig. 5.80 (a) Simple class C amplifier topology and (b) its equivalent circuit.

high. For a sufficiently high input voltage, the transistor reaches another saturated state with small r_{DS}.

We can, therefore, model the transistor in saturated class C operation simply by a switch in series with its residual differential resistance r_{DS}. Figure 5.80(a) shows a simple class C amplifier stage. The load is embedded in an LC resonant circuit which acts as a bandpass filter to suppress the harmonic frequency components other than the fundamental. The RF choke (L_{choke}) provides a constant current, at least on the time scale of interest. This circuit can also be realised with an LC parallel resonant circuit, by the way.

Replacing the choke with a constant current source, and the FET with a periodically actuated switch and its series resistance r_{DS}, we arrive at the equivalent circuit in Figure 5.80(b). The class C amplifier operates in this configuration by periodically shunting current away from the load.



(a)

(b)

Fig. 5.81 Class C amplifier: (a) power factor α and (b) drain efficiency β as a function of frequency.

A detailed analysis of class C operation is found in [17]. First, note that due to the RF choke, the average voltage across the load is the supply voltage V_{DD}. The peak voltage is $(1+\alpha)V_{DD}$ and the minimum voltage $(1-\alpha)V_{DD}$, where

$$\alpha(\theta) = \frac{4\sin\left(\frac{\theta}{2}\right)}{\theta + \sin(\theta) + \frac{2\pi r_{DS}}{R_L}}. \tag{5.207}$$

Here, θ is the conduction angle. Note that for $r_{DS} \to 0$, $\alpha_{max} = \alpha(\theta = \pi) = 1.27$ – the maximum voltage across the transistor can, therefore, exceed the supply voltage by a factor of 2.27.

The output power in saturated class C operation is

$$P_{out} = \frac{(\alpha V_{DD})^2}{2R_L}. \tag{5.208}$$

The drain efficiency is

$$\eta(\theta) = \pi \frac{r_{DS}}{R_L} \frac{\alpha^2}{\theta - 2\alpha \sin\left(\frac{\theta}{2}\right)}. \tag{5.209}$$

Both the power factor α and the drain efficiency η are shown in Figure 5.81 as a function of the conduction angle θ. Note that the output power always peaks at $\theta = \pi$, but that the efficiency has its maximum at much lower conduction angle. The normalised on-resistance of the FET, r_{DS}/R_L, has a significant influence on both the output power and the drain efficiency.

Class D amplifier

Above, we interpreted the class C amplifier as a resonant circuit driven by current pulses, where for maximum efficiency the current flow angle was $\theta < \pi$. We can, of course, not only turn the current on and off, but actually reverse it, as shown schematically in Figure 5.82.

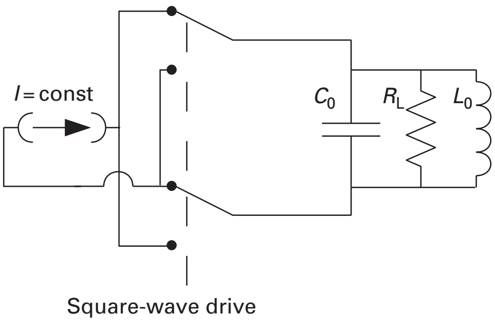


Fig. 5.82 Class D amplifier principle.

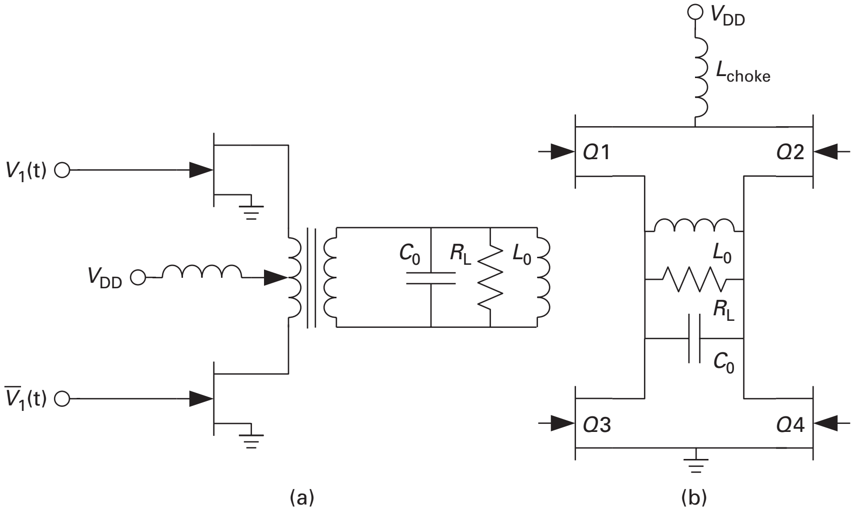


Fig. 5.83 Example implementations of class D amplifiers. After [17].

Instead of a single-pole, single-throw switch, the equivalent circuit shows a double-pole, double-throw switch which periodically reverses the current through the load. The parallel resonant circuit again eliminates all harmonic frequency components except the fundamental one.

In practice, the switches are realised with transistors, of course. Figure 5.83 shows two examples. In Figure 5.83(a), the current reversal is achieved using a transformer where the current is fed into the centre tap, and the ends of the primary coil are connected alternatingly to ground. This is a very common solution at lower frequencies.

At microwave frequencies, the transformers are difficult to realise, and in any case they do not lend themselves easily to monolithic integration. The circuit in Figure 5.83(b) is then more practical – it avoids transformers altogether; however, now we have four transistors to control instead of two: in this bridge configuration, transistors Q1 and Q4, and Q2 and Q3 conduct alternately to achieve the current phase reversal across the load.

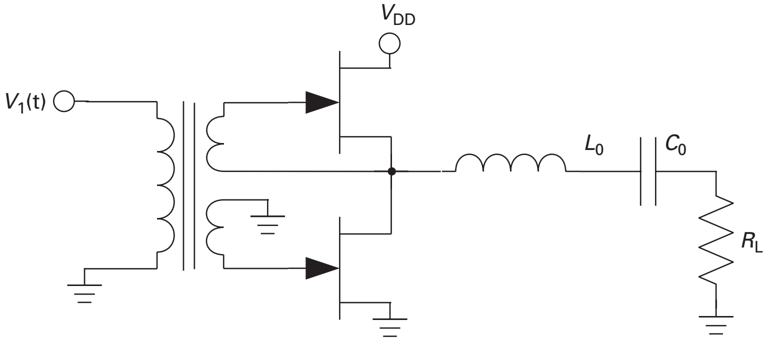


Fig. 5.84 Class D amplifier example using a series-fed load. After [17].

Note that in both implementations, the load floats – it has no direct ground reference, which is often a problem for microwave systems where ground-referenced (single-ended) transmission is more common. This can be avoided in a class D amplifier if a series-fed load is applied. This is shown in Figure 5.84. The bandpass function is now realised with a series resonant circuit (L_0, C_0), and the voltage is alternated, not the current. Still a balun is needed at the input, unless $V_1(t)$ is already available in differential form.

The class D amplifier output power is in saturation [17]:

$$P_{\text{out}} = \frac{2V_{\text{DD}}^2}{\pi^2 R_{\text{L}}}, \tag{5.210}$$

while the drain efficiency is

$$\eta = \frac{R_{\text{L}}}{R_{\text{L}} + r_{\text{DS}}}, \tag{5.211}$$

at least for the circuits according to Figures 5.83(a) and 5.84. r_{DS} is again the channel resistance of the FET for low V_{DS} ('on-resistance'). For the circuit in Figure 5.83(b), the efficiency is lower because the switch resistance doubles.

Class D amplifiers place high demands on the ideality of the switches and on the timing. This is particularly true for circuits such as in Figure 5.83(b) or 5.84, where switches are stacked – they must never conduct at the same time, not even for brief periods. Therefore, class D amplifiers are mostly restricted to lower RF frequencies.

Class E and F amplifiers

Class E and F amplifiers are derived from class C. The idea in a class E amplifier [38] is to make sure that the drain-source voltage of the switching transistor (see Figure 5.80) is zero when the transistor switches, leading to a reduction of losses due to capacitive charging. This can be achieved by a modified output network. Consider Figure 5.85. At first glance, it looks like a class C amplifier with a series resonant feed, with the addition of shunt capacitor C_2. Additionally, the series resonant filter is tuned to a frequency below the intended frequency of operation ω_0 by increasing L_1. Adjusting L_1 and C_2, the voltage-free switching condition is found and the efficiency is maximised.

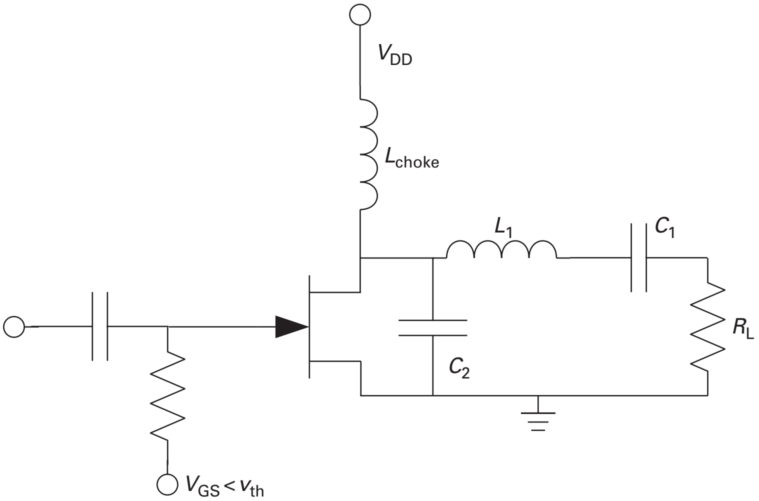


Fig. 5.85 Simplified class E amplifier schematic.

An in-depth analysis of class E operation can be found, e.g. in [8]. The inductance L_1 is chosen as

$$L_1 = \frac{1 + 1.153\omega_0 C_1 R_L}{\omega_0^2 C_1}. \tag{5.212}$$

The capacitance C_2 is

$$C_2 = \frac{2}{3.467\omega_0 \pi R_L}. \tag{5.213}$$

A problem may be that the maximum drain-source voltage is even higher than in the class C amplifier, $V_{DS,max} \approx 3.56 V_{DD}$.

The class F amplifier increases the efficiency by appropriately terminating the harmonics. The idea is to achieve square-wave voltage excitation with respect to the drain-source voltage, as in case of the series-fed class D amplifier (Figure 5.84).

A well-known fact from Fourier analysis is that a square wave (rectangular signal with 50% duty cycle) in the time domain produces only odd harmonics in the frequency domain. We must, therefore, make sure that all odd-numbered harmonics ($n = 1, 3, 5, \ldots$) are still present in the drain-source voltage. The load should, therefore, present an open to the transistors at these frequencies.

A way to achieve this is to use the transforming properties of quarter-wave transmission lines, which we discussed much earlier; see Equation (5.171) on p. 349. Assume that a transmission line, which is a quarter wavelength long at the fundamental frequency, is terminated by a short at all harmonic frequencies except the fundamental. Then an open will appear at the input for all odd-numbered harmonics, while a short results for all even-numbered harmonics, where the electric length of the transmission line is a multiple of $\lambda/2$. The modification of the original class C topology (Figure 5.80) is quite straightforward, as Figure 5.86(a) shows. λ is the wavelength at the frequency of operation ω_0. The resonant circuit formed by L_0, C_0 is resonant at ω_0; C_0 then acts

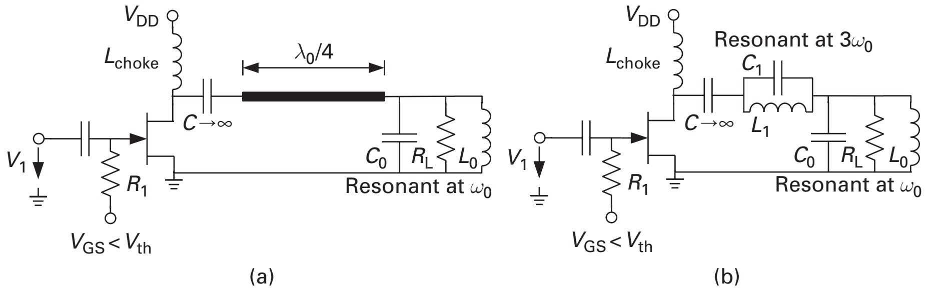


Fig. 5.86 Topology of a class F amplifier: (a) using a quarter-wave transformer and (b) using a parallel resonant trap tuned to the third harmonic.

as a short at the higher-order harmonics. This short is transformed into an open by the quarter-wave transformer at all odd-numbered harmonics.

In monolithic integration, the transmission line transformer is frequently much too long, and it may introduce significant losses. In many cases, it is perfectly acceptable to just use the third harmonic. This is shown in Figure 5.86(b). Here, a simple parallel LC circuit blocks the third harmonic ($3\omega_0$), while it acts as a short for all other harmonics and the fundamental frequency.

5.5 Oscillators

Oscillators are crucial components in almost any microwave system. Their fundamental task is to generate AC energy at a well-defined frequency from DC sources. A typical use of an oscillator is shown in the generic receiver block diagram of Figure 5.87, where it converts the input signal to a lower intermediate frequency. The mixer circuit, which is also needed for the frequency translation, will be discussed in the next section.

The class of oscillators discussed here has three aspects in common:

(i) a *resonator* to set the frequency of oscillation,
(ii) the generation of *instability* to allow the onset of oscillation, and
(iii) *amplitude control* to establish a stable amplitude in steady state.

Simple relaxation-type oscillators, such as found in simple digital timing circuits, will not be covered here.

5.5.1 Resonators – a brief overview

The resonator has the task of setting the oscillator's natural frequency.

The most common resonator types are

- lumped-element *LC resonators*, which come in either series or parallel resonance forms;
- *transmission-line resonators*;

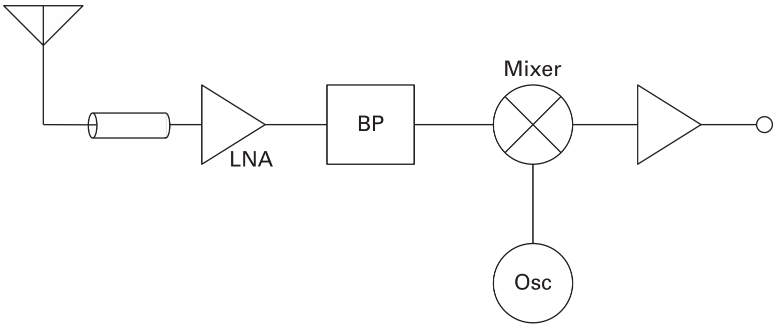


Fig. 5.87 Generic receiver block diagram.

- *cavity resonators* using waveguide elements;
- *dielectric resonators*, which use high ϵ_r ceramics and are typically combined with transmission line coupling structures;
- *quartz crystals* and similar devices which use the piezoelectric effect to derive electrical from a mechanical resonance – surface acoustic wave (SAW) and bulk acoustic wave (BAW) resonators also fall into this category.

Other resonator types, such as the magnetically tuned YIG (yttrium iron garnet) resonators, have only very limited use in speciality applications.

Critical aspects for resonators are

- the *quality factor*, which will be discussed in more detail below;
- the *reproducibility* of the resonant frequency – this can be a significant problem in BAW and SAW resonators;
- the *stability* of the resonant frequency against changes in temperature, mechanical shock and aging;
- the *tunability* of the resonant frequency – mostly established using variable capacitance elements.

For fixed-frequency oscillators, quartz crystals can still be considered to be the gold standard. Replacement of quartz resonators by elements which can be monolithically integrated is highly desirable and a hot research topic.

Quality factor

A very generic definition of the quality factor compares the stored and the dissipated energy in the resonator [20]:

$$Q = 2\pi \frac{\text{stored energy in the resonator}}{\text{dissipated energy during one cycle}}, \tag{5.214}$$

for $\omega = \omega_0$.

Let us consider RLC resonators (Figure 5.88) as an important example – via equivalent circuits, other resonator types can be transferred into RLC type resonators, at least in the immediate vicinity of the resonant frequency.

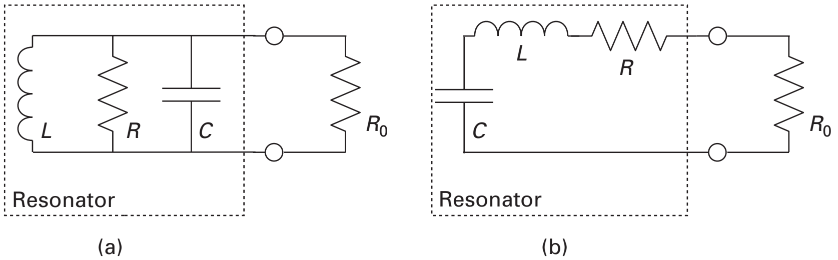


Fig. 5.88 RLC resonator circuits: (a) parallel topology and (b) series topology.

For the parallel resonant circuit (Figure 5.88(a)) at resonance $\omega = \omega_0 = 1/\sqrt{LC}$, the impedance is purely resistive and the dissipated energy during one cycle is

$$E_{\mathrm{diss}} = \frac{1}{2}\frac{\hat{I}^2 R}{\frac{\omega_0}{2\pi}} = \pi\frac{\hat{I}^2 R}{\omega_0}, \tag{5.215}$$

where $\hat{I}$ is the amplitude of the sinusoidal current flowing through the resonator.

The stored energy moves back and forth between the inductor and the capacitor; therefore, it suffices to calculate it for the capacitor:

$$E_{\mathrm{stored}} = \frac{1}{2}C\hat{V}^2 = \frac{1}{2}C\left(\hat{I}R\right)^2. \tag{5.216}$$

Inserting Equations (5.215) and (5.216) into (5.214) yields the Q factor for the parallel RLC resonator. This quality factor is called the *unloaded* Q because the loading resistor (R_0) has not been taken into account:

$$Q_{\mathrm{u}} = \omega_0 RC = \frac{R}{\omega_0 L} = \frac{R}{\sqrt{\frac{L}{C}}}. \tag{5.217}$$

The *loaded* Q is calculated by connecting R_0 in parallel to R:

$$Q_{\mathrm{l}} = \omega_0 C\frac{RR_0}{R+R_0} = \frac{Q_{\mathrm{u}}}{1+\frac{R}{R_0}}. \tag{5.218}$$

Similarly, we can calculate the unloaded Q for the series resonator (Figure 5.88):

$$Q_{\mathrm{u}} = \frac{\omega_0 L}{R} = \frac{\sqrt{\frac{L}{C}}}{R}, \tag{5.219}$$

and the loaded Q as

$$Q_{\mathrm{l}} = \frac{Q_{\mathrm{u}}}{1+\frac{R_0}{R}}. \tag{5.220}$$

5.5.2 Self-excitation criteria

Early in this chapter, we considered two-port stability from the viewpoint of avoiding parasitic oscillations in amplifiers. Now the task is to deliberately create instability in a certain frequency range.

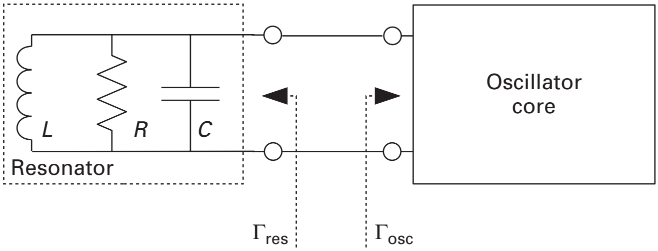


Fig. 5.89 Connection of a resonator to an oscillator core. A parallel RLC resonator is chosen as an example.

In the above resonator examples, the resonators always had a dissipative element associated with it. In practice, this is indeed always the case as resistive and radiative losses are never fully avoidable. In terms of the reflection coefficient seen looking into the resonator, this means

$$|\Gamma_{res}| < 1.$$

The stability boundary can be written as

$$\Gamma_{res}\Gamma_{osc} = 1, \tag{5.221}$$

where Γ_{osc} is the reflection coefficient looking into the oscillator core. Because the reflection coefficients are generally complex entities, Equation (5.221) has to be decomposed into a magnitude and a phase condition:

$$|\Gamma_{res}| \cdot |\Gamma_{osc}| = 1 \tag{5.222}$$

$$\angle(\Gamma_{res}) + \angle(\Gamma_{osc}) = 0. \tag{5.223}$$

The oscillator core will therefore necessarily have to provide $|\Gamma_{osc}| > 1$. Because we have seen in the Smith chart discussion that for all $\mathrm{Re}(Z) = 0, \ldots, \infty$, $|\Gamma| \leq 1$, this means that the real part of the oscillator core input impedance will have to be negative.

Another way of determining the proper conditions for oscillation is to use the Barkhausen self-excitation criterion. The block diagram for this discussion is shown in Figure 5.90. The system with positive feedback has the transfer function:

$$s(\omega) = \frac{1}{1 - F(\omega)}, \tag{5.224}$$

where $F(\omega)$ is the open loop gain. Obviously, the transfer function grows beyond all bounds – becomes unstable – for

$$F(\omega) = 1. \tag{5.225}$$

As $F(\omega)$ is a complex function, two conditions need to be fulfilled:

$$\mathrm{Re}\{F(\omega)\} = 1 \tag{5.226}$$

$$\mathrm{Im}\{F(\omega)\} = 0. \tag{5.227}$$

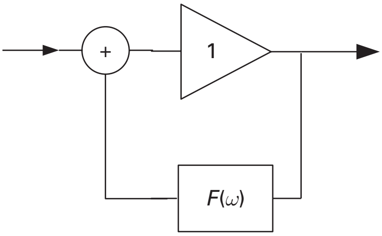


Fig. 5.90 Barkhausen self-excitation criterion: system model with positive feedback.

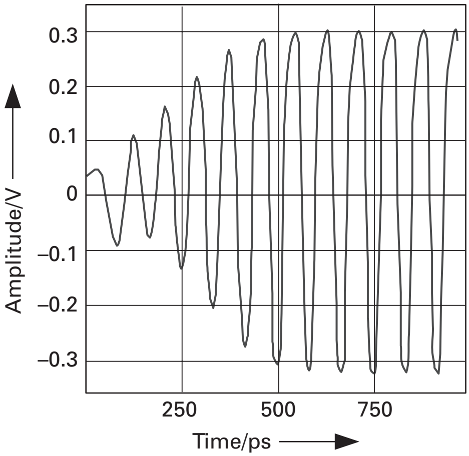


Fig. 5.91 Time-domain simulation of oscillator start-up behaviour.

5.5.3 Non-linearity in oscillators

The self-excitation criteria introduced so far assumed that the systems under investigation were all linear. In practice, however, this would not lead to the desired result of well-controlled sinusoidal signal generation.

Using the Γ criterion (Equation (5.222)), the initial $|\Gamma_{osc}|$ should be significantly (10–20%) higher than $|\Gamma_{res}|$ for the reliable onset of oscillation. However, the oscillation amplitude would then grow beyond all bounds, or in practice until it is limited by the supply voltage.

Fortunately, the active components we are using in the oscillator core to generate the negative resistance all exhibit gain saturation, i.e. the differential gain decreases with increasing signal amplitude. This leads to a self-stabilisation of the oscillation amplitude.

Figure 5.91 shows a time-domain simulation of oscillator start-up behaviour using a non-linear active device model (here, a Si/SiGe HBT). Note how the device linearity leads to a steady-state oscillatory behaviour after only a few cycles. Looking carefully, you will also notice that the initial frequency of oscillation is different from the one in steady state – this is caused by the reactive component of the oscillator core impedance, which also shows a non-linear behaviour and varies as the amplitude increases. The effect, which is undesirable, is frequently referred to as *chirp*.

5.5.4 Oscillator topologies

The negative resistance necessary to fulfil the condition $|\Gamma_{\mathrm{osc}}| > 1$ for the oscillator core can be generated in a variety of ways:

- The active devices in the oscillator core could have a negative differential resistance of its quasi-stationary I–V characteristics. Examples are tunnel diodes, Gunn diodes or exotic devices such as real-space transfer transistors. Except for Gunn diodes, which are still being used in inexpensive microwave modules (e.g. motion detectors), they have no longer (or never had) any commercial significance.
- The active device could incorporate a transit-time region which leads to a phase lag between the applied voltage and the current through the device. If this phase shift is larger than $\pi/2$ at a given frequency, the resulting impedance at that frequency has a negative real part. An example of such a device is the IMPact ionization Avalanche Transit-Time (IMPATT) diode, which is still being used extensively in millimetre-wave oscillators.
- The most common way to generate negative resistance is the use of positive feedback around an amplifying device.

The last item will be discussed in more detail here.

There are several ways of introducing positive feedback around an amplifier. Four of them are shown in Figure 5.92.

The configuration in Figure 5.92(a) uses magnetic coupling in a transformer around a common-source (or common-emitter) stage. As the common-source amplifier is inverting, a reversal of the winding sense in the transformer is necessary to generate the required positive feedback. Realised with vacuum tubes, this circuit was known very early in the history of radio and is called *Armstrong* (or Meissner) oscillator.

The *Hartley* topology (Figure 5.92(b)) uses a tapped-inductor feedback path around a non-inverting common-drain stage. The *Colpitts* oscillator (Figure 5.92(c)), uses a similar concept, but a capacitive voltage divider instead of the tapped inductor – it is therefore easier to realise in integrated form, and probably the most popular topology for MMIC implementations. When the inductor is replaced by a series LC circuit, the *Clapp* topology results (Figure 5.92(d)).

Due to its popularity, we will examine the Colpitts topology in more detail (Figure 5.93).

The current through capacitor C_2 is

$$i_{C2} = v_{\mathrm{gs}}\left[g_{\mathrm{m}} + \jmath\omega(C_1 + C_{\mathrm{gs}})\right], \tag{5.228}$$

which leads to the input voltage

$$v_{\mathrm{in}} = v_{\mathrm{gs}} + \frac{i_{C2}}{\jmath\omega C_2} = v_{\mathrm{gs}}\left(1 + \frac{C_1 + C_{\mathrm{gs}}}{C_2} - \jmath\frac{g_{\mathrm{m}}}{\omega C_2}\right). \tag{5.229}$$

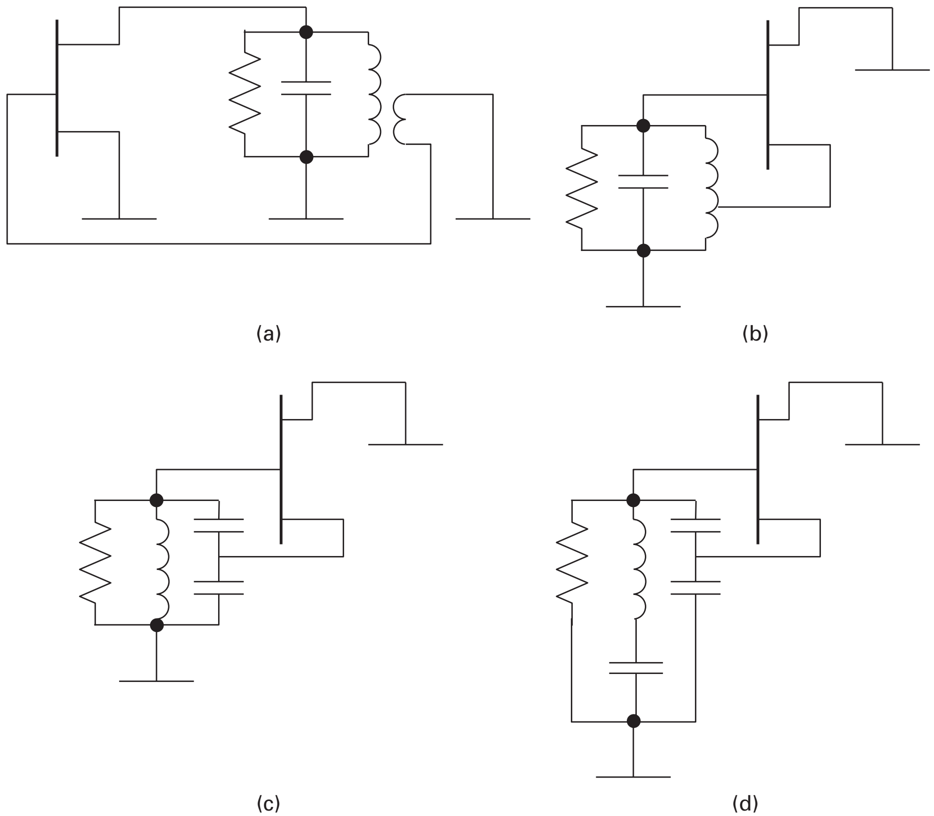


Fig. 5.92 Common oscillator topologies: (a) Armstrong or Meissner, (b) Hartley, (c) Colpitts and (d) Clapp.

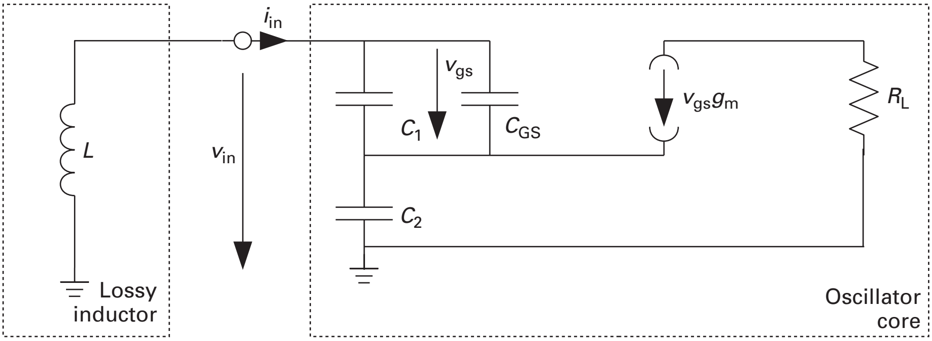


Fig. 5.93 Simple equivalent circuit of the Colpitts oscillator topology.

The input current is

$$i_{\text{in}} = v_{\text{gs}} \jmath\omega (C_1 + C_{\text{gs}}). \tag{5.230}$$

This allows us to calculate the input impedance of the oscillator core:

$$Z_{\text{in}} = \frac{v_{\text{in}}}{i_{\text{in}}} = -\frac{g_{\text{m}}}{\omega^2 C_2 (C_1 + C_{\text{gs}})} + \frac{1}{\jmath\omega} \frac{C_2 (C_1 + C_{\text{gs}})}{C_1 + C_2 + C_{\text{gs}}}. \tag{5.231}$$

The negative real part of the input impedance is obvious; the imaginary part is simply the series connection of the capacitances at the input.

Another aspect of the Colpitts oscillator is apparent: especially at high frequencies, the capacitor C_1 is not really necessary, the C_{gs} of the device suffices.

A very popular topology, especially for RF CMOS circuits, is the *cross-coupled pair*, shown in Figure 5.94(a). It is a differential pair with the gate of each transistor connected to the drain of the opposite transistor. Because of the 180° phase shift between both branches, this forces the small-signal gate voltages to be

$$v_{\text{gs1}} = -v_{\text{gs2}}. \tag{5.232}$$

Remember that in a differential pair under perfectly differential excitation, all nodes along the vertical centre plane are virtual grounds. Then the circuit behaviour can be completely described using the half-circuit in Figure 5.94(b). To maintain symmetry, C_0 has been replaced by the series connection of two capacitors of $2C_0$, while L_0 is divided into the series connection of two inductors with $L_0/2$.

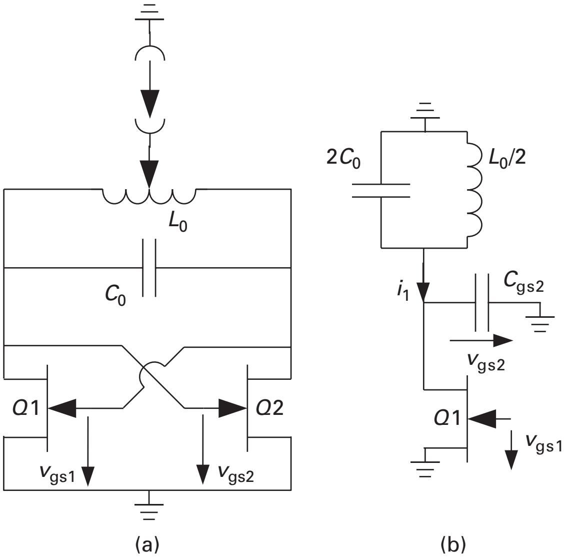


Fig. 5.94 (a) Simplified circuit of a cross-coupled pair oscillator and (b) its equivalent half-circuit.

Let us now calculate the admittance seen by the resonator towards the oscillator core:

$$\underline{Y}_1 = \frac{i_1}{v_{\mathrm{gs2}}} = \frac{g_{\mathrm{m}} v_{\mathrm{gs1}} + v_{\mathrm{gs2}} \jmath\omega C_{\mathrm{gs2}}}{v_{\mathrm{gs2}}} = -g_{\mathrm{m}} + \jmath\omega C_{\mathrm{gs2}}, \tag{5.233}$$

using Equation (5.232). The negative real part is hence $-g_{\mathrm{m}}$, and the oscillation frequency is

$$\omega_0 = \frac{1}{\sqrt{L_0(C_0 + C_{\mathrm{gs}})}}, \tag{5.234}$$

assuming $C_{\mathrm{gs1}} = C_{\mathrm{gs2}} = C_{\mathrm{gs}}$.

Finally, negative resistance can also be generated using common-gate (or common-base) configurations.

In the common-gate amplifier stage shown in Figure 5.95, note the inductance L_0 inserted into the gate lead. We will calculate the small-signal input impedance for this circuit. The input current is

$$i_1 = -v_{\mathrm{gs}}(g_{\mathrm{m}} + \jmath\omega C_{\mathrm{gs}}). \tag{5.235}$$

The current through L_0 is $v_{\mathrm{gs}} \jmath\omega C_{\mathrm{gs}}$, so the input voltage v_1 is (after a short calculation)

$$v_1 = -v_{\mathrm{gs}}(1 - \omega^2 L_0 C_{\mathrm{gs}}). \tag{5.236}$$

The input impedance is then, after separation into its real and imaginary parts,

$$\underline{Z}_1 = \frac{v_1}{i_1} = \frac{g_{\mathrm{m}}(1 - \omega^2 L_0 C_{\mathrm{gs}})}{g_{\mathrm{m}}^2 + \omega^2 C_{\mathrm{gs}}^2} - \jmath\omega C_{\mathrm{gs}} \frac{1 - \omega^2 L_0 C_{\mathrm{gs}}}{g_{\mathrm{m}}^2 + \omega^2 C_{\mathrm{gs}}^2}. \tag{5.237}$$

Now remember that the transit frequency $\omega_{\mathrm{T}} \approx g_{\mathrm{m}}/C_{\mathrm{gs}}$ and assume that

$$\frac{1}{L_0 C_{\mathrm{gs}}} \ll \omega^2 \ll \omega_{\mathrm{T}}^2.$$

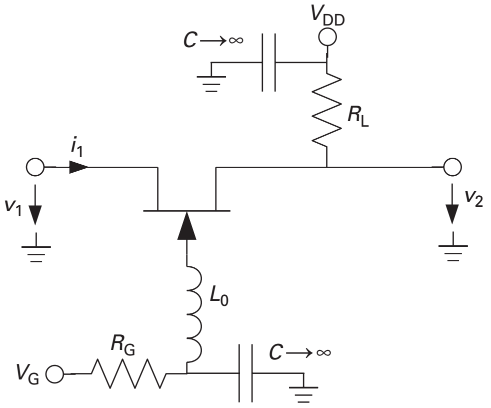


Fig. 5.95 Oscillator core using a common-gate stage.

Then, we obtain

$$\underline{Z}_1 = -\frac{\omega^2 L_0}{\omega_\mathrm{T}} + \jmath\omega L_0 \frac{\omega^2}{\omega_\mathrm{T}^2}. \tag{5.238}$$

Provided that the frequency is larger than the resonant frequency of L_0, C_gs, this circuit will, therefore, also generate a negative resistance. The imaginary part is inductive.

Voltage-controlled oscillators

Electronic control of the oscillation frequency is a very common requirement. While both variable inductance and variable capacitance concepts are in principle feasible, variable inductance approaches suffer from poor integrability. Attempts to realise integrated variable inductors using micro-electro-mechanical structures (MEMS) exist, but have not found practical use yet.

In practical applications, variable capacitors, in turn, are always realised using *varactor diodes*, which may build upon p–n junction diodes, Schottky diodes, or MOS diodes, depending on the underlying semiconductor technology. In each case, the capacitance of a blocking diode structure is varied by changing the voltage across the diode. Variable capacitors using MEMS have been investigated quite extensively, but again have not reached sufficient maturity for commercial applications so far.

The voltage-controlled oscillator (VCO) example in Figure 5.96 uses the cross-coupled differential pair topology introduced in Figure 5.94(a). The fixed capacitor has been replaced by two varactor diodes. The tuning voltage is connected to a virtual ground point, which facilitates decoupling between the RF and the DC control paths.

The series resistance of the varactor diodes should not be overlooked – it may substantially lower the overall resonator quality factor. The total quality factor can be shown to be

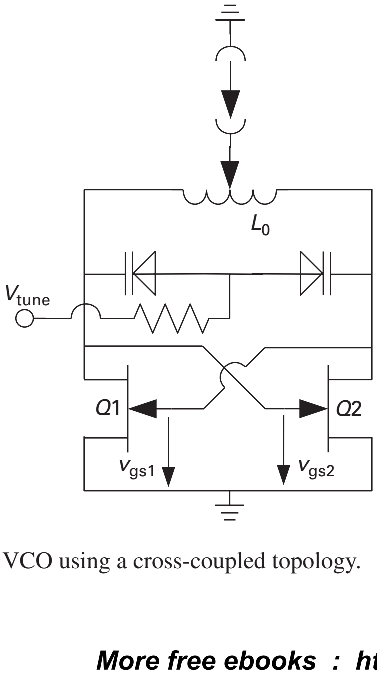


Fig. 5.96 VCO using a cross-coupled topology.

$$Q_{\text{total}} = \frac{Q_{\text{C}} Q_{\text{L}}}{Q_{\text{C}} + Q_{\text{L}}}, \tag{5.239}$$

where

$$Q_{\text{L}} = \frac{\omega_0 L}{R_{\text{L}}} = \sqrt{\frac{L}{C}} \frac{1}{R_{\text{L}}}, \; Q_{\text{C}} = \frac{1}{\omega_0 C R_{\text{C}}} = \sqrt{\frac{L}{C}} \frac{1}{R_{\text{C}}}$$

are the quality factors of an inductor with series resistance R_{L} and a capacitor with series resistance R_{C}, respectively, in a parallel resonant circuit. The assumption of weak losses has been made:

$$\frac{\omega_0^2 L^2}{R_{\text{L}}^2} \gg 1, \, \omega_0^2 R_{\text{C}}^2 C^2 \ll 1.$$

5.5.5 Noise in oscillators

The noise phenomena in active devices of course also affect oscillator performance. In principle, two things may happen:

- The oscillation amplitude may fluctuate randomly with time – *amplitude noise*.
- The phase of the sinusoidal signal may fluctuate randomly with time – *phase noise*.

Of the two, amplitude noise is the least critical. First of all, the gain compression effect which leads to a stable oscillation condition in the first place also reduces random amplitude fluctuations. Also, in frequency translation applications (frequency up- or down-conversion), oscillators typically work into switch-type mixers where the oscillator amplitude has little effect on the conversion efficiency, provided it is still sufficient for switching operation (see Section 5.6, p. 396).

For these reasons, we will restrict our discussions to phase noise, which has a much stronger impact on system performance.

Phase noise describes the random fluctuations of the oscillator phase with time. Mathematically,

$$s(t) = A \sin\left[\omega t + \phi(t)\right]. \tag{5.240}$$

In the frequency domain, these phase fluctuations are manifest in *noise sidebands* close to the carrier (see Figure 5.97). The spectral power density of the noise sidebands decreases with increasing distance to the carrier. They have a number of detrimental effects, for example

- in transmitters, they lead to interference in nearby channels;
- in receivers, phase noise increases the perceived in-channel noise due to a phenomenon known as *reciprocal mixing*, which is an intermodulation effect between a strong interferer and the local oscillator noise sidebands.

The customary figure of merit for phase noise suppression is the ratio between the carrier power and the power of the noise sidebands in a 1 Hz bandwidth at an offset Δf from the carrier. It is typically expressed in 'dBc/Hz'.

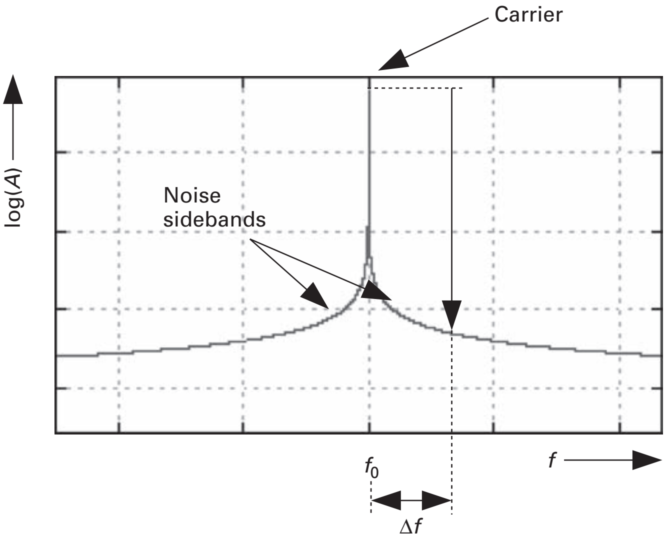


Fig. 5.97 Phase noise sidebands around a carrier in the frequency domain.

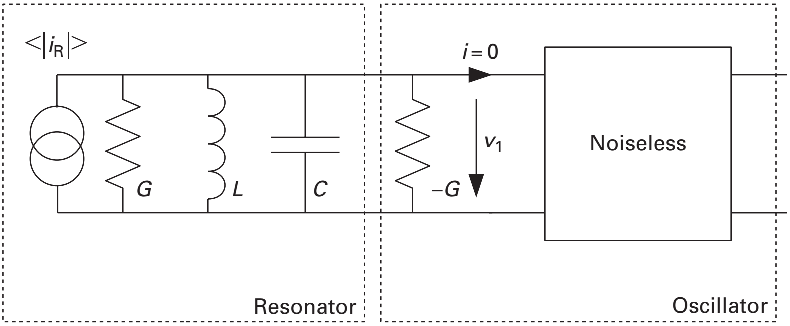


Fig. 5.98 Noise equivalent circuit of an ideal oscillator core with a lossy resonator.

To appreciate the effect of the resonator on the oscillator noise performance, consider Figure 5.98. The oscillator core itself is assumed to be noiseless (a strong over-simplification). The imaginary part of its input admittance is merged into the LC resonator; at the resonant frequency, the oscillator core presents a real part $-G$ which just offsets the resonator loss G.

The resonator dissipative element, G, creates a thermal noise current whose noise phasor is (see Equation (5.82))

$$\left\langle |i_\mathrm{R}|^2 \right\rangle = 8kTBG,$$

where B is the measurement bandwidth.

The voltage v_1 is $i_\mathrm{R} \cdot Z_\mathrm{res}$, where Z_res is the resonator impedance. Because the resonator losses are exactly compensated, it is the impedance of an ideal parallel LC resonator:

$$\underline{Z}_{\text{res}} = \frac{\jmath\omega L}{1-\omega^2 LC} = \frac{\jmath\omega L}{1-\frac{\omega^2}{\omega_0^2}}. \tag{5.241}$$

Taylor series expansion of the denominator, aborted after the linear term, leads to an approximate impedance for small deviations $\delta\omega$ from the resonance ω_0:

$$\underline{Z}_{\text{res}}(\omega_0 \pm \Delta\omega) \approx \mp\frac{\jmath\omega_0 L}{\frac{2\Delta\omega}{\omega_0}} = \mp\jmath\frac{\omega_0}{2\Delta\omega}\frac{1}{QG}, \tag{5.242}$$

using $Q = R/\omega_0 L = 1/(\omega_0 LG)$.

The phasor of the noise-generated voltage v_1 is then

$$\left\langle |v_1|^2 \right\rangle = \left\langle |i_{\text{R}}|^2 \right\rangle \left|\underline{Z}_{\text{res}}\right|^2 \approx \frac{8kTB}{GQ^2\left(\frac{2\Delta\omega}{\omega_0}\right)^2}, \tag{5.243}$$

using the approximation in Equation (5.242).

Comparing the noise power in bandwidth B to the signal power P_{S}, finally we obtain the phase noise suppression:

$$L(\Delta\omega) = \frac{2kTB}{P_{\text{S}}}\left(\frac{\omega_0}{2\Delta\omega Q}\right)^2. \tag{5.244}$$

Note the quadratic dependence of the phase noise suppression on the resonator Q.

In practical cases, the oscillator core is not noiseless. Furthermore, we need to take low-frequency noise phenomena into account, which lead to a stronger increase in phase noise close to the carrier. Leeson (1966) [26] introduced the following semi-empirical formula, which builds upon Equation (5.244):

$$L(\Delta\omega) = \frac{2FkTB}{P_{\text{S}}}\left[1+\left(\frac{\omega_0}{2\Delta\omega Q}\right)^2\right]\left(1+\frac{2\pi f_{\text{c}}}{|\Delta\omega|}\right). \tag{5.245}$$

The additional factors are F – accounting for the additional noise in the oscillator core – and f_{c}, the cutoff frequency for low-frequency ('$1/f$') noise.

Figure 5.99 shows an example calculation using Equation (5.245). We clearly distinguish three different regions:

(i) Close to the carrier, the noise power drops with -30 dB/decade. Here, the low-frequency noise increase in the oscillator core's active devices dominate.
(ii) Further out, the decay is -20 dB/decade, corresponding to the earlier calculations. Here, white noise sources (such as the thermal noise provided by the lossy resonator) dominate.
(iii) Far away from the carrier, a noise floor is visible, but this is generally not very relevant, unless F is very large.

If low phase noise oscillators are a requirement, both high quality factor resonators and active devices with low-frequency noise are a must. In general, bipolar devices (including HBTs) will fare much better than FETs. On-chip resonators generally have much lower Q than off-chip resonators can achieve.

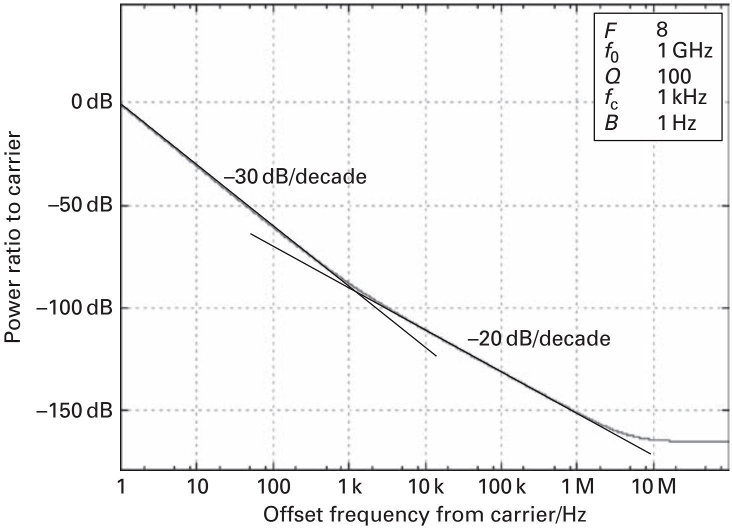


Fig. 5.99 Simulated phase noise sideband of an oscillator using the Leeson equation.

5.6 Mixers

Mixers are generally frequency-translation components, with a variety of applications in analogue signal processing, such as frequency shifting of signals (up and down conversion), phase shifting, modulation and demodulation. A special class of mixers – four-quadrant multipliers – can also be used in correlators, for example in impulse-radio ultra-wideband receivers.

The mathematics behind mixer operation has been reviewed already in the context of non-linear amplification (p. 370 and following). Recall that if we take a signal consisting of two sinusoidal components of different frequencies ω_1 and ω_2 and feed it into a non-linear two-port, the output $h(t)$ can be described by a Taylor series expansion:

$$h(t) = k_1\,[a\sin(\omega_1 t) + b\sin(\omega_2 t)] + k_2\,[a\sin(\omega_2 t) + b\sin(\omega_2 t)]^2 + \dots \tag{5.246}$$

The quadratic term expands to

$$\begin{aligned} &k_2\,[a\sin(\omega_2 t) + b\sin(\omega_2 t)]^2 \\ &= k_2\left[a^2\sin^2(\omega_1 t) + b^2\sin^2(\omega_2 t) + 2ab\sin(\omega_1 t)\sin(\omega_2 t)\right]. \end{aligned} \tag{5.247}$$

The product term can be expressed as

$$2ab\sin(\omega_1 t)\sin(\omega_2 t) = ab\,\{\cos[(\omega_1 - \omega_2)t] - \cos[(\omega_1 + \omega_2)t]\}\,. \tag{5.248}$$

Any non-linear system under two-tone excitation will therefore produce spectral components at the sum and difference of the input signals. We also conclude that an *analogue multiplier* would be the ideal mixer, because it *only* produces the sum and difference of the input spectral components.

Incidentally, the interaction used here for frequency mixing purposes, was called *second-order intermodulation* in the context of non-linear amplifiers.

5.6.1 Transconductance multiplier

A simple analogue multiplier can be realised using a bipolar differential amplifier. It utilises the fact that in bipolar transistors, the small-signal transconductance is linearly dependent on the large-signal collector current in the operating point. Now take a differential pair Q_1, Q_2 with a common current source transistor Q_3 (Figure 5.100). The RF input voltage v_1 is fed differentially into the top transistor pair (superimposed on the bias voltage V_0), while the oscillator current signal, I_{LO}, is fed single-endedly into the base of Q_3. Let Q_3 have a current gain B. Then, the differential output signal is

$$v_2 = v_1 g_{\mathrm{m}} R_{\mathrm{L}} = v_1 I_{\mathrm{LO}} \frac{B R_{\mathrm{L}}}{2V_{\mathrm{T}}}, \tag{5.249}$$

because

$$g_{\mathrm{m}} = \frac{B I_{\mathrm{LO}}}{2V_{\mathrm{T}}}.$$

The output voltage is therefore proportional to the product of the input voltage and the local oscillator current.

This simple circuit, however, has a number of drawbacks. First of all, it will only work for $I_{\mathrm{LO}} > 0$. Secondly, the input voltage has to be much smaller than V_{T} (or 26 mV at room temperature) to fulfil the small-signal assumption. Finally, it will only work with bipolar transistors.

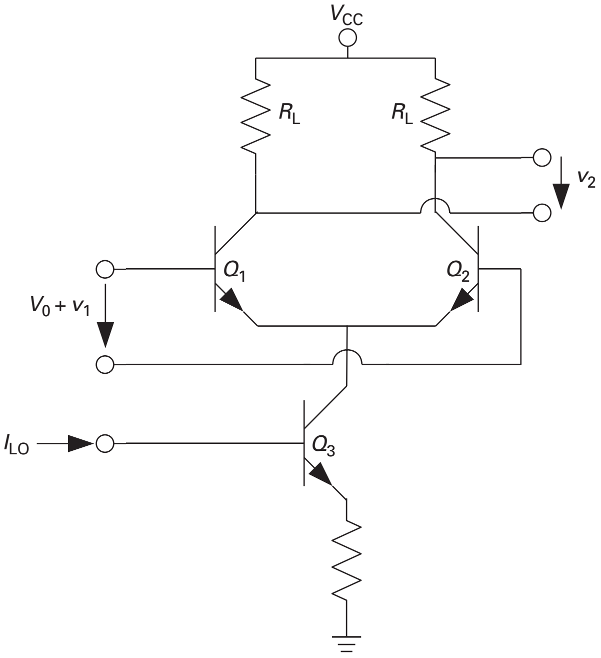


Fig. 5.100 Transconductance multiplier circuit.

5.6.2 Resistive mixer

After looking at a mixer circuit which works only with bipolar transistors, let us briefly consider a circuit which will work only with FETs. To appreciate its mode of operation, examine the output I–V characteristics of a FET (see Figure 5.101).

At low V_{DS}, the relationship between I_{D} and V_{DS} is an approximately linear one and can therefore be accurately described by the channel conductance g_{DS}. Assuming a simple Statz–Curtice model for the FET:

$$I_{\mathrm{D}}\,(V_{\mathrm{GS}}, V_{\mathrm{DS}}) = \frac{\beta(V_{\mathrm{GS}} - V_{\mathrm{P}})^2}{1 + \alpha(V_{\mathrm{GS}} - V_{\mathrm{P}})} \tanh(\gamma V_{\mathrm{DS}}),$$

where V_{P} is the pinch-off or threshold voltage and α, β and γ are model parameters, we find

$$g_{\mathrm{DS}} = \gamma \frac{\beta(V_{\mathrm{GS}} - V_{\mathrm{P}})^2}{1 + \alpha(V_{\mathrm{GS}} - V_{\mathrm{P}})} \left[1 - \tanh^2(\gamma V_{\mathrm{DS}})\right] \approx \gamma \frac{\beta(V_{\mathrm{GS}} - V_{\mathrm{P}})^2}{1 + \alpha(V_{\mathrm{GS}} - V_{\mathrm{P}})}, \qquad (5.250)$$

provided that $\tanh^2(\gamma V_{\mathrm{DS}}) \ll 1$ or $\gamma V_{\mathrm{DS}} < 0.3$.

Now the FET is placed in a circuit as shown in Figure 5.102. The inductor L enforces a steady-state bias point $V_{\mathrm{DS},0} = 0$, but is invisible at the LO, RF or IF frequencies. The gate bias can be set to a suitable V_{GS} for optimum mixer operation. The FET, together with the generator resistance of the RF port R_{RF}, forms a resistive voltage divider whose division ratio is modulated by the gate potential:

$$V_{\mathrm{DS}}(t) = \frac{V_{\mathrm{RF}}(t)}{1 + g_{\mathrm{DS}}(t) R_{\mathrm{RF}}}. \qquad (5.251)$$

The periodic variation of g_{DS} by the gate potential $V_{\mathrm{GS}} = V_{\mathrm{GS},0} + V_{\mathrm{LO}} \sin(\omega_{\mathrm{LO}} t)$:

$$g_{\mathrm{DS}} = \beta\gamma \left[V_{\mathrm{GS},0} - V_{\mathrm{P}} + V_{\mathrm{LO}} \sin(\omega_{\mathrm{LO}} t)\right]^2 \qquad (5.252)$$

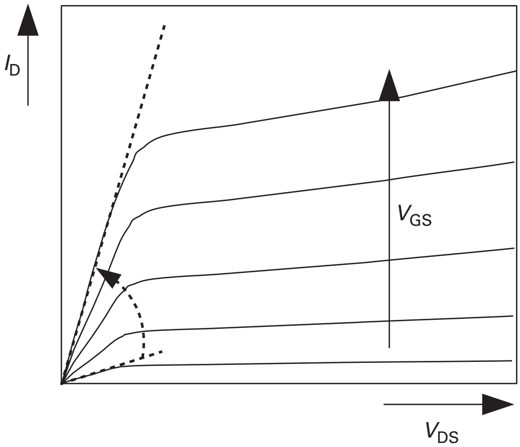


Fig. 5.101 FET output I–V characteristics with indication of variable resistor operation.

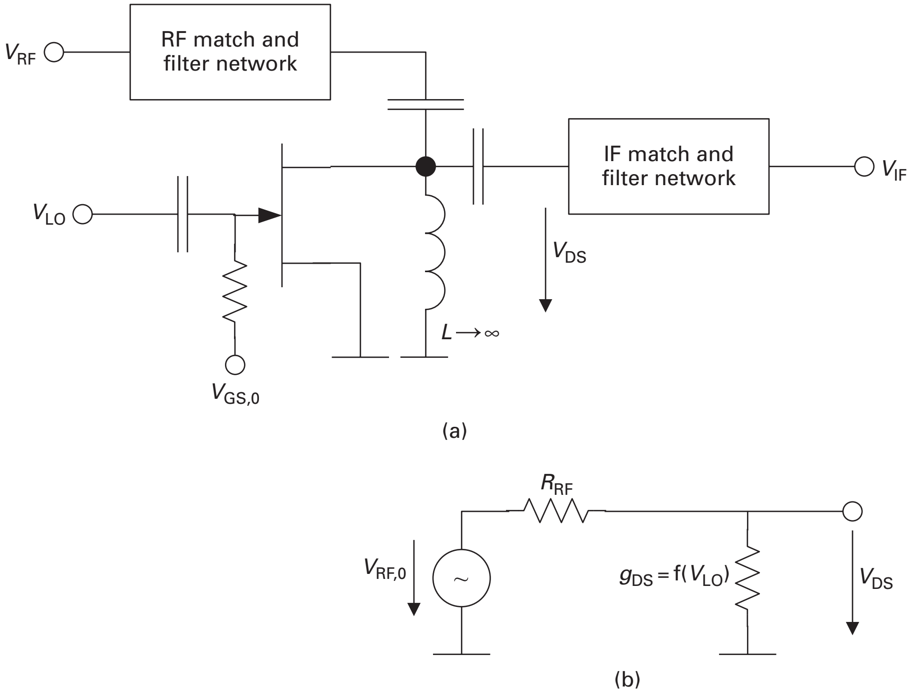


Fig. 5.102 (a) Resistive mixer configuration and (b) its simplified equivalent circuit.

results in the desired mixing action (pure square law behaviour, i.e. $\alpha = 0$, is assumed here, see Equation (5.250)). Note that Equation (5.252) only holds for $V_{LO} < V_{GS,0} - V_P$.

Frequently, the local oscillator voltage will be chosen such that $V_{LO} \geq V_{GS,0} - V_P$ – the channel conductance is then switched between two saturated states, $g_{DS} = 0$ and a high state essentially limited by the source and drain series resistances, which were initially omitted in the simplified discussion. Under this condition, and assuming $V_{RF}(t) = V_{RF}\sin(\omega_{RF}t)$, the time-dependent drain-source voltage becomes

$$V_{DS}(t) = \frac{V_{RF}\sin(\omega_{RF}t)}{R_S + R_D + R_{RF}}\left[R_S + R_D + R_{RF}\mathrm{rect}(\omega_{LO}t)\right]. \tag{5.253}$$

Since the Fourier series of the rect function is

$$\mathrm{rect}(\omega_{LO}t) = \frac{1}{2} + \frac{2}{\pi}\left[\sin(\omega_{LO}t) + \frac{\sin(3\omega_{LO}t)}{3} + \ldots\right], \tag{5.254}$$

this switching mode operation leads to the desired multiplication with the fundamental frequency of the local oscillator, but also with higher-order harmonics – the latter operation is referred to as *sub-harmonic pumping*.

Figure 5.103 shows as an example the measured conversion gain of a resistive mixer, as a function of the local oscillator power. At low P_{LO}, the conversion gain

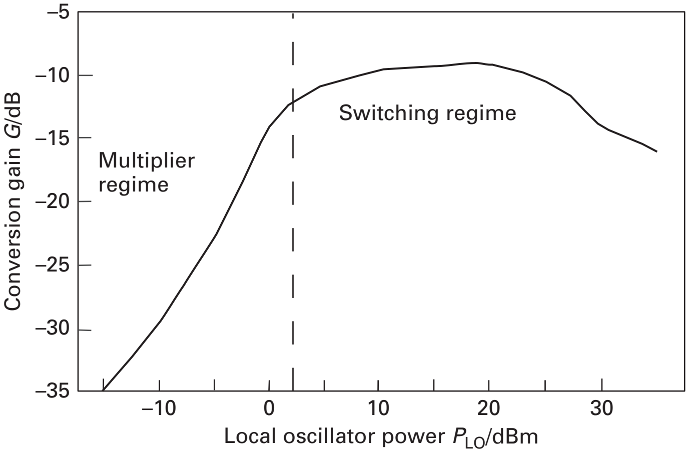


Fig. 5.103 Measured conversion gain of a resistive mixer.

increases proportionally with the local oscillator power – the mixer acts like a multiplier. For higher P_{LO}, the mixer enters the switching regime – the conversion gain is roughly independent of the local oscillator power. This behaviour can be seen, in variations, for any of the mixer circuits discussed here. In standard wireless systems, mixers are operated in the switching regime, because independence from local oscillator power fluctuations is highly desirable. For example, it reduces the effect of oscillator amplitude noise. There are applications, however, where operation in the multiplier regime is wanted. An excellent example are correlation receivers for impulse-radio ultra-wideband systems, where the formation of a cross-correlation between the received pulse and a template pulse in the receiver calls for a true multiplier.

The resistive mixer has the advantage of high simplicity, and, most prominently, very high linearity with respect to the RF port. The latter is especially true when local oscillator power leaking through the gate-drain capacitance is short-circuited to ground, so that the drain potential is not modulated at the LO frequency [27].

Disadvantages are the significant required local oscillator voltage swing and the conversion loss inherent to the voltage divider principle.

An important issue for practical mixers is *port isolation* – ideally, power fed into the LO port should not be present at the RF and IF ports, and power fed into the RF port should not be present at the IF port in downconverters, while power fed into the IF port should not leak to the RF port in upconverters. In the resistive mixer, the isolation between the RF and IF ports is realised by filters only, which is a significant disadvantage. The LO-to-RF and LO-to-IF isolations are somewhat better because the leakage path is via the gate-drain capacitance, but it may still be too high.

Mixer concepts where port isolation is assisted by destructive interference are much better, in this respect. They will be treated next.

5.6.3 Single-balanced mixer

Consider the circuit in Figure 5.104. Transistor Q_1 is a common-source amplifier stage fed by the RF signal. Its load is formed by the differential pair Q_2, Q_3 with load resistances R_L. R_2 and R_3 are for biasing only. Q_2 and Q_3 are driven by the local oscillator – but due to the balun transformer,[7] their gate signals are exactly 180° out of phase.

Assume that the local oscillator signal is large enough so that the Q_2 and Q_3 are being switched off alternately. Then, the differential voltage V_IF can be written as

$$V_\mathrm{IF} = R_\mathrm{L} I_\mathrm{D1} \left[2 \cdot \mathrm{rect}(\omega_\mathrm{LO} t) - 1\right], \tag{5.255}$$

where I_D1 is the drain current of Q_1. Provided that Q_1 is operated in the small-signal regime, the small-signal IF output voltage is, considering the fundamental frequency component of the local oscillator signal only,

$$V_\mathrm{IF} = \hat{V}_\mathrm{RF} \frac{4 g_{\mathrm{m},1} R_\mathrm{L}}{\pi} \sin \omega_\mathrm{LO} \sin \omega_\mathrm{RF} t, \tag{5.256}$$

using Equation (5.254) and $V_\mathrm{RF} = \hat{V}_{LO} \sin(\omega_\mathrm{LO} t)$.

The desired multiplication is again visible, producing spectral components at $\omega_\mathrm{RF} \pm \omega_\mathrm{LO}$.

The mixer uses a special property already discussed in the context of differential amplifiers: the common-source connection of transistors Q_1 and Q_2 is a virtual

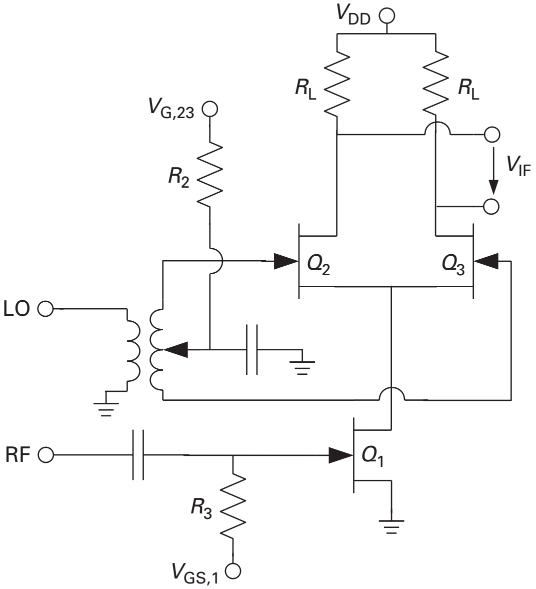


Fig. 5.104 Single-balanced mixer circuit.

[7] Note that baluns are drawn as transformers in this and the following circuit diagrams. At micro- and millimetre-wave frequencies, they are rarely transformers, but may be realised using transmission line segments, or as active baluns.

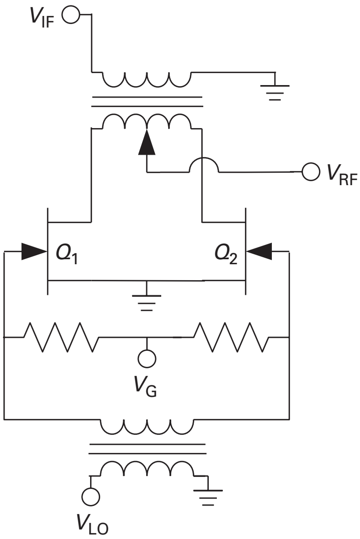


Fig. 5.105 Single-balanced upconversion mixer with FETs in shunt configuration.

ground for purely differential excitation. Therefore, the local oscillator and RF ports are highly decoupled by virtue of destructive interference. LO-to-RF leakage is very critical, because the local oscillator signal must not radiate via the antenna.

The circuit shown in Figure 5.105 is also a single-balanced mixer, used to upconvert a signal at an intermediate frequency to a higher frequency (RF). As opposed to the circuit shown in Figure 5.104, where the switches were in series configuration, they are in shunt here, alternately connecting the two ends of the top balun to ground, periodically changing the sign of the IF signal at the RF port and thus producing the desired multiplication of the LO signal with IF. As upconversion is what we desire, a subsequent filter will have to suppress the difference frequency and pass only the sum. Because the RF port is connected to the virtual ground connection with respect to fully differential excitation of Q_1 and Q_2, destructive interference of potential LO leakage at the IF port will again ensure a very high LO-to-RF isolation.

5.6.4 Double-balanced mixer

The configuration shown in Figure 5.104 is a building block of the double-balanced mixer, probably the most popular active mixer configuration in MMIC design today. These topologies are commonly referred to as *Gilbert cells* [12].

In the double-balanced mixer (Figure 5.106), all signals have to be applied in a differential fashion – this is indicated by the presence of three baluns. Transistors Q_1 and Q_2 form a differential amplifier, connected to another pair of differential amplifiers Q_3-Q_6. In the original publication by Gilbert, the configuration is operated as a true four-quadrant multiplier, while in most applications, the top four transistors are switched by the LO signal between two states – they are jointly referred to as the *switching quad*. Only the latter mode shall be discussed here. In this case, the double-balanced

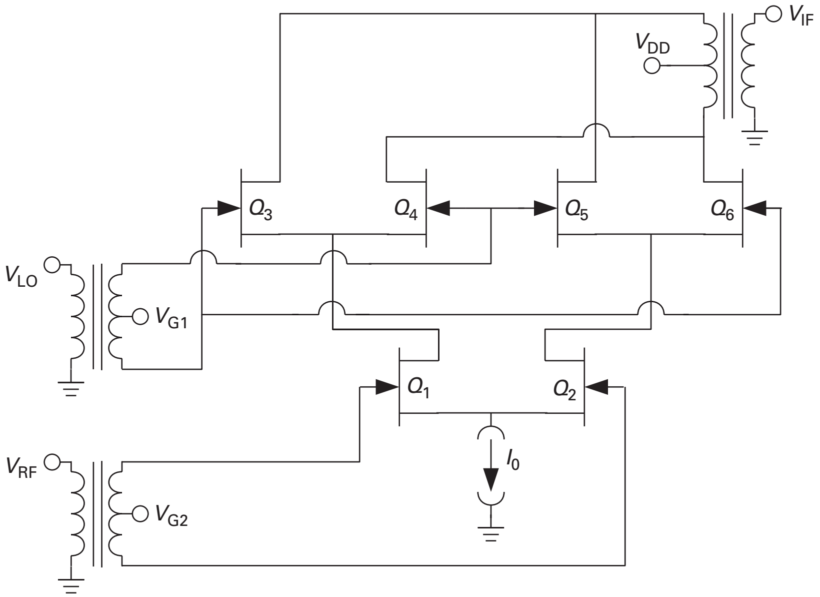


Fig. 5.106 Double-balanced mixer with Gilbert cell topology.

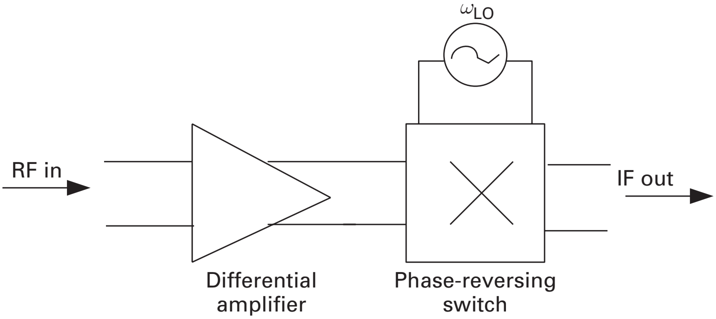


Fig. 5.107 Block diagram of a double-balanced mixer of the Gilbert cell type, partitioned into a differential amplifier and a phase-reversing switch.

mixer can be viewed as the cascade of a differential amplifier and a phase-reversing switch, such that the amplified RF signal changes its phase by 180° with the period of the LO signal. This is shown schematically in Figure 5.107.

The big advantage of the double-balanced mixer is that all ports are now decoupled by destructive interference. For example, the differential RF signal leakage cancels out at both joint gate connections of the switching quad. Likewise, the LO leakage cancels at the source connections of Q_3, Q_4 and Q_5, Q_6, respectively.

A disadvantage of the circuit shown in Figure 5.106 is the rather large voltage headroom required. On top of the required minimum drain-source voltage of the switching quad and the differential amplifier Q_1, Q_2, additional headroom is required for the current source I_0. The latter is frequently replaced by a parallel resonant circuit, tuned to

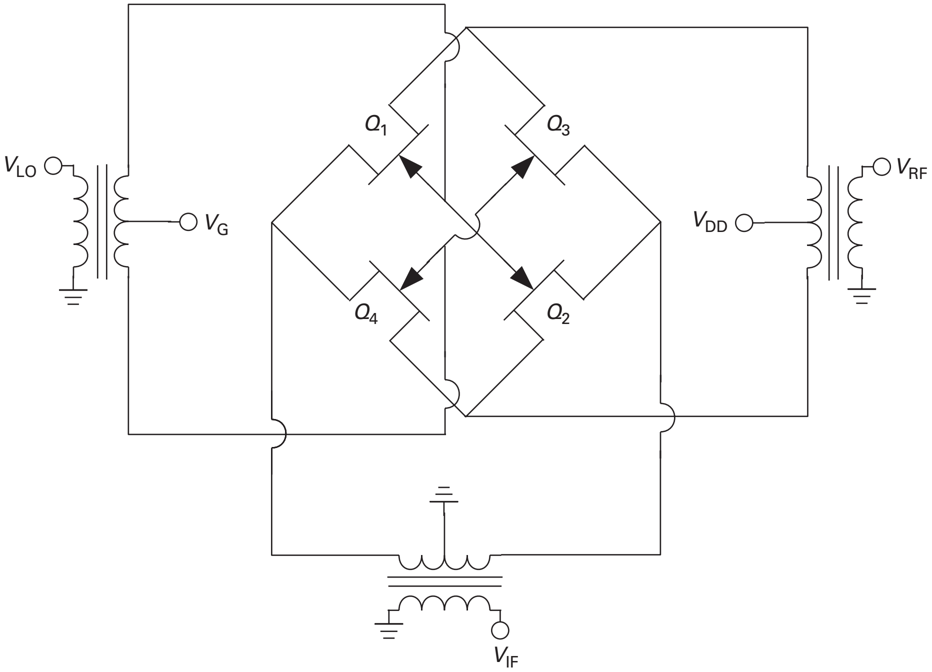


Fig. 5.108 Double-balanced mixer with a ring topology.

ω_{RF}. It provides the required high impedance for the RF signal, but provides a zero DC resistance, lowering the supply voltage requirements.

Double-balanced mixers can also be realised using resistive mixer principles (see Section 5.6.2). This has the advantage of very low power consumption, but of course the circuit will not be able to produce any conversion gain.

Figure 5.108 shows an example. Depending on the choice of the drain potential V_{DD}, this circuit can be operated as a resistive mixer (V_{DS} low, in the linear regime), or the transistors may act as current switches (V_{DS} in the saturated region). The RF signal is applied to the drains of the transistors. The LO signal alternately turns on transistors Q_1, Q_2 and Q_3, Q_4, leading to a periodic phase reversal of RF signal at the IF port.

5.6.5 Micromixer

The micromixer concept, shown in Figure 5.109, evolved from the Gilbert multiplier topology and was also published by B. Gilbert [13]. The transistors $Q_4 - Q_7$ are the switching quad, as before – the circuit is drawn using bipolar transistors here, but would work as well with FETs. The topology is somewhat simplified, for example the baluns are not included.

The local oscillator signal is again applied in differential format. The RF signal, however, is single-ended and applied to the amplifier structure formed by the transistors $Q_1 - Q_3$. Q_1 is a common-base amplifier stage and Q_3 a common-emitter stage. Because the former provides a non-inverting and the latter an inverting voltage gain,

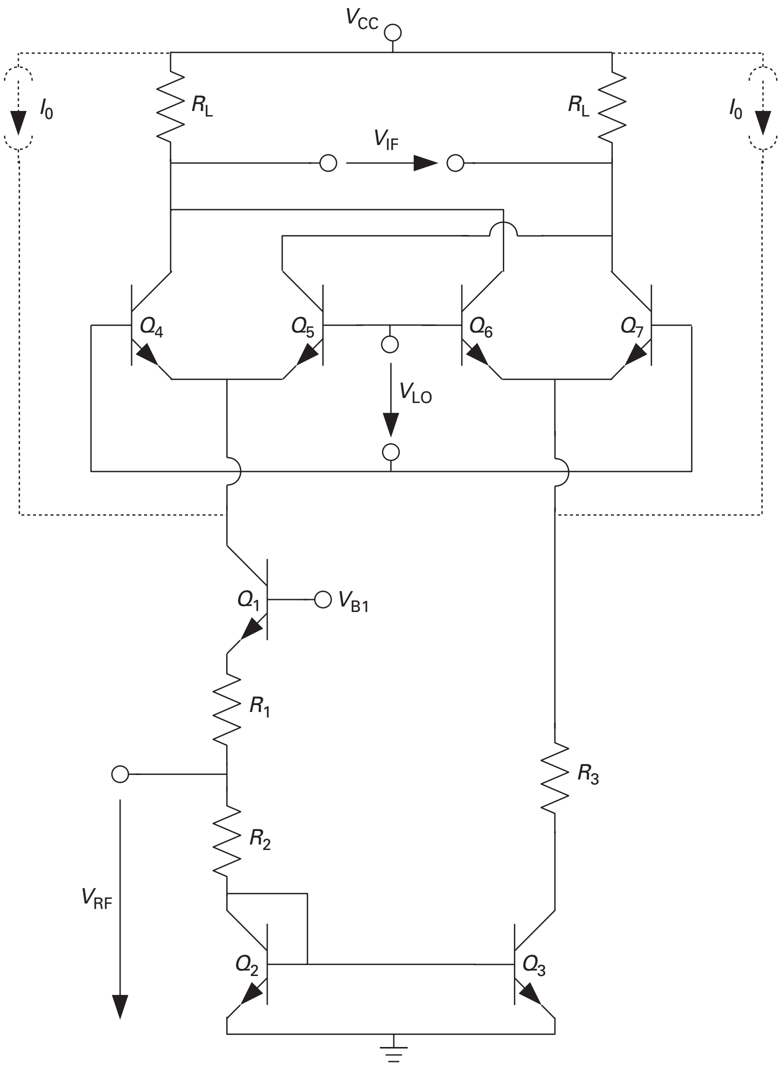


Fig. 5.109 Simplified micromixer topology using bipolar transistors.

the signals fed into the two branches of the switching quad are 180° out of phase – the circuit doubles as an single-ended-to-differential converter. Q_2 forms a current mirror with Q_3 such that the currents in both branches are equal.

Frequently, the gain of the common-base/common-source amplifier/balun is increased by injecting additional current into the two branches. This is indicated in Figure 5.109 by the two dashed current sources I_0.

Compared to the original Gilbert multiplier, the input stage of the micromixer can be made more linear, and can be designed to achieve a broadband match, using the resistors R_1 and R_2.

5.6.6 Diode-based mixers

While the circuit design descriptions in this chapter (and, for that matter, the whole book) emphasise concepts with active components, the use of diodes in mixers must be mentioned here because they are still very commonly employed, especially at millimetre-wave frequencies where transistors lose their usefulness. The diode of choice is the Schottky (metal–semiconductor) diode due to the absence of carrier storage effects when switching from forward to reverse bias.

As in the case of operating a FET as a resistive or switching mixer, the principle is the change of the differential resistance. For a diode, this is shown in Figure 5.110. Clearly, the differential diode conductance $g_D = dI_D/dV_D$ is very low for bias point A (or equally for $V_D < 0$) and very high for bias point B. We will now use the local oscillator again to periodically change the diode between these two states.

Consider Figure 5.111, which depicts the very popular *diode ring mixer*. The circuit is very similar to the FET ring mixer shown earlier (Figure 5.108). Assume that the RF signal is always much smaller than the local oscillator (LO) signal. Then, the diode state will depend on the applied LO signal only. Either the left or the right diode pair may conduct and exhibit a high differential conductance. The diode ring acts as a phase-reversing switch, which connects the RF signal to the intermediate frequency (IF) load with periodically alternating polarity.

Similarly, single-balanced and single-ended (unbalanced) mixer topologies can be realised using diodes. This shall not be further expanded here.

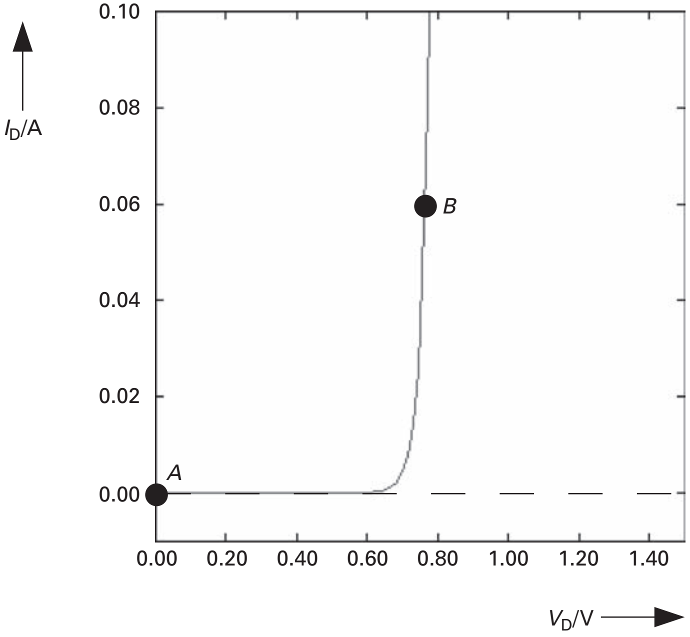


Fig. 5.110 Example Schottky diode I–V characteristics, with bias point indicated for small (A) and high (B) differential conductance.

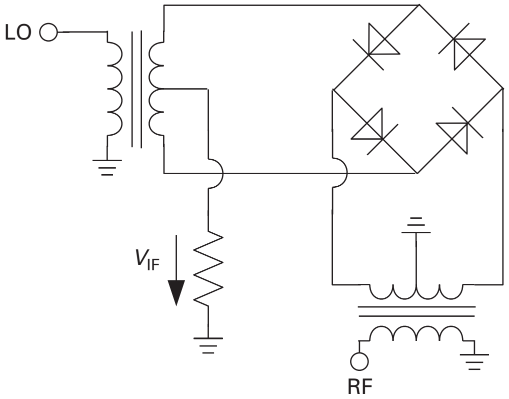


Fig. 5.111 Diode ring mixer.

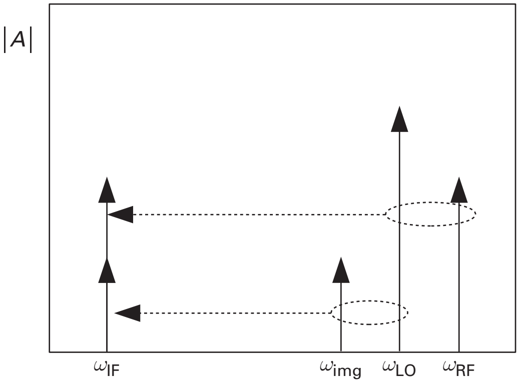


Fig. 5.112 The image frequency problem in downconversion mixing: ω_{RF} and ω_{img} both produce components at ω_{IF} when mixed with ω_{LO}.

5.6.7 Image-rejection mixer topologies

One problem inherent to any of the mixers discussed here still needs to be mentioned. When operated as a downconverter (converting an RF signal at a higher frequency to a lower IF), it uses the fact that the multiplication of sinusoidal signals produces a spectral component at the difference of the two initial frequencies, as was discussed in Equation (5.248).

$$2\sin(\omega_1 t)\sin(\omega_2 t) = \cos\left[(\omega_1 - \omega_2)t\right] - \cos\left[(\omega_1 + \omega_2)t\right].$$

Consider now a case where $\omega_1 = \omega_{RF}$ is the RF signal and $\omega_2 = \omega_{LO}$ is the LO frequency, which is lower than ω_{RF}. The difference is the IF: $\omega_{RF} - \omega_{LO} = \omega_{IF}$. If, however, there is another signal present at the RF input, with a frequency $\omega_{img} = \omega_{RF} - 2\omega_{IF}$, it will *also* mix with ω_{LO} to produce a spectral component at the right IF: $\omega_{LO} - \omega_{img} = \omega_{IF}$. The problem is schematically shown in Figure 5.112. ω_{img} is called the *image frequency*.

The signal at the image frequency will now overlay the downconverted RF signal and in most cases cause unacceptable interference. The most common way to avoid this is to make sure, by appropriate filtering, that no signal is in fact present at ω_{img}.

In upconversion mixers, the image problem also applies. Upconverting a signal at ω_{IF} by means of an LO at a higher frequency ω_{RF} results in two spectral components at $\omega_{\text{LO}} \pm \omega_{\text{IF}}$. Again, one of these components will have to be suppressed, classically by filtering.

However, in modern receiver and transmitter concepts, there is a trend to lower IFs where digital/analogue conversion can be achieved very inexpensively. Consequently, the frequency distance $2\omega_{\text{IF}}$ between the desired signal and the image is getting smaller, requiring very rigid filtering. The necessary filter qualities are rarely possible on chip, which is another significant drawback, as off-chip filters are costly in production and assembly.

A more convenient solution is the use of an *image-rejection* mixer. The image-rejection mixer topology can use any of the fundamental mixer circuits we discussed.

The block diagram shown in Figure 5.113 shows a possible implementation for a downconversion mixer. At the input, two signals shall be present, at the RF and image frequencies:

$$\omega_{\text{RF}} = \omega_{\text{LO}} + \omega_{\text{IF}}, \omega_{\text{img}} = \omega_{\text{LO}} - \omega_{\text{IF}}$$

The signal is split into two signals with equal amplitude, but 90° phase shift (*in quadrature*). We assume that this can be done equally for the RF and the image signals – this is a good assumption because the concept is especially important if RF and image signals are located in close spectral proximity and cannot be separated easily by filtering. For simplicity, we assume also that both RF and image signals are sinusoidal with an amplitude of 1.

The oscillator signal, at ω_{LO}, shall also have an amplitude of 1. It is split into two signals with equal amplitude and phase.

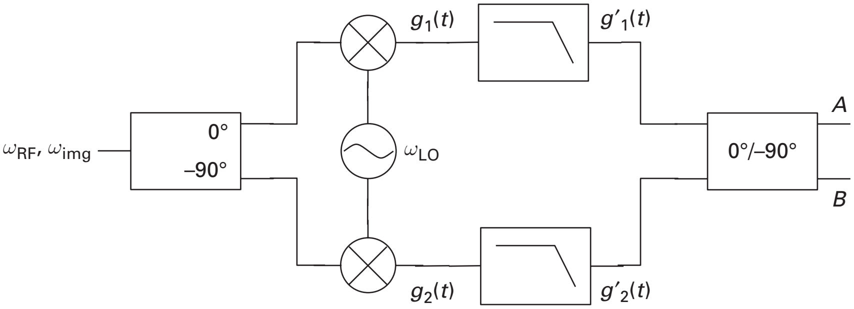


Fig. 5.113 Block diagram of an image reject mixer.

Let us first consider the RF signal only (the image signal is turned off). The signal at the output of the top mixer is

$$\begin{aligned} g_1(t) &= \frac{1}{4}\sin(\omega_{\mathrm{RF}}t)\sin(\omega_{\mathrm{LO}}t) \\ &= \frac{1}{8}\left\{\cos[(\omega_{\mathrm{RF}}-\omega_{\mathrm{LO}})t] - \cos[(\omega_{\mathrm{RF}}+\omega_{\mathrm{LO}})t]\right\}. \end{aligned}$$

The high frequency component at $\omega_{\mathrm{RF}}+\omega_{\mathrm{LO}}$ is easily suppressed in the low-pass filter (LP). After the low-pass filter in the upper branch,

$$g_1'(t) = \frac{1}{8}\cos[(\omega_{\mathrm{RF}}-\omega_{\mathrm{LO}})t] = \frac{1}{8}\cos(\omega_{\mathrm{IF}}t) = \frac{1}{8}\sin\left(\omega_{\mathrm{IF}}t + \frac{\pi}{2}\right). \tag{5.257}$$

In the lower branch, the RF signal is delayed by $-\pi/2$.

Using $\sin(\omega_{\mathrm{RF}} - \pi/2) = -\cos(\omega_{\mathrm{RF}}t)$, we obtain at the output of the bottom mixer:

$$\begin{aligned} g_2(t) &= -\frac{1}{4}\cos(\omega_{\mathrm{RF}}t)\sin(\omega_{\mathrm{LO}}t) \\ &= -\frac{1}{8}\left[\sin(\omega_{\mathrm{RF}}+\omega_{\mathrm{LO}})t - \sin(\omega_{\mathrm{RF}}-\omega_{\mathrm{LO}})t\right]. \end{aligned}$$

The high frequency component is removed in the low-pass filter:

$$g_2'(t) = \frac{1}{8}\sin(\omega_{\mathrm{IF}}t). \tag{5.258}$$

The four-port at the right of the block diagram in Figure 5.113 is a 90° *hybrid coupler*. Signals fed into the inputs emerge at outputs A and B with equal amplitude, but the signal entering at the top experiences an additional $-\pi/2$ phase shift at output B, while the signal entering at the bottom experiences an additional phase shift of $-\pi/2$ at output B.

Using Equations (5.257) and (5.258), we then find at A:

$$\frac{1}{8}\left[\sin\left(\omega_{\mathrm{IF}}t + \frac{\pi}{2}\right) + \sin\left(\omega_{\mathrm{IF}}t - \frac{\pi}{2}\right)\right] = 0.$$

At B:

$$\frac{1}{8}\left[\sin(\omega_{\mathrm{IF}}t) + \sin(\omega_{\mathrm{IF}}t)\right] = \frac{1}{4}\sin(\omega_{\mathrm{IF}}t).$$

The IF signal due to the wanted (RF) signal hence only appears at output B.

Let us consider the image frequency, which is below ω_{LO} such that $\omega_{\mathrm{IF}} = \omega_{\mathrm{LO}} - \omega_{\mathrm{img}}$. Using $\sin(-\alpha) = -\sin\alpha$ and $\cos(-\alpha) = \cos\alpha$, we find at the output of the upper low-pass filter:

$$g_1'(t) = -\frac{1}{8}\sin\left(\omega_{\mathrm{IF}}t - \frac{\pi}{2}\right), \tag{5.259}$$

while at the output of the lower low-pass filter:

$$g_2'(t) = -\frac{1}{8}\sin(\omega_{\mathrm{IF}}t). \tag{5.260}$$

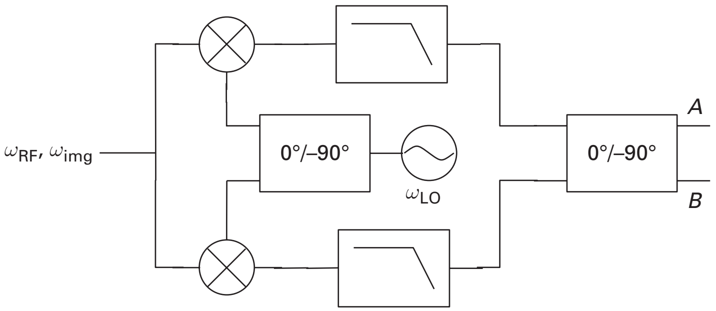


Fig. 5.114 Image reject mixer topology with quadrature LO.

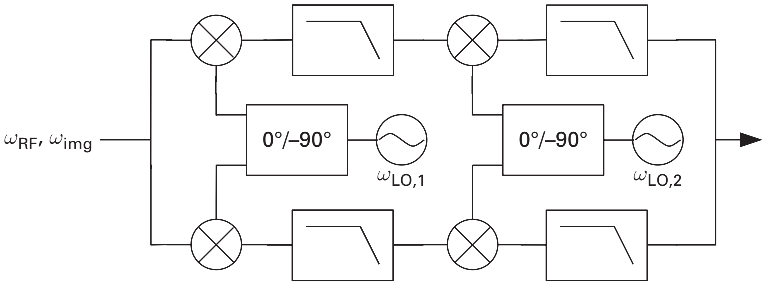


Fig. 5.115 A 'third method' image-rejection mixer topology using two mixing steps.

These signals combine at output *A* as

$$-\frac{1}{8}\left[\sin\left(\omega_{\mathrm{IF}}t - \frac{\pi}{2}\right) + \sin\left(\omega_{\mathrm{IF}}t - \frac{\pi}{2}\right)\right] = -\frac{1}{4}\sin\left(\omega_{\mathrm{IF}}t - \frac{\pi}{2}\right).$$

At *B*:

$$-\frac{1}{8}\left[\sin(\omega_{\mathrm{IF}}t - \pi) + \sin(\omega_{\mathrm{IF}}t)\right] = 0.$$

The IF signal due to the image frequency only appears at port *A*. The circuit in Figure 5.113 hence separates the IF components due to the RF and image signals at the input.

The topology in Figure 5.114 serves the same purpose and was described by Hartley already in 1928 [18]. It has the advantage that oscillators can be constructed so that they directly generate quadrature output signals, which maintain 90° phase shift over a wide frequency band.

The final topology, introduced by Weaver in 1956, eliminates the 90° hybrid at the IF, but uses a second down-conversion step [39]. It is shown in Figure 5.115.

While it requires two additional mixers and low-pass filters plus an extra LO, it eliminates the need for the IF hybrid which, for low IFs, is very difficult to integrate.

5.6.8 Mixer noise figure

The existence of an image frequency also complicates the noise figure assessment of mixers; while the intended signal is present at only one frequency (ω_{RF}), noise present at the image frequency ω_{img} is also converted to the IF and decreases the signal-to-noise ratio at the mixer output.

In principle, noise figure definitions for mixers use Friis' definition of the two-port noise figure (Equation (5.71)); the noise figure is calculated as the quotient of the signal-to-noise ratios at the input and the output of the two-port, assuming that the noise temperature of the input is $T_0 = 290\,\mathrm{K}$.

The issue here is how to calculate the signal-to-noise ratio at the input.

We may consider the signal-to-noise ratio only for the RF frequency. If the signal power is S, the signal-to-noise ratio before the mixer is

$$\mathrm{SNR}_{\mathrm{inp}} = \frac{S}{kT_0\Delta f}, \tag{5.261}$$

where Δf is the measurement bandwidth.

The mixer noise sources are combined in an equivalent noise source $kT_{\mathrm{n}}\Delta f$, placed at the input. T_{n} is the noise temperature of the mixer.

Generally, the mixer's gain at the RF and the image frequency can have different values, which we call G_{RF} and G_{img}, respectively.[8]

The image frequency also contributes a noise power of $kT_0\Delta f$. The signal-to-noise ratio at the output is then

$$\mathrm{SNR}_{\mathrm{out}} = \frac{G_{\mathrm{RF}}S}{k\Delta f[G_{\mathrm{RF}}(T_0 + T_{\mathrm{n}}) + G_{\mathrm{img}}T_0]}. \tag{5.262}$$

The ratio of the expressions (5.261) and (5.262) is called the *single-sideband noise figure*:

$$F_{\mathrm{SSB}} = 1 + \frac{T_{\mathrm{n}}}{T_0} + \frac{G_{\mathrm{img}}}{G_{\mathrm{RF}}} \tag{5.263}$$

We may also simply apply Equation (5.73), developed originally for two-ports, to the mixer. The result is the *IEEE single-sideband noise figure*:

$$F_{\mathrm{SSB,IEEE}} = 1 + \frac{T_{\mathrm{n}}}{T_0} \tag{5.264}$$

The IEEE definition is similar to (5.263), provided that $G_{\mathrm{img}} \ll G_{\mathrm{RF}}$. However, we can also calculate the signal-to-noise ratio before the mixer taking the thermal noise at the image frequency into account. Then,

$$\mathrm{SNR}_{\mathrm{inp}} = \frac{S}{2kT_0\Delta f}. \tag{5.265}$$

[8] Note that all gains are available gains, i.e. power match is assumed for all ports.

Now using (5.265) and (5.262) to calculate the noise figure, we obtain the *double-sideband noise figure*:

$$F_{\mathrm{DSB}} = \frac{F_{\mathrm{SSB}}}{2} = \frac{F_{\mathrm{SSB,IEEE}} + \frac{G_{\mathrm{img}}}{G_{\mathrm{RF}}}}{2}. \tag{5.266}$$

On a logarithmic scale, F_{SSB} is thus 3 dB larger than F_{DSB}.

For the assessment of receiver systems, F_{SSB} is the appropriate entity because we need to compare signal-to-noise ratios at the intended frequency only. However, measurements of mixer noise figure invariably yield F_{DSB}, unless the image frequency is suppressed at the input of the mixer by appropriate filtering. When using mixer noise figure data, this has to be carefully observed.

5.7 Baluns, unbals and hybrids

In the preceding sections, reference was frequently made to mysterious building blocks which convert signals from single-end to differential (two signals of equal amplitude, but 180° out of phase), or splitting a signal into two parts 'in quadrature' (90° out of phase). Due to their importance, they also deserve a brief section of their own.

The term *balun* is a contraction of *balanced to unbalanced*, describing the function this component performs; they convert a balanced (differential) signal, which can be understood as two ground-referenced signals of equal amplitude, but 180° phase difference, to an unbalanced signal, which is a single ground-referenced form. The *unbal* does exactly the opposite – it converts an *unbalanced* signal to a *balanced* (differential) one.

5.7.1 Passive baluns and unbals

Passive baluns and unbals are essentially the same components, operated in different directions. Therefore, it is common to call them by the name 'balun' only, irrespective of the actual role.

A very common balun/unbal at lower RF frequencies is the centre-tapped transformer (Figure 5.116). Due to the symmetric centre tap of the secondary winding, the balanced output voltage is formed by two voltages v_1 and $\bar{v}_1$ of equal amplitude, but opposite phase. At lower frequencies, the centre-tapped transformer balun is frequently constructed as a *toroidal* transformer, using ring-shaped magnetic cores. As the frequency of operation increases, this will rapidly not be possible anymore as the useful frequency

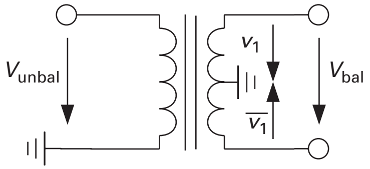


Fig. 5.116 Center-tapped transformer as a balun/unbal.

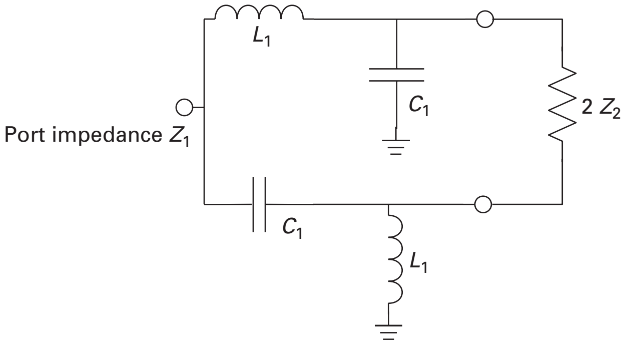


Fig. 5.117 Lumped-element balun circuit.

range of the core material is exceeded and also the parasitic capacitances of the windings become excessive.

A rather straightforward balun implementation is shown in Figure 5.117. It uses a combination of LC low-pass and high-pass filters to provide phase shifts of −90° and +90°, respectively, at the design frequency ω_0. It can provide impedance transformation at the same time. Both high-pass and low-pass filters provide the 90° phase shifts if

$$L_1 C_1 = \frac{1}{\omega_0^2}. \tag{5.267}$$

At this frequency, they need to additionally fulfil:

$$\frac{L_1}{C_1} = 2 Z_1 Z_2 \tag{5.268}$$

to transform between the single-ended port impedance Z_1 and the differential port impedance $2Z_2$. Solving Equations (5.267) and (5.268) yields

$$L_1 = \frac{\sqrt{Z_1 Z_2}}{\omega_0} \tag{5.269}$$

$$C_1 = \frac{L_1}{2 Z_1 Z_2}. \tag{5.270}$$

As all lumped-element transformation circuits, the lumped-element balun will be quite narrow band.

At microwave frequencies, a rather large number of possible implementations of *transmission line* baluns exist. We will restrict our discussion to one example.

A very popular transmission line balun, whose operation is also quite easy to understand, is the *rat-race* coupler structure shown in Figure 5.118. It consists of a transmission line ring with a circumference of 1.5 times the wavelength λ, with four ports arranged as shown in Figure 5.118. If equal power distribution to the coupled ports is desired, the transmission line ring must have a characteristic impedance of $\sqrt{2}Z_0$, where Z_0 is the port impedance. If we feed a signal into port 1, it will split into two partial signals, which travel clockwise and counter-clockwise through the ring. They will interfere constructively at port 3, 2 and 3, while the phase difference is π at port 4, leading to destructive interference there. The signal at port 3 lags port 1 by $\pi/2$, while the

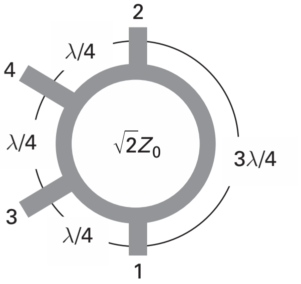


Fig. 5.118 A 180° hybrid ('rat race') coupler structure.

signal at port 2 leads port 1 by $\pi/2$ (well, it lags by $3\pi/2$, but that is the same). Because port 4 is decoupled, and we have equal power distribution, the scattering parameters with respect to port 1 are then:

$$S_{21} = +\jmath\frac{1}{\sqrt{2}}$$
$$S_{31} = -\jmath\frac{1}{\sqrt{2}}$$
$$S_{41} = 0.$$

The phase difference between ports 2 and 3 is π or 180° – a single-ended signal into port 1 is converted into a differential signal between ports 2 and 3. The component is reciprocal, of course – a differential signal applied between ports 2 and 3 will combine into a single-ended signal out of port 1.

The rat-race coupler has another interesting property – a signal inserted into port 4 will split into in-phase components out of ports 2 and 3, while port 1 is isolated. In scattering matrix terms,

$$S_{24} = -\jmath\frac{1}{\sqrt{2}}$$
$$S_{34} = -\jmath\frac{1}{\sqrt{2}}$$
$$S_{14} = 0.$$

5.7.2 Active baluns and unbals

A major disadvantage of the passive baluns discussed so far is their narrow bandwidth of operation, and at lower frequencies their potentially large chip area consumption, due to either the size of the transmission line segments or the size of the necessary reactances.

Active circuits can provide balanced-to-unbalanced and unbalanced-to-balanced-conversions over a wide operational bandwidth, and often in a very small chip area. They are, therefore, frequently used in IC implementations of microwave circuits. Disadvantages are the additional noise due to the active components, potential non-linearities and the added power consumption, which can be considerable if high linearity is required.

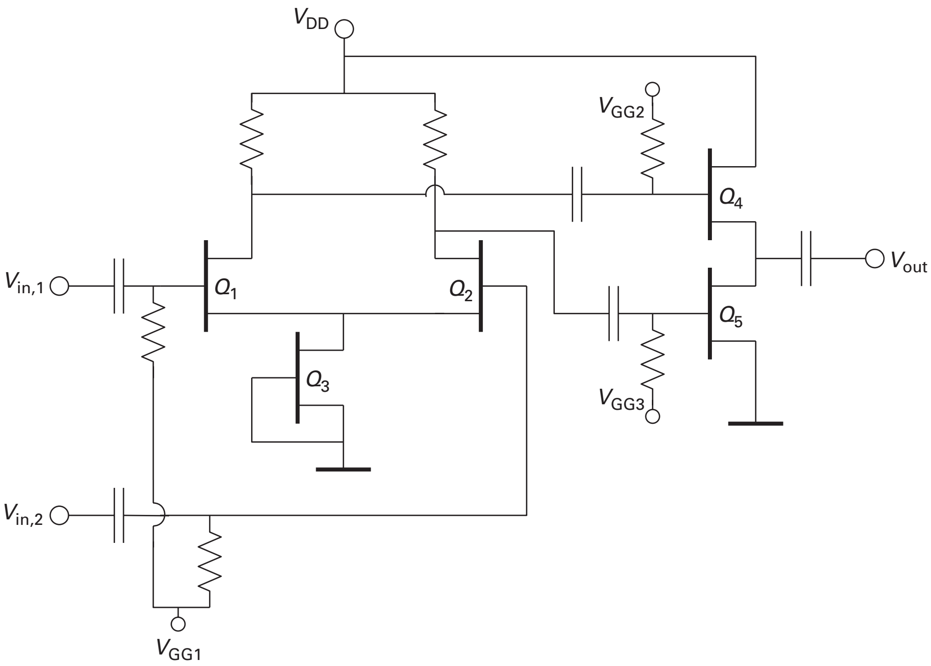


Fig. 5.119 Active balun circuit.

Figure 5.119 shows a typical active balun conversion circuit. Transistors $Q_1{-}Q_3$ form a differential amplifier, which is used here to suppress any common-mode components between the input voltages $V_{\text{in},1}$ and $V_{\text{in},2}$. Transistors Q_4 and Q_5 form the balun proper – Q_4 acts in common-drain (source follower) configuration, while Q_5 is in common-source configuration, both working against the common single-ended output.

The simplest active unbal (Figure 5.120), uses a transistor which is simultaneously configured in common-source and common-drain topology. Output 1 is at the source, hence the voltage gain is in good approximation +1. Output 2 is at the drain, and the resistors R_1 and R_2 must be chosen such that the voltage gain is −1, leading to the desired balanced output signal V_{out}. Quasistatically, this is very simple – the small-signal output voltage $v_{\text{out},2}$ is

$$v_{\text{out},2} = -R_1 \frac{v_{\text{out},1}}{R_2},$$

so that for equal magnitudes of $v_{\text{out},1}$ and $v_{\text{out},2}$,

$$R_1 = R_2.$$

For higher frequencies, the performance deteriorates, especially due to leakage through the transistor's gate-drain capacitance. The useful range can be extended by cascading a differential amplifier with good common-mode rejection.

Figure 5.121 shows an alternative active unbal, which uses common-source and common-gate amplifiers in parallel. Q_1 is a common-gate topology, with a quasi-static

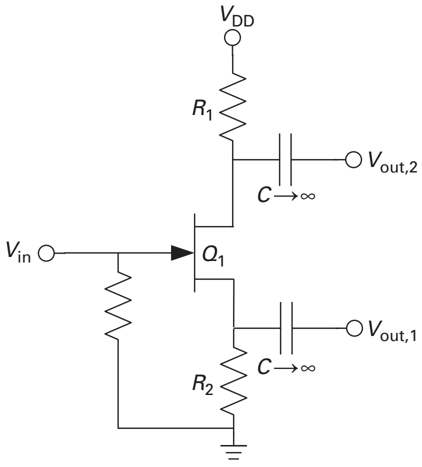


Fig. 5.120 Active unbal using a transistor in common-source/common-drain configuration.

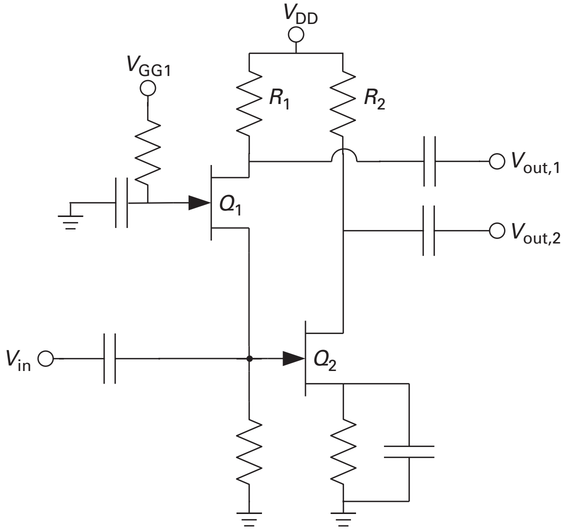


Fig. 5.121 Active unbal using a combination of common-source and common-gate topologies.

small-signal gain of $g_m R_1$,[9] while Q_2 is a common-source amplifier with a voltage gain of $-g_m R_2$. Provided $R_1 = R_2$ and equal transistor transconductances g_m, both output voltages will have the same magnitude, but opposite phase: $V_{out,1} = V_{out,2}$.

Compared to the earlier circuit (Figure 5.120), this circuit has the advantage that the input impedance is significantly reduced, leading to a higher bandwidth, while the output impedances are equal for both ports.

[9] Neglecting the transistor output conductance.

5.7.3 Quadrature generation

The generation of two signals with 90° phase shift is another very frequent task, as we have seen in the discussion of image-rejection mixer topologies.

The simplest generation of quadrature signals is realised using the simple RC network shown in Figure 5.122. The output voltages are

$$V_{\mathrm{Q}} = V_0 \frac{1}{1 + \jmath\omega R_1 C_1} \tag{5.271}$$

$$V_{\mathrm{I}} = V_0 \frac{\jmath\omega R_1 C_1}{1 + \jmath\omega R_1 C_1}. \tag{5.272}$$

At $\omega_0 = 1/(R_1C_1)$, V_{Q} lags V_0 by 45°, while V_{I} leads V_0 by 45° – V_{Q} and V_{I} are hence in quadrature. The 90° phase difference between V_{Q} and V_{I} is maintained over a wide frequency range, but the amplitudes are equal only at ω_0. Note also that this is a lossy network – at ω_0, $|V_{\mathrm{I}}| = |V_{\mathrm{Q}}| = V_0/\sqrt{2}$.

The polyphase filter family, an example of which is shown in Figure 5.123, uses also passive RC elements to generate quadrature output signals from a differential input

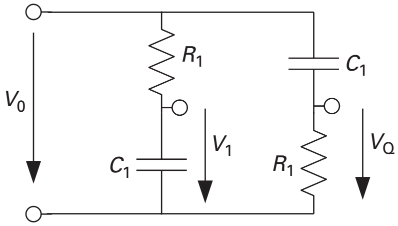


Fig. 5.122 Quadrature generation using a simple RC network.

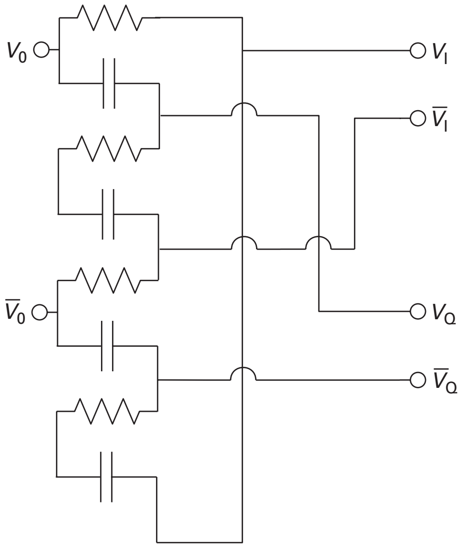


Fig. 5.123 RC polyphase network for differential quadrature generation.

signal. It is very frequently used in fully differential receiver concepts. All RC filters have significant loss and should therefore not be used in noise-critical signal paths – a common use is for quadrature generation from differential LOs, or in quadrature IF combiners at the output of image-rejection mixers such as the one shown in Figure 5.114. An in-depth description of polyphase filter operation can be found in [3].

Another common quadrature generator, this time using transmission lines, is the 90° hybrid, shown in Figure 5.124.

All transmission line segments are electrically a quarter wavelength long. A signal fed into port 1 will emerge at 2 with 90° phase lag, and with 180° at port 3. The signals at 2 and 3 are hence in quadrature. The signal from port 1 will interfere destructively at port 4, which is hence ideally decoupled, provided that 2 and 3 are terminated with the proper impedance Z_0.

Quadrature signals can also be generated digitally where linear operation is not mandated – following an oscillator.

The circuit depicted in Figure 5.125(a) needs a clock frequency at four times the intended output frequency. As state changes occur only on the rising edge of the clock, it

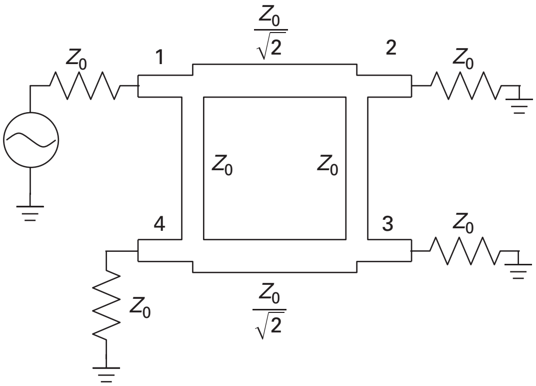


Fig. 5.124 Transmission line 90° hybrid.

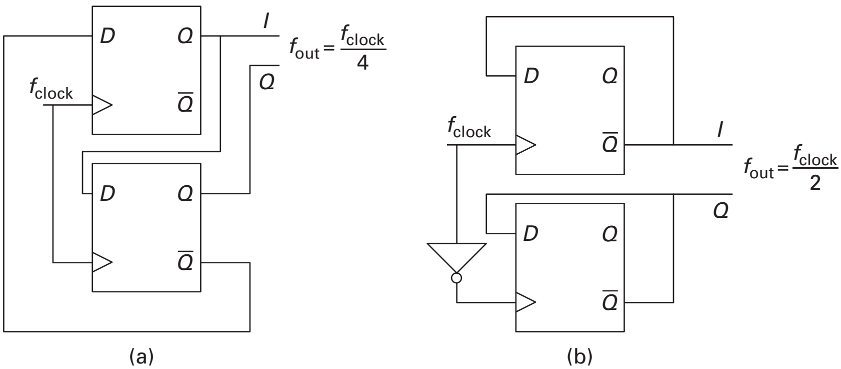


Fig. 5.125 Quadrature signal generation using flipflops.

is insensitive to the clock's duty cycle. However, the very high clock frequency requirement makes it unsuitable for many applications. The circuit in Figure 5.125(b) needs only twice the output frequency as a clock, but as it triggers on both the rising and the falling edges of the clock, it is very sensitive to the clock's duty cycle. The time delay in the inverter may lead to additional problems as the clock frequency increases.

Finally, quadrature clock signals may also be generated in special oscillator circuits. This, however, shall be beyond the scope of this book.

5.8 Problems

(1) Demonstrate that the power delivered to a load of arbitrary impedance can be expressed as the difference in the squared magnitudes of the incident and reflected normalised power waves – see Equation (5.3).

(2) Consider a two-port whose scattering matrix is known. Calculate the power delivered to an arbitrary load Z_L as a function of the available power of the generator, whose source impedance shall be equal to the normalising impedance Z_0.

(3) You have to design a common-source amplifier with a FET technology whose transit frequency f_T is 50 GHz. The load resistance of the amplifier is given to be 100 Ω, the voltage gain shall be $A_V = -10$. Calculate the input capacitance – hint: use the common rule of thumb that $C_{GD} \approx 0.1 \cdot C_{GS}$. You may also neglect g_{DS}.

(4) Consider the circuit in Figure 5.126. The transistors Q_1 and Q_2 shall have a transconductance of $g_m = 20$ mS and a transit frequency $f_T = 50$ GHz. Calculate the input admittance of this circuit.

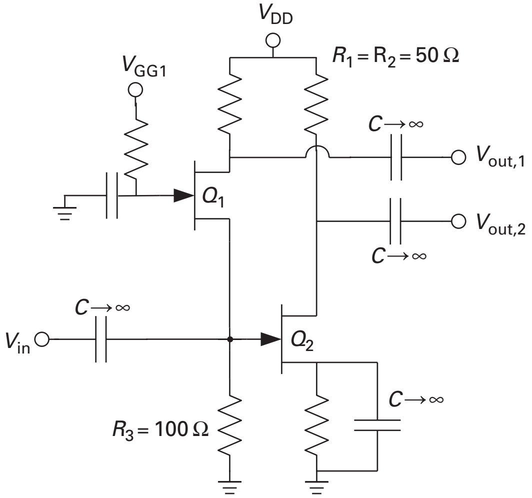


Fig. 5.126 Active unbal circuit for Problem 4.

(5) Suggest a simple one-transistor circuit capable of creating an impedance (referenced to ground) with a real part $\mathrm{Re}\{Z_1\} = -400\,\Omega$ at $f = 10\,\mathrm{GHz}$. The FET technology you are using has an f_T of 100 GHz and the transistor transconductance shall be $g_\mathrm{m} = 50\,\mathrm{mS}$.

(6) The circuit in Figure 5.127 shall be used to realise a wideband voltage amplifier stage. The transistor Q_1 has an $f_\mathrm{T} = 10\,\mathrm{GHz}$ and transconductance of $g_\mathrm{m} = 100\,\mathrm{mS}$ in a bias point $I_\mathrm{D} = 10\,\mathrm{mA}$, $V_\mathrm{GS} = -0.5\,\mathrm{V}$ and V_DD is 10 V. The voltage gain shall be $a_\mathrm{V} = -10$.
(a) Calculate R_3.
(b) What is the optimum choice for C_1 with respect to maximum bandwidth, if the transistor can be modelled using the simple equivalent circuit in Figure 5.20?
(c) Calculate R_2.
(d) What is V_DS under these circumstances?

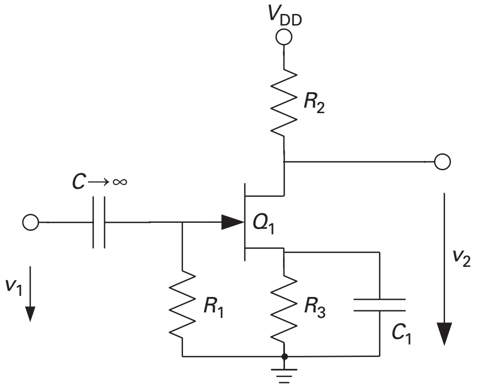


Fig. 5.127 Simple wideband amplifier.

(7) Consider the circuit in Figure 5.32. Given that all transistors have the same size, why do Q_1 and Q_2 have the same collector current, provided that the current gain $\beta \gg 1$? Let now $I_\mathrm{C} = 5\,\mathrm{mA}$ for all transistors, $\beta = 100$, $f_\mathrm{T} = 50\,\mathrm{GHz}$, $R_\mathrm{L} = 100\,\Omega$. Draw the small-signal equivalent circuit and calculate the input impedance and the voltage gain.

(8) Discuss three different ways of eliminating the feedback capacitance in amplifiers.

(9) The impedance seen in the input of a common-source amplifier stage at $f = 10\,\mathrm{GHz}$ is $Z_1 = (5 - \jmath 159)\,\Omega$. The output admittance of the preceding stage is $Y_2 = 10 + \jmath 3\,\mathrm{mS}$. Suggest a suitable matching network so that the available power is transferred from the first to the second stage. Hint: the Smith chart is very helpful here.

(10) You have the task to design a simple distributed amplifier. The transistors to use have a transconductance $g_\mathrm{m} = 400\,\mathrm{mS}$ and $f_\mathrm{T} = 70\,\mathrm{GHz}$. The gain cell shall be a simple common-source stage. The gate width of each transistor is $W_\mathrm{G} = 100\,\mu\mathrm{m}$. For the drain-gate capacitance, use $C_\mathrm{DG} = 0.1 C_\mathrm{GS}$.

The unloaded gate line has a characteristic impedance $Z_\mathrm{u} = 80\Omega$ and an inductance per unit length of $L' = 0.8\,\mathrm{nH\,mm^{-1}}$.

Calculate the necessary length of the gate line segments so that the characteristic impedance of the loaded gate line is 50Ω. The drain line shall equally have this impedance.

(11) In the schematic in Figure 5.63, identify the function of each of the transistors shown. What is the purpose of the RC combination attached to the emitters of two of the transistors?

(12) A communications system can tolerate a minimum third-order intermodulation distance of 60 dB; the maximum input power is −20 dBm. What is the necessary input-referred third-order intercept point?

(13) Explain why a high resonator quality factor is especially important in oscillator using active devices with little low-frequency noise, such as Si/SiGe HBTs.

(14) Show how the Hartley image reject topology (Figure 5.114) achieves separation of the signal and image frequencies to output A and B.

References

[1] Abele P., Kallfass I., Zeuner M. *et al.* (2003). 32 GHz MMIC distributed amplifier based on n-channel SiGe MODFETs. *Electron. Lett. 39*, 1448–1449.

[2] Battjes C. R. (1980). Monolithic wideband amplifier. USA Patent 4, 236, 119.

[3] Behbahani F., Kishigami Y., Leete J., Abidi A. (2001). CMOS mixers and polyphase filters for large image rejection. *IEEE J. Solid-State Circ. 36*, 873–887.

[4] Beyer J., Prasad S. N., Becker R., Nordman J., Hohenwarter G. (1984). MESFET distributed amplifier design guidelines. *IEEE Trans. Microw. Theory Tech. MTT-32*, 268–275.

[5] Chartier S., Schleicher B., Korndörfer F., Glisic S., Fischer G., Schumacher H. (2007). A fully integrated fully differential low-noise amplifier for short range automotive radar using a SiGe:C BiCMOS technology. In *Proc. 2nd European Microwave IC Conference (EuMIC), Munich, Germany, 8–10 October, 2007*.

[6] Chartier S., Sönmez E., Schumacher H. (2006). Millimeter-wave amplifiers using a 0.8 μm Si/SiGe HBT technology. In *Proc. Silicon Monolithic Integrated Circuits in RF Systems, San Diego, CA, 18–20 January 2006*. IEEE.

[7] Darlington S. (1953). Semiconductor translating device. USA Patent 2, 663, 806.

[8] Ellinger F. (2007). *Radio Frequency Integrated Circuits and Technologies*, 1st edn. Springer.

[9] Fleming J. A. (1905). Instrument for converting alternating electric currents into continuous currents. USA Patent 803, 684.

[10] Forest L. D. (1908). Space telegraphy. USA Patent 879, 532.

[11] Friis H. T. (1944). Noise figure of radio receivers. *Proc. IRE 32*, 7, 419–422.

[12] Gilbert B. (1968). A precise four-quadrant multiplier with subnanosecond response. *IEEE J. Solid-State Circ. 3*, 365–373.

[13] Gilbert B. (1997). The Micromixer: a highly linear variant of the Gilbert mixer using a bisymmetric class-AB input stage. *IEEE J. Solid-State Circ. 32*, 1412–1423.

[14] Gonzalez G. (1997). *Microwave Transistor Amplifiers, Analysis and Design*. Prentice Hall.

[15] Gupta M. S. (1992). Power gain in feedback amplifiers, a classic revisited. *IEEE Trans. Microw. Theory Tech. MTT-40*, 5 (May), 864–879.

[16] Häfele M., Trasser A., Beilenhoff K., Schumacher H. (2005). A GaAs distributed amplifier with an output voltage of 8.5 V_{pp} for 40 Gb/s modulators. *Proc. 13th GAAS Symposium, Paris, France, 3–4 October 2005,* 345–348.

[17] Hagen J. B. (1996). *Radio-Frequency Electronics*, 1st edn. Cambridge University Press.

[18] Hartley R. V. L. (1928). Modulation system. USA Patent 1, 666, 206.

[19] Haus H., Adler R. (1958). Optimum noise performance of linear amplifiers. *Proc. IRE 46*, 1519–1533.

[20] Hoffmann M. (1997). *Hochfrequenztechnik*, 1st edn. Springer Berlin Heidelberg.

[21] Hunt F., Hickman R. (1939). On electronic voltage stabilizers. *Rev. Sci. Instrum. 10*, 1 (January), 6–21.

[22] Kline R. (1993). Harold Black and the negative-feedback amplifier. *IEEE Contr. Syst. Mag. 13*, 4, 82–85.

[23] Ku W. H. (1966). Unilateral gain and stability criteria of active two ports in terms of scattering parameters. *Proc. IEEE 54*, 11 (November), 1617–1618.

[24] Kurokawa K. (1965). Power waves and the scattering matrix. *IEEE Trans. Microw. Theory Tech. MTT-13*, 3 (March), 194–202.

[25] Lee T. H. (1997). *The Design of CMOS Radio-Frequency Integrated Circuits*, 1st edn. Cambridge University Press, 181.

[26] Leeson D. (1966). A simple model for feedback oscillator noise spectrum. *Proc. IEEE 54*, 329–330.

[27] Maas S. A. (1987). A GaAs MESFET mixer with very low intermodulation. *IEEE Trans. Microw. Theory Tech. MTT-35*, 425–429.

[28] Mason S. J. (1954). Power gain in feedback amplifiers. *Trans. IRE, Prof. Group on Circ. Theory CT-1*, 2 (June), 20–25.

[29] Miller J. (1920). Dependence of the input impedance of a three-electrode vacuum tube upon the load in the plate circuit. *Scientific Papers of the Bureau of Standards 15* (351), 367–385.

[30] Percival W. S. (1937). Thermionic valve circuits. Britain Patent 460, 562.

[31] Rothe H., Dahlke W. (1955). Theorie rauschender Vierpole. *AEÜ 9*, 117–121.

[32] Rothe H., Dahlke W. (1956). Theory of noisy fourpoles. *Proc. IRE 44*, 811–815.

[33] Schick C., Feger T., Sönmez E., Schad K., Trasser A., Schumacher H. (2005). Broadband Si/SiGe HBT amplifier concepts for 40 GBit/s fibreoptic communication systems. In *Proc. 35th European Microwave Conf., Paris, France, October 2005*.

[34] Schmitt O. (1937). A simple differential amplifier. *Rev. Sci. Inst. 8*, 4 (April), 126–127.

[35] Schumacher H., Abele P., Sönmez E., Schad K., Trasser A. (2003). Low-cost circuit solutions for micro- and millimeter-wave systems using commercially available SiGe technologies. In *Proc. 1st International SiGe Technology and Device Meeting, Nagoya, Japan, 14–17 January 2003*.

[36] Smith P. H. (1931). Transmission line calculator. *Electronics 12*, 1 (January), 29–31.

[37] Smith P. H. (1944). An improved transmission line calculator. *Electronics 17*, 1 (January), 130.

[38] Sokal N., Sokal A. (1975). Class E – a new class of high-efficiency tuned single-ended switching power amplifiers. *IEEE J. Solid-State Circ. 10*, 3 (June), 168–176.

[39] Weaver D. (1956). A third method of generation and detection of single-sideband signals. *Proc. IRE 44*, 1703–1705.

Index