OPTICAL NETWORKS
Biswanath Mukherjee *Series Editor*

Mohammad Azadeh

Fiber Optics Engineering

Springer

CYTEK V. BECKMAN
IPR2025-01319
BECKMAN 2039

Fiber Optics Engineering

For further volumes:
http://www.springer.com/series/6976

Optical Networks

Series Editor: Biswanath Mukherjee
University of California, Davis
Davis, CA

Mohammad Azadeh

Fiber Optics Engineering

Springer

Mohammad Azadeh
Source Photonics, Inc.
20550 Nordhoff St.
Chatsworth, CA 91311
USA
azadeh@sourcephotonics.com

Series Editor
Biswanath Mukherjee
University of California
Davis, CA
USA

ISSN 1935-3839
ISBN 978-1-4419-0303-7 e-ISBN 978-1-4419-0304-4
DOI 10.1007/978-1-4419-0304-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009929311

© Springer Science+Business Media, LLC 2009
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

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

Preface

Within the past few decades, information technologies have been evolving at a tremendous rate, causing profound changes to our world and our ways of life. In particular, fiber optics has been playing an increasingly crucial role within the telecommunication revolution. Not only most long-distance links are fiber based, but optical fibers are increasingly approaching the individual end users, providing wide bandwidth links to support all kinds of data-intensive applications such as video, voice, and data services.

As an engineering discipline, fiber optics is both fascinating and challenging. Fiber optics is an area that incorporates elements from a wide range of technologies including optics, microelectronics, quantum electronics, semiconductors, and networking. As a result of rapid changes in almost all of these areas, fiber optics is a fast evolving field. Therefore, the need for up-to-date texts that address this growing field from an interdisciplinary perspective persists.

This book presents an overview of fiber optics from a practical, engineering perspective. Therefore, in addition to topics such as lasers, detectors, and optical fibers, several topics related to electronic circuits that generate, detect, and process the optical signals are covered. In other words, this book attempts to present fiber optics not so much in terms of a field of "optics" but more from the perspective of an engineering field within "optoelectronics." As a result, practicing professionals and engineers, with a general background in physics, electrical engineering, communication, and hardware should find this book a useful reference that provides a summary of the main topics in fiber optics. Moreover, this book should be a useful resource for students whose field of study is somehow related to the broad areas of optics, optical engineering, optoelectronics, and photonics.

Obviously, covering all aspects of fiber optics in any depth requires many volumes. Thus, an individual text must out of necessity be selective in the topics it covers and in the perspectives it offers. This book covers a range of subjects, starting from more abstract basic topics and proceeding towards more practical issues. In most cases, an overview of main results is given, and additional references are provided for those interested in more details. Moreover, because of the practical character of the book, mathematical equations are kept at a minimum, and only essential equations are provided. In a few instances where more mathematical details are given and equations are derived, an elementary knowledge of calculus is sufficient for following the discussion, and the inconvenience of having to go through the math is well rewarded by the deeper insights provided by the results.

The logical flow of the book is as follows. The first three chapters act as a foundation and a general background for the rest of the book. Chapter 1 covers basic physical concepts such as the nature of light, electromagnetic spectrum, and a brief overview of fiber optics. Chapter 2 provides an overview of important networking concepts and the role of fiber optics within the telecommunication infrastruc-

ture. Chapter 3 provides an introduction to fiber optics from a signal viewpoint. This includes some basic mathematical background, as well as characterization of physical signals in the electrical and optical domains.

Chapters 4–7 cover the main elements of a fiber optic link in more depth. Chapter 4 is dedicated to diode lasers which are the standard source in fiber optics. Chapter 5 deals with propagation of optical signals in fibers and signal degradation effects. PIN and APD detectors that convert photons back to electrons are the topic of Chapter 6. Thus, these three chapters deal with generation, propagation, and detection of optical signals. Chapter 7, on the other hand, deals with light coupling and passive components. Therefore, Chapter 7 examines ways of transferring optical signals between elements that generate, detect, and transport the optical signals.

The next two chapters, Chapters 8 and 9, essentially deal with electronic circuits that interface with diode lasers and optical detectors. In particular, Chapter 8 examines optical transmitter circuits and various electronic designs used in driving high-speed optical sources. Chapter 9 examines the main blocks in an optical receiver circuit as well as ways of characterizing the performance of a receiver. A feature of this book is that in addition to traditional CW transceivers, burst mode transmitter and receiver circuits, increasingly used in PON applications, are also discussed.

The final three chapters of the book cover areas that have to do with fiber optics as a viable industry. Chapter 10 presents an overview of reliability issues for optoelectronic devices and modules. A viable fiber optic system is expected to operate outside the laboratory and under real operating conditions for many years, and this requires paying attention to factors outside pure optics or electronics. Chapter 11 examines topics related to test and measurement. In an engineering environment, it is crucial not only to have a firm grasp on theoretical issues and design concepts, but also to design and conduct tests, measure signals, and use test instruments effectively. Finally, Chapter 12 presents a brief treatment of fiber optic related standards. Standards play a crucial rule in all industries, and fiber optics is no exception. Indeed, it is oftentimes adherence to standards that enables a device or system to go beyond a laboratory demonstration and fulfill a well-defined role in the jigsaw of a complex industry such as fiber optics.

* * *

I am greatly indebted to many individuals for this project. In particular, I would like to thank Dr. A. Nourbakhsh who inspired and encouraged me to take on this work. I would also like to acknowledge my past and present colleagues at Source Photonics for the enriching experience of many years of working together. In particular, I would like to thank Dr. Mark Heimbuch, Dr. Sheng Zheng, Dr. Near Margalit, Dr. Chris LaBounty, and Dr. Allen Panahi, for numerous enlightening discussions on a variety of technical subjects. Without that experience and those discussions, this book could not have been created. I would also like to thank Springer for accepting this project, and in particular Ms. Katelyn Stanne, whose guidance was essential in bringing the project to its conclusion.

Mohammad Azadeh

Contents

Chapter 1 Fiber Optic Communications: A Review..... 1
1.1 Introduction 1
1.2 The nature of light 3
1.2.1 The wave nature of light 4
1.2.2 The particle nature of light..... 8
1.2.3 The wave particle duality 9
1.3 The electromagnetic spectrum 10
1.4 Elements of a fiber optic link..... 13
1.5 Light sources, detectors, and glass fibers..... 15
1.5.1 Optical sources..... 15
1.5.2 Optical detectors 18
1.5.3 The optical fiber..... 19
1.6 Advantages of fiber optics 20
1.7 Digital and analog systems 21
1.8 Characterization of fiber optic links 22
1.9 Summary 25

Chapter 2 Communication Networks 29
2.1 Introduction 29
2.2 Network topologies..... 29
2.3 Telecommunication networks..... 33
2.4 Networking spans 38
2.4.1 Local area networks (LANs)..... 38
2.4.2 Metropolitan area networks (MANs) 38
2.4.3 Wide area networks (WANs) 39
2.5 Hierarchical structure of networks..... 40
2.5.1 Open System Interconnect (OSI) model 40
2.5.2 Datalink layer..... 42
2.5.3 Network layer..... 43
2.5.4 Higher layers..... 43
2.6 Circuit switching and packet switching networks..... 43
2.6.1 Circuit switching..... 44
2.6.2 Packet switching 45
2.7 SONET/SDH 46
2.8 WDM networks 49
2.9 Passive optical networks (PONs)..... 55
2.10 Summary..... 57

Chapter 3 Signal Characterization and Representation 61
3.1 Introduction 61
3.2 Signal analysis 61
3.2.1 Fourier transform 62
3.2.2 Fourier analysis and signal representation 63

3.2.3 Digital signals, time and frequency domain representation 65
3.2.4 Non-return-to-zero (NRZ) and pseudorandom (PRBS) codes 65
3.2.5 Random and pseudo-random signals in frequency domain.....67
3.3 High-speed electrical signals 68
3.3.1 Lumped and distributed circuit models.....68
3.3.2 Transmission lines 70
3.3.3 Characteristic impedance 71
3.3.4 Microstrip and striplines 73
3.3.5 Differential signaling 76
3.4 Optical signals 79
3.4.1 Average power.....80
3.4.2 Eye diagram representation.....81
3.4.3 Amplitude parameters.....82
3.4.4 Time parameters.....84
3.4.5 Eye pattern and bathtub curves 86
3.5 Spectral characteristics of optical signals 88
3.5.1 Single-mode signals.....88
3.5.2 Multimode signals.....90
3.6 Summary 91

Chapter 4 Semiconductor Lasers.....95
4.1 Introduction 95
4.2 Optical gain and optical oscillation 95
4.3 Physical processes for optical amplification.....98
4.4. Optical amplification in semiconductors 100
4.5 Rate equation approximation.....103
4.5.1 Carrier density rate equation 104
4.5.2 Photon density rate equation 106
4.5.3 Steady-state analysis 107
4.5.4 Temperature dependence of LI curve 110
4.5.5 Small signal frequency response.....111
4.5.6 Time response 113
4.5.7 Frequency chirp 114
4.5.8 Large signal behavior.....115
4.6 Semiconductor laser structures 117
4.6.1 Heterostructure laser 118
4.6.2 Quantum well lasers.....119
4.6.3 Distributed feedback (DFB) lasers.....120
4.6.4 Vertical surface emitting lasers (VCSELs) 121
4.7 Summary 123

Chapter 5 Optical Fibers 127
5.1 Introduction 127
5.2 Optical fiber materials, structure, and transmission windows 127
5.3 Guided waves in fibers 131

5.3.1 Guided modes, ray description 131
5.3.2 Guided modes, wave description 133
5.3.3 Signal degradation in optical fibers 135
5.4 Attenuation 135
5.4.1 Absorption 137
5.4.2 Scattering 137
5.5 Dispersion 138
5.5.1 Modal dispersion 139
5.5.2 Chromatic dispersion 140
5.5.3 Waveguide dispersion 142
5.5.4 Polarization dispersion 143
5.6 Nonlinear effects in fibers 144
5.6.1 Self- and cross-phase modulation (SPS and XPM) 144
5.6.2 Four Wave Mixing (FWM) 146
5.6.3 Stimulated Raman scattering (SRS) 147
5.6.4 Stimulated Brillouin Scattering (SBS) 148
5.7 Fiber amplifiers 149
5.8 Summary 151

Chapter 6 PIN and APD Detectors 157
6.1 Introduction 157
6.2 The PIN diode and photon-electron conversion 157
6.2.1 PIN diode, static characteristics 158
6.2.2 PIN diode, dynamic characteristics 161
6.3 Avalanche photodiode (APD) 162
6.4 Noise in photodetectors 166
6.4.1 Shot noise 166
6.4.2 Thermal noise 167
6.4.3 Signal-to-noise ratio (SNR) 168
6.5 Photodetector materials and structures 170
6.5.1 Photodetector materials 170
6.5.2 PIN diode structures 172
6.5.3 APD structures 172
6.6 Summary 173

Chapter 7 Light Coupling and Passive Optical Devices 177
7.1 Introduction 177
7.2 Coupling light to and from a fiber 177
7.2.1 Direct coupling 178
7.2.2 Lensed fibers 180
7.2.3 Fiber coupling via lens 180
7.3 Fiber-to-fiber coupling 182
7.3.1 Connectorized couplings 183
7.3.2 Fiber finish 185
7.3.3 Fiber splicing 186

7.4 Passive components 188
7.4.1 Splitters and couplers 188
7.4.2 Attenuators 190
7.4.3 Isolators 191
7.4.4 Optical filters 193
7.5 Summary 193

Chapter 8 Optical Transmitter Design 199
8.1 Introduction 199
8.2 Transmitter optical subassembly (TOSA) 200
8.3 Biasing the laser: the basic LI curve 201
8.4 Average power control (APC) 203
8.4.1 Open loop average power control schemes 204
8.4.2 Closed loop power control 206
8.4.3 Thermal runaway 208
8.5 Modulation circuit schemes 209
8.5.1 Basic driver circuit 209
8.5.2 Transmission line effects 211
8.5.3 Differential coupling 212
8.5.4 High current drive circuits: ac coupling 213
8.6 Modulation control, open loop vs. closed loop schemes 216
8.6.1 Open loop modulation control 216
8.6.2 Closed loop modulation control: Pilot tone 217
8.6.3 Closed loop modulation control: high bandwidth control 218
8.7 External modulators and spectral stabilization 219
8.8 Burst mode transmitters 221
8.9 Analog transmitters 224
8.10 High frequency design practices 227
8.10.1 Power plane 227
8.10.2 Circuit layout 229
8.11 Summary 232

Chapter 9 Optical Receiver Design 235
9.1 Introduction 235
9.2 Receiver optical subassembly (ROSA) 235
9.2.1 Transimpedance amplifier (TIA) 236
9.2.2 Detector/TIA wire bonding in optical subassemblies 238
9.2.3 APD receivers 240
9.3 Limiting amplifier 242
9.4 Clock and data recovery 245
9.5 Performance of optical receivers 246
9.5.1 Signal-to-noise ratio (SNR) and bit error rate (BER) 247
9.5.2 Sensitivity 249
9.5.3 Overload 252
9.6 Characterization of clock and data recovery circuits 253

9.6.1 Jitter transfer 253
9.6.2 Jitter tolerance 256
9.7 Burst mode receivers 257
9.7.1 Dynamic range challenges in burst mode traffic 257
9.7.2 Design approaches for threshold extraction 258
9.7.3 Burst mode TIAs 260
9.8 Summary 261

Chapter 10 Reliability 265
10.1 Introduction 265
10.2 Reliability, design flow, and design practices 266
10.2.1 Design flow 267
10.2.2 Modular approach 268
10.2.3 Reliability design practices and risk areas 269
10.3 Electrical issues 271
10.3.1 Design margin 271
10.3.2 Printed circuit boards (PCBs) 274
10.3.3 Component selection 275
10.3.4 Protective circuitry 276
10.4 Optical issues 277
10.4.1 Device level reliability 277
10.4.2 Optical subassemblies 278
10.4.3 Optical fibers and optical coupling 278
10.5 Thermal issues 279
10.5.1 Power reduction 280
10.5.2 Thermal resistance 281
10.6 Mechanical issues 282
10.6.1 Shock and vibration 282
10.6.2 Thermal induced mechanical failures 284
10.6.3 Mechanical failure of fibers 285
10.7 Software issues 285
10.7.1 Software reliability 286
10.7.2 Failure rate reduction 287
10.8 Reliability quantification 288
10.8.1 Statistical models of reliability: basic concepts 288
10.8.2 Failure rates and MTTF 289
10.8.3 Activation energy 291
10.9 Summary 293

Chapter 11 Test and Measurement 297
11.1 Introduction 297
11.2 Test and measurement: general remarks 297
11.3 Optical power 299
11.4 Optical waveform measurements 301
11.4.1 Electrical oscilloscopes with optical to electrical converter 301

11.4.2 Digital communication analyzers (DCA) 302
11.4.3 Amplitude related parameters 305
11.4.4 Time-related parameters 306
11.4.5 Mask measurement 308
11.5 Spectral measurements 309
11.5.1 Optical spectrum analyzer (OSA) 309
11.5.2 Wavelength meters 312
11.6 Link performance testing 313
11.6.1 Bit error rate tester (BERT) 313
11.6.2 Sensitivity measurement 315
11.6.3 Sensitivity penalty tests 316
11.7 Analog modulation measurements 317
11.7.1 Lightwave signal analyzer (LSA) 317
11.7.2 Signal parameter measurements 319
11.8 Summary 322

Chapter 12 Standards 327
12.1 Introduction 327
12.2 Standards development bodies 327
12.2.1 International Telecommunication Union (ITU) 327
12.2.2 International Electrotechnical Commission (IEC) 328
12.2.3 Institute of Electrical and Electronics Engineers (IEEE) 328
12.2.4 Telecommunication Industry Association (TIA) 329
12.2.5 ISO and ANSI 329
12.2.6 Telcordia (Bellcore) 330
12.2.7 Miscellaneous organizations 330
12.3 Standards classification and selected lists 331
12.3.1 Standards related to components 332
12.3.2 Standards related to measurements and procedures 335
12.3.3 Reliability and safety standards 339
12.3.4 Networking and system standards 341
12.4 Fiber standards 345
12.5 Laser safety 346
12.6 SFF-8472 digital monitoring interface 347
12.6.1 Identification data (A0h) 347
12.6.2 Diagnostic data (A2h) 348
12.7 Reliability standards 349
12.8 Networking standards 351
12.8.1 SONET/SDH 352
12.8.2 Ethernet 353
12.8.3 Passive optical networks (PON) 355
12.9 Summary 356

Appendix A Common Acronyms 361

Appendix B Physical Constants 363

Index 365

Chapter 1
Fiber Optic Communications: A Review

1.1 Introduction

There is no doubt that telecommunication has played a crucial role in the makeup of the modern world. Without the telecommunication revolution and the electronic foundations behind it, the modern life would be unimaginable. It is not hard to imagine why this is the case, after all, it is communication that shapes us as human beings and makes the world intelligible to us. Our daily life is intricately intertwined with telecommunication and its manifestations. While we are driving, we call our friend who is traveling on the other side of the world. We can watch events live as they are unfolding in another continent. We buy something in Australia and our credit card account is charged in United States. We send and receive emails with all kinds of attachments in a fraction of second. In short, under the effects of instant telecommunications, the world is shrinking from isolated lands separated by vast oceans to an interconnected global village.

If telecommunication is a product of modern technology, communication itself in the form of language has been with humans as long as there have been humans around. Whether we are talking about human language or electronic communications, there are certain common fundamental features at work. To begin with, let us consider the case of common spoken language. Let us assume I am talking to my friend who is sitting next to me in a restaurant.

Conceptually, we can break down the process to certain stages. In the first stage, the process starts in my mind. I have some thoughts, memories, or concepts that I like to share with my friend. We can call these initial stages the processing layers (Fig. 1.1). The term "processing layers" in its plural form is meant to represent the extremely complex set of functions that logically constitute some sort of hierarchy. Data processing takes place in these various layers until eventually in the last stage the intended message is converted to a serial stream of data, for instance in the form of a sentence. The next process involves converting this stream of data into a physical signal. I can achieve this task by using my speech organs, through which I can produce and modulate sound waves in a precise manner, representing the sentence that I intend to communicate to my friend. Effectively, my vocal system acts as an audio transmitter.

Next comes the transmission of the physical signal. Sound waves carrying energy travel through the medium of the air until they reach my friend. The next step is receiving the physical signal and converting it back to a format that can be processed by the brain. This is achieved by the ear, which acts as the receiver. The function of the receiver is to convert the physical signal back to a form that is suitable for further processing by the nervous system and ultimately the brain.

M. Azadeh, *Fiber Optics Engineering*, Optical Networks,
DOI 10.1007/978-1-4419-0304-4_1, © Springer Science+Business Media, LLC 2009

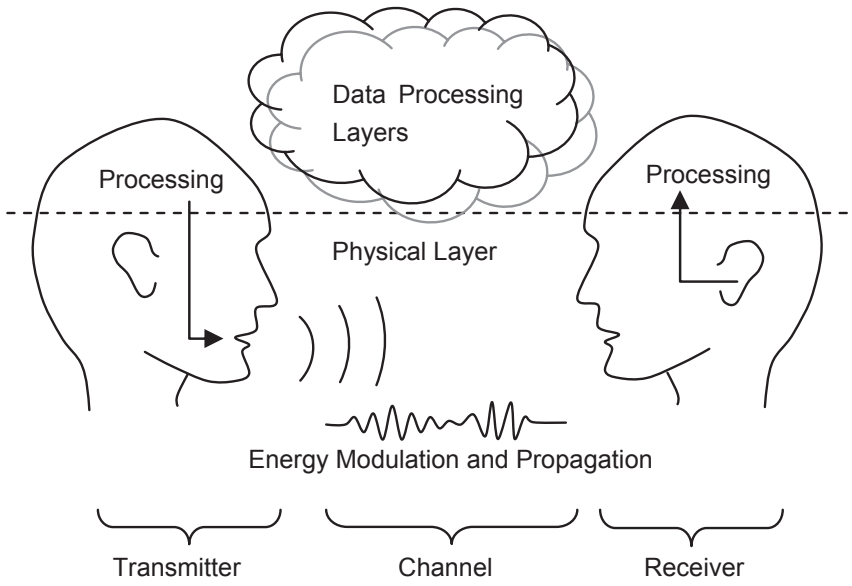

Fig. 1.1. Simple model depicting the essential elements in a communication link

Several crucial features of a communication link are evident from this example. Let us spend some time discussing these points, as they are directly relevant to all communications systems, including fiber optic links.

- First, note that the whole process can be divided into two separate domains which we have called "date processing layers" and "physical layer." This emphasizes the hierarchical nature of the process. More specifically, the physical layer does not care about the nature and details of the processing that takes place in the brain. The physical layer represents the process of converting data to a physical signal and sending that signal to the destination effectively without losing data.
- The information that needs to be transferred has to be processed and eventually converted into a representation that is understood by both sides. In the case of humans, this would be the common language. In a communication system, this would mean some form of formatting or coding, as well as a set of standards known by both the transmitter and the receiver that govern those formats and codes.
- Communication requires modulation of energy. Indeed, it is precisely the modulation of some form of energy, and not just the existence of energy (in this case sound waves), that can represent information. A constant monotonic sound cannot represent any information, even though it has energy and it travels from one point to another, unless some aspect of it is changed. For instance, it must be turned off and on. In every case, a transmitter must modulate some form of energy that carries the information from the sender to the receiver and thus enables communication. No modulation, no communication!

- Another feature is that the more information I intend to transmit to my friend in a given time, the faster I must modulate the sound waves. If I want to tell a long story about all the interesting things that I saw in my trip to South America, I need to talk much faster than if I just wanted to complain about the weather. This obvious fact lies behind much of the efforts to achieve higher modulation speeds in communication links.
- The modulated energy must travel through a medium or a channel. This channel should support the propagation and transfer of the modulated energy: obviously sound waves can propagate in the air. Electromagnetic waves do not need a physical medium, as they can propagate in free space. So in that case vacuum can also act as the channel. Moreover, anytime information is transferred in the form of modulated energy through any medium some form of degradation takes place. It may get weaker as it propagates through the medium, or it could be mixed with other unwanted noise signals, or the waveform of the signal itself may distort.
- In order to combat these degradations, either the transmitter or the receiver (or both) should somehow compensate them. If my friend is sitting further away from me, or if the restaurant is too busy and the noise level is high, I must talk louder. If he misses something that I say, he may ask me to repeat what I said or I may have to talk slower. If neither the transmitter nor the receiver is willing to modify itself to accommodate the signal degradation, the communication link may break down.

These observations, simple as they seem, are directly relevant to all practical telecommunication systems. We will revisit these concepts throughout this book on different occasions, especially as they apply to fiber optic links.

1.2 The nature of light

The main distinction of fiber optic links is that they use light as the form of energy that they modulate, and they use optical fibers to propagate that energy from the source to the destination. Indeed the main advantage of using light energy for communication is the ease with which light can be modulated with high-speed signals and transported over long distances in an optical fiber with minimal degradation. Thus, in order to understand the nature of optical communications we must start with a brief discussion about the nature of light.

In spite of the abundance of our various experiences and encounters with light, the actual nature of light remains elusive and mysterious. In ancient times, the interest in light was mainly expressed as a fascination with one of the most amazing optical instruments, i.e., the eye. Thus, there was debate between philosophers about the nature of vision, how it takes place, and how it results in the perception of shapes and colors. For instance, Aristotle, who exercised great influence on scientific thinking for centuries, explained light and vision in terms of his theoretical

concepts like potentiality, actuality, and form and matter. He thought that the form of an object, as opposed to its matter, can somehow travel through space in the form of an image and be received by the viewer. Perception then takes place when this form is impressed upon the soul. Furthermore, transparency is a potentiality in some substances, and brightness (i.e., light) is the actualization of that potential [1]. The Atomists on the other hand, and chief among them Democritus, believed that everything consisted of atoms. Therefore they thought in terms of "atoms of light" [2]. There were also theories that regarded vision and light as rays emanating from the eye and reaching toward the objects. Plato, for instance, believed that light emanating in the form of rays from the eye combines with the light of day, and the result, in the form of a ray, will reach the object [3][1].

These views, alien as they may seem today, in some ways remind us of the modern views of light. For example, Aristotle's theories about an image traveling in the air have some resemblance to the modern theory of imaging. More notably, the belief in the particle nature of light has resurfaced in twentieth century quantum physics [4,5].

In the past few centuries, discussion on the nature of light had divided scientists in mainly two camps. On one side was the particle or corpuscular theory of light. One of the most prominent supporters of this view was Sir Isaac Newton. Partly due to Newton's influence, the corpuscular theory held sway for almost a century after him. On the opposite side was the wave theory, a main proponent of which was Christian Huygens. Eventually, however, Newton's name and prestige was insufficient to overcome the weight of experimental evidence favoring the wave theory, and this is how the wave theory became the first thoroughly scientific theory that was able to explain all the known phenomena at the time.

1.2.1 The wave nature of light

Numerous optical phenomena such as diffraction and interference provide strong evidence for the wave nature of light. One of the first scientists responsible for the wave theory in its modern form was Thomas Young who is famous for his experiments on interference [6]. Eventually, however, the wave theory of light found its most eloquent expression in nineteenth century by James Clerk Maxwell and his electromagnetic theory. Maxwell combined all known phenomena related to electricity and magnetism and summarized the results in his famous four equations, which in their differential form are as follows[2] [7]:

---

[1] Although it must be said that Plato always talks through other characters, most notably Socrates. So we should be careful in assigning a view to Plato directly.

[2] In fact, the four equations that are commonly known as Maxwell's equations are more properly called Maxwell–Heaviside, as their current formulation is due to Oliver Heaviside. Maxwell's own formulation was much more cumbersome [8].

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{1.1}$$

$$\nabla \times \mathbf{H} = J + \frac{\partial \mathbf{D}}{\partial t} \tag{1.2}$$

$$\nabla . \mathbf{D} = \rho \tag{1.3}$$

$$\nabla . \mathbf{B} = 0 \tag{1.4}$$

In these equations **E** is the electric field vector, **H** is the magnetic field vector, **D** is the electric flux density vector, **B** is the magnetic flux density vector, ρ is the charge density, and ∇ is the differential operator representing space derivatives. These equations are based on earlier works from Faraday, Ampere, Coulomb, and Gauss, among others. Notice that in these equations we have two magnetic variables and two electric variables. These variables are further related to each other according to the constitutive relations[3]:

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{1.5}$$

$$\mathbf{B} = \mu \mathbf{H} \tag{1.6}$$

Here ε is the permittivity and μ is the permeability of the medium. It is through these constitutive relations that the dynamics of bound charges and currents in real materials come into play. Before moving on, we should make a few comments.

These equations summarize a body of experimental evidence, obtained earlier by scientists like Faraday, Ampere, and Gauss. More specifically, setting the time derivatives to zero, we get the familiar electrostatic and magnetostatic field equations. Moreover, there are certain symmetries between magnetic and electric fields, with one major exception, which is clearly evident from Eqs. (1.3) and (1.4). Equation (1.3) states that electric fields can "diverge" from a charge density. In other words, electric field lines can have beginnings and ends. However, Eq. (1.4) states that magnetic fields do not begin or end, which means they always have to be in the form of loops. Physically, this means that no magnetic charge or magnetic monopole exists. The lack of a "magnetic current" term similar to J in Eq. (1.1) is another aspect of this fact. This curious lack of symmetry has prompted an extensive search for magnetic monopoles, but with no results so far.

An elegant consequence of Maxwell's theory is that light is identified as a form of electromagnetic waves, and indeed all classical optics can be driven from Maxwell's equations. It can be seen from these equations that the field variables E and

[3] Technically these relationships hold only for linear, homogenous materials. More complex forms of constitutive relations must be used for more exotic materials.

H are coupled together. Thus, the equations can be combined to arrive at a single equation for a single variable. For example, in free space (or in a non-conducting dielectric) the current term J in Eq. (1.2) is zero. We can then take the curl of Eq. (1.1) and, after some manipulation including using the constitutive relations, arrive at [9]

$$\nabla^2 \mathbf{E} = \varepsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{1.7}$$

which is the well-known wave equation for electric field. The same wave equation can be obtained for the magnetic field. The general solution for wave equation can conveniently be expressed in the mathematical form using the complex notation

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{A}(\mathbf{r}) e^{j(\mathbf{k}.\mathbf{r} - \omega t)} \tag{1.8}$$

where $\mathbf{r}$ is the space vector and non-harmonic space variations are lumped into the vector variable $\mathbf{A}(\mathbf{r})$. Harmonic time variations are represented by the angular frequency ω, and harmonic space variations are represented by the wave vector $\mathbf{k}$.

The solution represented by Eq. (1.8) describes a wave propagating in the direction of the vector $\mathbf{k}$, with a speed that can be shown to be

$$c = \frac{1}{\sqrt{\varepsilon\mu}} \tag{1.9}$$

Maxwell himself realized that the speed calculated from Eq. (1.9) was remarkably similar to the available experimental measurements of speed of light. He concluded that light waves are transverse electromagnetic waves which like any other wave can be described by amplitude, frequency, and wavelength. Therefore, electromagnetic waves in general and light in particular are in fact one class of solutions to Maxwell's equations.

Because electric and magnetic fields are vectors, in general they have components in x, y, and z directions (in the Cartesian coordinates). However, in fiber optic applications, we are generally interested in waves that propagate along a single direction, for instance along the fiber, or along the optical axis of a device. In such cases we can write the wave solution as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{A}(\mathbf{r}) e^{j(kz - \omega t)} \tag{1.10}$$

where now z is the direction of wave propagation, and k is the component of wave vector **k** in the z direction.[4] Equation (1.10) can describe a plane wave traveling in the z direction or a guided wave propagating along a wave guide such as an optical fiber. We will revisit this equation in Chapter 5 where we discuss light propagation in optical fibers. Note that in Eq. (1.10) the electric field **E** (and the field amplitude **A**) is still a vector and a function of three spatial coordinates.

Let us now analyze the term ($kz-\omega t$) in Eq. (1.10). The angular frequency ω is given by $\omega = 2\pi f$, where f is the frequency in Hertz and is related to the oscillation period T as $\omega=2\pi/T$. Moreover, k is related to the wavelength λ as $k = 2\pi/\lambda$. Therefore, k represents the periodicity of the waves in space, and ω represents the periodicity in time. To better appreciate the meaning of these quantities, we should remember that a phase front is a point in the wave where the phase does not change. In terms of Eq. (1.10), this means all the points in space where the argument of the exponential function is constant. Setting the argument equal to a constant and differentiation with respect to time, we obtain

$$\omega - k\frac{dz}{dt} = 0 \tag{1.11}$$

The term dz/dt can be recognized as the speed of the wave along the z direction, which is in fact the speed of light. After substituting the wave vector and angular frequency with wavelength and frequency, we arrive at the simple but important relationship between the speed of light, its wavelength, and its frequency:

$$c = f\lambda \tag{1.12}$$

Equation (1.12) simply states that the speed of light is equal to the number of wavelengths that pass from a point in space in a second (frequency) times the length of one wavelength. Of the two quantities on the right-hand side of Eq. (1.12), f is determined by the source and, under normal circumstances, remains constant as the wave propagates through various media. In other words, we can think of frequency as an inherent property of the light, i.e., a property that is determined by the source and (usually) remains the same once the light is generated. On the other hand, λ changes as the wave goes from one medium to the next.

From an optical point of view a medium is characterized by its index of refraction, n. The index of refraction of vacuum is one, but other media have indices of refraction greater than unity. If we represent the vacuum speed of light and wave length by c_0 and λ_0, respectively, we obtain $c=c_0/n$ and $\lambda=\lambda_0/n$. In other words, both the speed of light and its wavelength decrease as it enters a medium with an

[4] Technically, we should add a subscript z to k and denote it as k_z as a reminder of the fact that it is the z component of the wave vector **k**, and that the wave vector in general has other components as well. However, to keep the notation simple, we will drop the z subscript whenever we are only interested in propagation along the z direction.

index of refraction of n. Therefore, the speed and wavelength are extrinsic properties of light, i.e., properties that are affected by the medium.

A harmonic wave described by Eq. (1.10) may seem too artificial. But we can think of it as a building block for constructing more complex waveforms. This is possible because of a very important property of Maxwell's equations: their linearity. This means that if we find two solutions for the equations, for example two plane waves with two different frequencies or amplitudes, any linear combination of those two plane waves is also a solution to the equations. In this way, complex waveforms with arbitrary profiles in space or time can be studied.

1.2.2 The particle nature of light

In spite of the success of Maxwell's theory in describing a wide range of optical phenomena, in twentieth century the picture once again changed, especially with the advent of quantum theory. Although a wide range of phenomena are best explained through the wave nature of light, there are other instances, such as the photoelectric effect, that are easier to explain by taking light to consist of individual packets of energy, called photons. Later in the twentieth century certain phenomena such as the Lamb shift and photon antibunching were discovered that, unlike the photoelectric effect, do not have any classical field explanation. As a result, although quantum field theories present fundamental challenges to our classical notions of reality, quantum optics is now a well-established and growing field [10–13]. In fact, quantum light states in which a precise and known number of photons are present can now be realized experimentally [14].

One of the first triumphs of quantum theory of light came when Max Planck realized that he could explain the problem of black body radiation spectrum by assuming that the electromagnetic energy could be radiated or absorbed only in multiples of a minimum amount of energy, or quantum of energy, whose value is directly proportional to the frequency of electromagnetic radiation [15]. The result is the well-known relationship:

$$E = hf \tag{1.13}$$

where E is the energy in Joules and h is the Plank's constant whose value in the international system is 6.623×10^{-34} Joules times seconds.

Note that by itself, all Eq. (1.13) claims is that the energy of a photon is proportional to its frequency. When it comes to interaction of light with atomic systems, the quantization of energy comes from another equation, called the Schrödinger equation, whose steady-state solutions in a potential field are discrete. One of the most important (and abundant!) examples of such a system is the atom. The nucleus provides the potential field, and the steady-state solutions of the Schrödinger equation result in the familiar discrete energy levels the electrons can occupy. According to Maxwell's equations an electron orbiting a nucleus is an accelerating

charge which must radiate electromagnetic energy. If it were not for energy quantization and if the electron could change its energy continuously, it would have to radiate all its energy in a flash and crash into the atom's nucleolus.

However, the electron can move between the allowed discrete levels in an atom. If an electron moves from a higher energy level to a lower energy level, conservation of energy requires that energy be released in some other way. A radiative transition is one where the energy difference is released in the form of a photon whose frequency is related to the energy difference according to Eq. (1.13). That is why this equation is of primary importance to lasers. We can think of a laser as a system with two energy levels. When the electrons are somehow pushed into the higher level, a situation known as population inversion is created. When they jump back to the lower level in a coherent manner, they produce coherent electromagnetic waves, or laser light. The frequency of the light is determined by Eq. (1.13). In semiconductors, the two levels are typically the valence band and the conduction band, and the energy difference between the two levels can be as high as a few electron-Volts.[5] By knowing this energy difference, we can calculate the frequency or the wavelength of the light generated by that laser through Eq. (1.13).

1.2.3 The wave particle duality

So is light wave or particle? This is for sure an interesting and still challenging question for physicists. As a result of both theoretical and experimental evidence and in spite of persistent mysteries in interpretation of the evidence, it has now been accepted that light has a dual nature in that it can behave both as particle and as wave [16–18]. In fact, such a duality is not limited to light. According to quantum theory, not only light waves can have particle-like properties, particles of matter can and do have a wave-like nature. One way to appreciate this dual nature is through the well-known de Broglie equation:

$$\lambda = \frac{h}{p} \tag{1.14}$$

which postulates a wavelength λ for *any* particle with a momentum of p. Thus, an electron that is moving with a certain speed can show wavelike behavior, something indeed verified experimentally. The reason that cars and billiard balls do not act like waves is that due to their large masses, and because of the small value of Planck's constant, the wavelength associated with them is so exceedingly small that for all practical purposes it can be neglected.

---

[5] An electron-Volt (eV) is the energy needed to move an electron up a potential barrier of 1 volt. We remember that increasing the potential of a charge of 1 Coulomb by 1 Volt requires 1 Joule. Thus, 1 eV is 1.60×10^{-19} J, because the electron charge is 1.60×10^{-19} C.

The denominator of Eq. (1.14) represents the momentum of a particle. When it comes to electromagnetic waves, Maxwell's theory indeed predicts a momentum for the wave. For uniform plane waves, we have [6]

$$p = \frac{E}{c} \tag{1.15}$$

where E is energy (per square meter per second), p is the momentum (per square meter of cross section), and c is the speed of light. If we substitute p from Eq. (1.15) for p in Eq. (1.14) and use Eq. (1.12) to convert wavelength to frequency, we arrive at the familiar Planck equation, Eq. (1.13). This shows that the pieces of the puzzle fit together once we recognize that the wave-like and particle-like behaviors are not contradictory but complementary.

Equation (1.14) yields another important insight. When the energy of a wave increases, so does its momentum, and an increase in momentum means a decrease in wavelength. Thus, we can expect high-energy waves to show particle-like behavior more clearly. In fact, gamma rays, which are the shortest wavelength and highest energy form of the electromagnetic spectrum, behave not like waves, but like rays of high-energy particles. On the other hand radio waves, which are in the lower side of the electromagnetic spectrum, have such low energies that for all practical purposes their particle nature can be neglected.

What is the practical relevance of all these concepts to engineering applications? From a practical point of view, the light behaves more like particles when it is being generated or detected. On the other hand, when it comes to propagation of light, it behaves more like waves. This approach has resulted in a view known as semiclassical theory of light. If we want to study the generation or detection of light in such devices as semiconductor lasers and detectors, we use the quantum physical approach and think of light as photons. When it comes to propagation of light in free space or other media, we use classical field theory, as described by Maxwell's equations.

1.3 The electromagnetic spectrum

As mentioned in the previous section, when it comes to light propagation, we can safely treat it as an electromagnetic wave. The electromagnetic spectrum covers a wide range of frequencies, and visible light occupies only a small fraction of it [19]. It is imperative for engineers to gain an overall understating of this spectrum. In fact, a large portion of electrical engineering deals with frequencies that correspond to the low side of this very same spectrum. In these lower frequencies, we can manipulate signals through electronic devices such as transistors. The optical frequency is far too high to be handled directly by electronic devices. Therefore, in fiber optic applications we are in fact dealing with two separate bands of the

spectrum: the optical frequency and the much lower modulation frequencies. Thus, it is doubly important for optical engineers to be familiar with the characteristics of electromagnetic waves at different regions of the spectrum.

An overview of the electromagnetic spectrum is shown in Fig. 1.2. Electromagnetic waves can be characterized according to their frequency or wavelength. Generally, at the low side of the spectrum working with frequency is more convenient. Thus, we have the AM radio band which functions in the range of hundreds of kilohertz or VHF and UHF TV signals that operate up to the range of several hundred megahertz. The microwave and millimeter range includes wavelengths in the range of roughly 1 cm–1 mm. The corresponding frequency varies roughly from a gigahertz to tens of gigahertz. Above these frequencies the ability of electronic circuits to modulate the electromagnetic waves starts to diminish. But the spectrum itself continues into the infrared region, where most fiber optic communication links operate. Here the frequencies start to get so large that it becomes more convenient to talk in terms of wavelength. That is why the higher regions, including the visible light, are usually characterized by wavelength. Beyond visible light and at shorter wavelengths, we have ultraviolet, X-rays, and finally gamma rays.

One way to gain a better insight into the behavior of electromagnetic waves in practical systems is to divide the spectrum into three regions, based on the ratio of the system's physical dimensions to the wavelength of signals of interest. Depending on this ratio, the behavior of signals can be studied by application of circuit theory, wave theory, or ray theory. If we denote the physical size of the system that we work with as D, the three regions are as follows:

- $D << \lambda \Rightarrow$ circuit theory
- $D \approx \lambda \Rightarrow$wave theory
- $D >> \lambda \Rightarrow$ ray theory

When the wavelength is much larger than the dimensions of our system, we are in the domain of circuit theory. In this regime we can neglect the wave nature of the electromagnetic energy, assume instantaneous energy propagation within the system (infinite wave speed), and use simplified lumped circuit analysis. This is where most of conventional electronic circuits operate. Thus, we can think of circuit theory as the low-frequency approximation to Maxwell's equation.

The other extreme is the case where the wavelength is much shorter than the dimensions of our system. In this case we can also neglect the wave nature of electromagnetic waves and treat them as rays. The best example of this approximation is geometrical optics. The mid range, however, is where we cannot utilize either of the above approximations. Here we have to use wave theory which in its most complete form is expressed by Maxwell's equations, although depending on the application oftentimes certain simplifying assumption are made here too.

The above categorization can also illuminate various modes of transmission in each region of spectrum, which is also shown in Fig. 1.2. Theoretically, all electromagnetic waves can propagate in free space. However, at low frequencies the most efficient mode of transferring electromagnetic energy is through conducting wires. This is the domain of circuits and lumped elements.

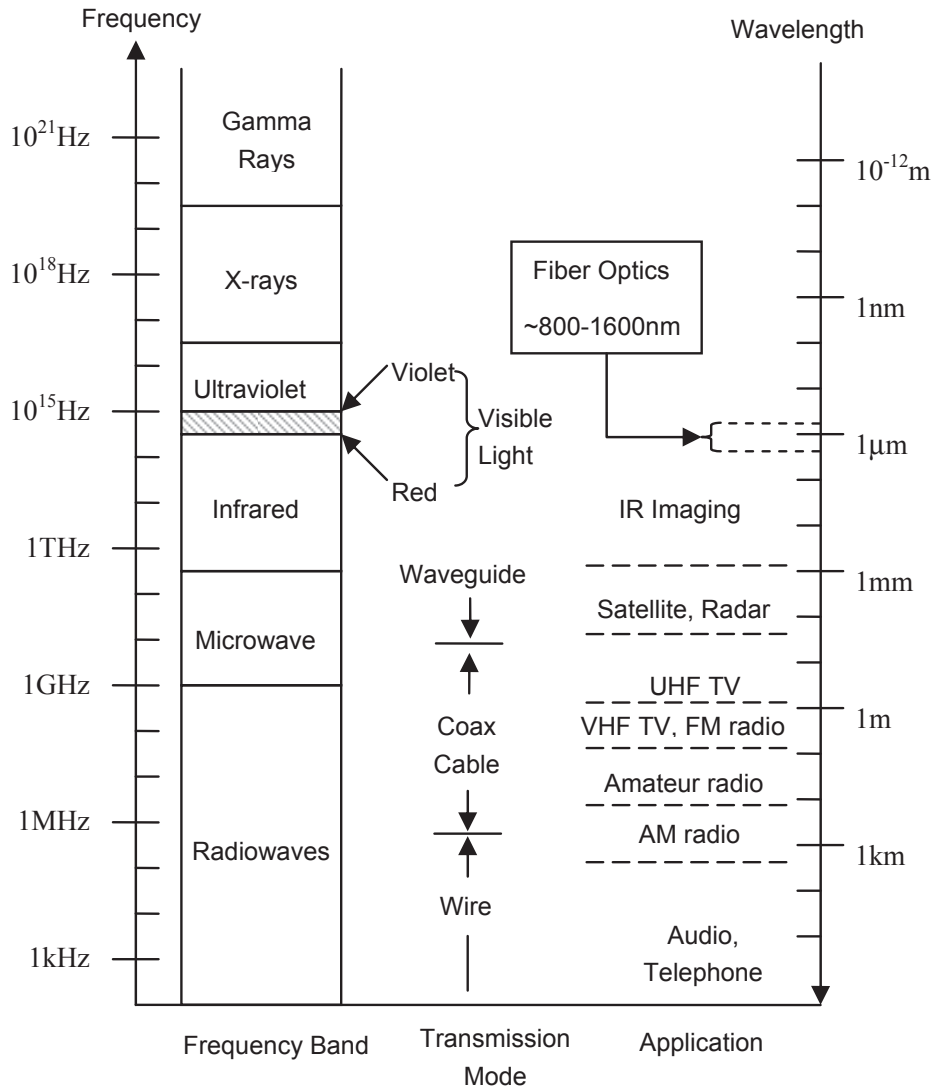

Fig. 1.2. The electromagnetic spectrum

But as the frequency is increased, the wavelength decreases, and the wave nature of the signals must be taken into account. That is why for transmission distances comparable to the wavelength we need to use a controlled impedance medium, such as a coaxial cable. Still at higher frequencies, a waveguide must be used. In a waveguide we have essentially electromagnetic waves subject to the boundary conditions forced by the geometry of the waveguide. This is the case for microwave frequencies where hollow pipes guide the electromagnetic energy in the desired direction. In fact, the optical fibers that are used to guide light waves at much higher optical frequencies are also waveguides that enforce boundary conditions on light waves and hence prevent them from scattering in space.

Figure 1.2 also shows the relatively small portion of the spectrum used in fiber optics. The most common wavelengths used in fiber optic communication range from 800 to 1600 nm, which happen to be mostly in the infrared range. The reasons these wavelengths are particularly attractive for optical communication have to do with both light sources and the fiber medium. Many useful semiconductor laser structures have bandgap energies that fall in this range, which makes them efficient light sources at these wavelengths. On the other hand, propagation losses in silica fibers reach their minimum values in this range. The availability of efficient light sources and suitable propagation properties of fibers make this range of wavelengths the optimal choice for fiber optic communications.

1.4 Elements of a fiber optic link

We started this chapter with a model of communication based on common language. With that background in mind, and with the insights into the electromagnetic spectrum, we can now narrow down the discussion further and go over the building blocks of a typical fiber optic link, as shown in Fig. 1.3.

First, notice that the figure is divided vertically into two domains: a physical layer and a set of data processing layer(s). This is meant to be a schematic reflection of the layered structure of most fiber optic systems. The processing layers are where complex signal processing functions such as multiplexing and demultiplexing, error detection, routing, and switching take place. This is equivalent to what happens in the brain in our language model of Fig. 1.1.

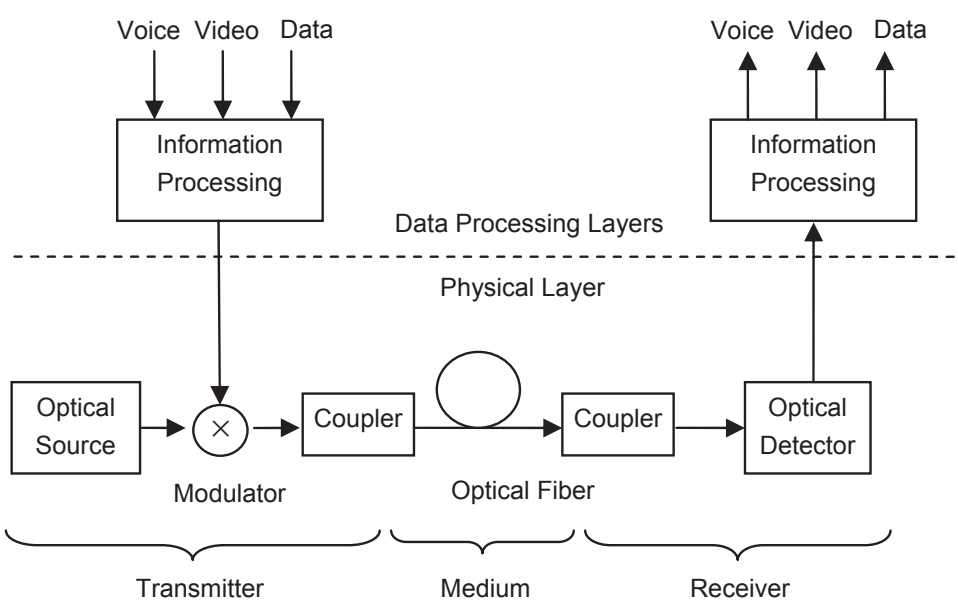

Fig. 1.3. Elements of a typical fiber optic link

The physical layer, on the other hand, does not care (at least directly) about the contents or formatting of the data. Its purpose is to convert the electrical data to optical data, send the optical signal over the fiber, receive it on the other side of the fiber, and convert it back to electrical data. Once this is done, the data will be sent back to the higher levels where further processing in the electrical domain will take place to recover the actual contents in their original format.

The exact architecture of the data processing layers varies and depends on the application. In a simple system, for instance, it may consist of a simple transducer that converts some physical quantity, for example, sound energy, into an electrical signal. That electrical signal then is converted to optical signal and sent over the fiber. On the other hand, in a complex system, the processing layers may include a complicated structure with many layers, each specialized at performing a specific task. Moreover, in such a system each layer makes the layers below itself "transparent" to layers above it. By doing so, each layer can focus on a specific task without having to worry about the details of what goes on below or above it.

As we move up and away from the physical layer, the functions become increasingly algorithmic and require complex data processing. Therefore, in general, moving down toward the physical layer means more hardware, and moving up toward processing layers means more software. In this book we are mainly concerned with the physical layer, although major concepts related to processing layers are reviewed in Chapter 2.

As can be seen from Fig. 1.3, the physical layer consists of an optical transmitter, an optical fiber or channel, and a receiver. The optical transmitter includes a light source along with a modulator. It provides the necessary optical energy that carries the information along the fiber. In practice, this source is either an LED or a semiconductor laser. The modulator's function is to modulate the light from this source with the serial sequence of data. Modulators can be divided into two broad categories. *Direct modulators* modulate the light by directly controlling the current that is running through the light source. As a result, in a direct modulation scheme the same device, say a semiconductor laser, both generates and modulates the light. Direct modulation schemes are very attractive because of simplicity and low cost and are preferred at lower modulation speeds. *External modulators*, on the other hand, do not generate the light themselves. Instead, they manipulate the constant or CW light that is generated by a separate optical source. As a result, external modulation is more complex and costly. However, external modulators can operate at higher speeds and provide better performance.

Once the optical signal with the desired modulation is generated, it must be coupled into the fiber. This is done through a coupler. The reason a separate block is dedicated to this function in Fig. 1.3 is to emphasize its non-trivial nature and the involved challenges. Unlike electrical signals that can easily be coupled from one conductor to the next through physical contact, transferring optical signals in fiber optic systems involves more mechanical challenges and requires, among other things, careful alignment. Even then, every time an optical signal is transferred from one medium to another a portion of it is lost.

Once the optical signal is coupled into the optical fiber, it can generally propagate for long distances with relatively little degradation. The exact nature and amount of these degradations is a function of the structure of the fiber, wavelength, and spectral width of the optical source. *Single-mode* fibers can generally support longer distances with much less degradation, while *multimode fibers* are suitable for shorter distances.

The final stage in the link is the receiver. Once the light reaches the other side of the link, it has to go through another coupler which directs the light to an optical detector. The detector converts the modulated optical signal to an electrical signal. However, the electrical signal coming out of a detector is generally too weak to be useful for further processing. As a result, the receiver must provide additional amplification in order to bring up the amplitude of the electrical signals within acceptable levels. This is typically done through adding a *preamplifier*, or a *transimpedance amplifier*, immediately after and at close physical proximity to the detector. The receiver may also perform further signal processing and conditioning after the preamplifier stage. At any rate, the receiver must provide a clean replica of the original signal at its output, something that can then be passed up to the processing layers, where various functions such as demultiplexing, error correction, and routing take place.

1.5 Light sources, detectors, and glass fibers

In the previous section we discussed some of the main characteristics of fiber optic links in general terms. In order to complete our general discussion, it is necessary to have a brief review of optical sources, optical detectors, and the fiber.

1.5.1 Optical sources

The role of the optical source is to generate light energy that once modulated carries the information across the fiber. Although there are a multitude of ways to generate light, in almost all fiber optic systems a semiconductor device is used for this purpose. These devices include diode lasers and light-emitting diodes (LEDs). There are many reasons for using these devices. For one thing, they are generally easy to operate and integrate with electronic circuits. From a circuit standpoint, diode lasers and LEDs all behave like a diode. In order to turn them on, a forward voltage must be applied to them, which in turn results in a current flow, which in turn will turn the device on. To increase the optical power, all we need to do is to increase the current. To turn the device off, we just need to turn off the diode by shutting down the current. Thus, in many applications the modulation of light is achieved by directly modulating the current flow through the device.

This method, called direct modulation, is very convenient and can be utilized up to modulation rates of several gigahertz. Figure 1.4 shows a simple circuit for converting an electric signal to an optical signal using a semiconductor laser.

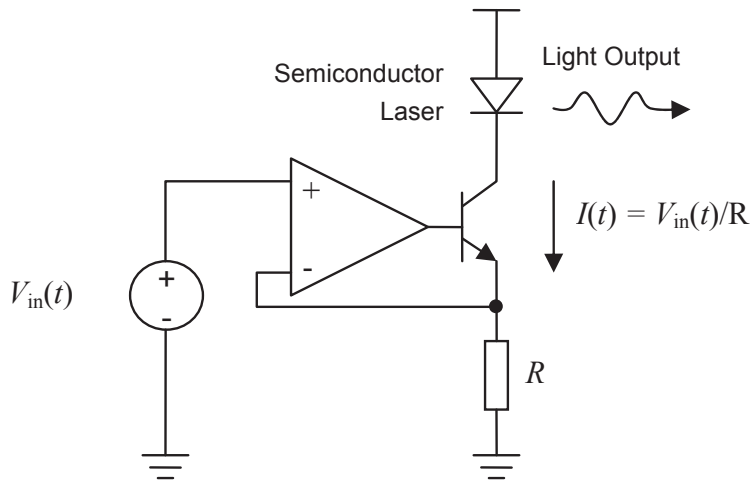

Fig. 1.4. Simplified laser driver circuit

The circuit converts the voltage of the source to a proportional current signal that flows through the laser. Because the laser behaves more or less linearly above its threshold current, the voltage signal is converted to a similar optical signal. Although this simple circuit has several shortcomings, it serves well to illustrate the basic idea behind a *laser driver*.

Another advantage of semiconductor lasers is that these devices are generally very efficient light sources. Semiconductor lasers are characterized by their *threshold current* and *slope efficacy*. A typical semiconductor laser optimized for high-speed communication applications may have a threshold current of 5 mA and a slope efficiency of 0.5 mW/mA at room temperature. This means that it will start generating light as soon as the current flowing through it exceeds 5 mA. After that, the optical output power increases by 0.5 mW for every milliampere increase in current. Thus, at 10mA above threshold, this device can generate 5 mW of optical power. If we assume a forward voltage of 2 V, this represents a power dissipation of 2 V×15 mA=30 mW. The overall efficiency of this device at this particular biasing point is 5/30=0.17, which is not bad at all.

Semiconductor devices also have an important advantage in terms of the optical wavelengths they produce. The wavelength generated or absorbed by a semiconductor is a function of its bandgap energy. The bandgap of semiconductors used in optoelectronics is in the range of 0.5–2 eV. This range corresponds to a wavelength range of approximately 500–4000 nm, which happens to include the wavelength windows of 1300 and 1500 nm, where the attenuation of glass fiber is minimum. Therefore, the light output of these devices can propagate in fibers for long distances with little attenuation. Semiconductor devices are also very small in

size, generally cheap, and very reliable. They can be produced in large quantities through wafer processing techniques similar to other semiconductor devices. All these advantages make these devices ideal sources for numerous optical communication applications.

We mentioned that semiconductor light sources used in optical communications can be either LEDs or lasers. LEDs are cheaper, and thus they are mainly used in low data rates or short-reach applications. The main disadvantage of LEDs is that their light output has a wide spectrum width, which in turn causes high dispersion as the light propagates in the fiber. Dispersion causes the smearing of sharp edges in a signal as it propagates in the fiber and is directly proportional to the spectral width of the source. This is why LEDs cannot be used for long distance or high modulation rates in optical communication. Semiconductor lasers, on the other hand, have much narrower spectral widths, and therefore they are usually preferred in high-speed or long-reach links. In this book, we focus our discussions on lasers.

The properties of these lasers depend on the materials used in constructing them as well as the physical and geometrical structures used in their design. Generally speaking, a laser is an optical oscillator, and an oscillator is realized by applying feedback to an amplifier. Semiconductor lasers can be divided into two main categories depending on the nature of this feedback.

In a Fabry–Perot (FP) laser, the feedback is provided by the two facets on the two sides of the active region. The optical cavity in FP lasers generally supports multiple wavelengths. Therefore, the output spectrum, although much narrower compared to an LED, still consists of several closely located peaks, or optical modes. A distributed feedback (DFB) laser, on other hand, includes additional structures that greatly attenuate all but one of those modes, and therefore a DFB laser comes closest to producing an ideal single wavelength output. This is why DFB lasers can minimize dispersion and support the longest attainable reaches.

Both FP and DFB lasers are *edge-emitting* devices, i.e., the light propagates in parallel to the semiconductor junction and comes out from the sides. A different structure is the vertical cavity surface emitting laser, or VCSEL. In a VCSEL the light output is perpendicular to the surface of the semiconductor. VCSELs are very attractive because of low threshold current and high efficiency. Moreover, many VCSELs can be integrated in the form of one-or two-dimensional arrays, a feature not available for edge-emitting devices.

However, VCSELs usually work at short wavelengths around 850 nm, a wavelength unfortunately unusable for long haul communication. Although research into production of VCSELs at longer wavelengths of 1300 and 1500 nm is intense, such devices are still not available in an industrial scale [20]. As a result, VCSELs like LEDs are usually used for short-distance links. We will come back to these various laser structures in Chapter 4 where we discuss their principles of operation and properties as well as their advantages and disadvantages in more depth.

1.5.2 Optical detectors

Like optical sources, the optical detectors used in fiber optics are almost exclusively semiconductor devices in the form of PIN diodes and avalanche photo diode (APD) detectors. From a circuit perspective, these are again diodes that like any other diode (including laser diodes or LEDs) can be forward or reverse biased. However, for them to act as detector, they must be reverse biased. A PIN diode can operate with a very low reverse bias. This makes the PIN diode attractive because it can be operated as an element within a standard electronic circuit that runs at a low supply voltage of say 3.3 V. Figure 1.5 shows a simple detector circuit. The output of the detector is a current that is proportional to the light power received by the detector. Thus, the detector can be considered as a light-controlled current source. A transimpedance amplifier (TIA) converts the photocurrent generated by the detector to a voltage.

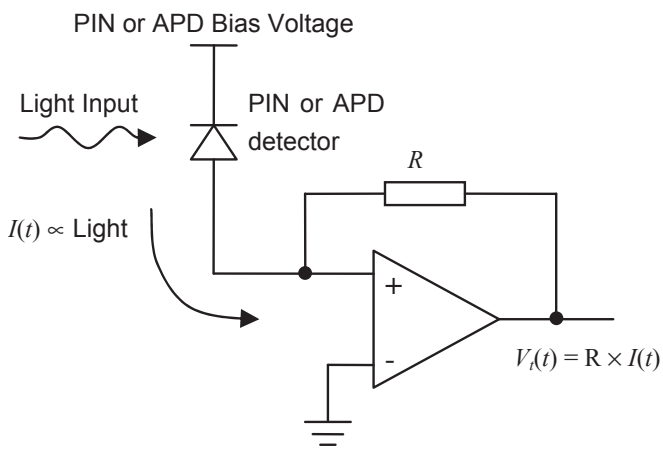

Fig. 1.5. Simplified detector using PIN diode or APD

APD detectors are not very different from a circuit point of view. However, they require a much higher reverse bias voltage to operate. Currently low-voltage APDs require at least 30–40 V of reverse bias. Moreover, the required optimal reverse bias voltage for an APD varies with temperature. As the temperature rises, the reverse bias applied to the APD must also increase to maintain the gain of the APD constant. Because of these complications, APDs are harder to use and more care must be taken in designing them. They are also more costly in terms of both the APD itself and the additional required support circuitry. However, APDs have a big advantage that makes them attractive for high-end receivers: unlike PIN detectors, an APD detector provides inherent current gain. This additional gain directly translates to improved performance. Typically an APD can work with a fraction of the optical power needed for a similar PIN diode, and this means longer links can be supported with an APD receiver.

1.5.3 The optical fiber

Perhaps the most critical factor that has made fiber optic communication possible is the development of low-loss silica fibers. The idea that light can be "guided" in a medium other than air is not inherently revolutionary. The physical principle behind such guiding is the refraction of light, as summarized by Snell's law:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{1.16}$$

where θ_1 and θ_2 are the angles of incidence and refraction, and n_1 and n_2 are the indices of refraction of the two media (Fig. 1.6). This means that when light enters from a medium with a lower index of refraction to a medium with a higher index of refraction, it bends toward the line normal to the two media. Going the opposite direction, the light bends away from the normal direction. In this case if the angle of incidence is more than what is known as the critical angle, θ_T, all the energy reflects back into the medium with the higher index of refraction. This phenomenon is called *total internal reflection*. This is what happens if you try to look outside from under the water. You can see things that are immediately above you, but as you look away, beyond a certain angle the surface of water acts like a mirror, reflecting light from under the water while blocking light from outside. The same principle is behind the light guiding properties of an optical fiber.[6] An optical fiber consists of a *core* with higher index of refraction and a *cladding* with a lower index of refraction. If light is launched into the fiber either in parallel with its axis or with a small angle, total internal reflection will trap all the energy inside the fiber and the optical wave will be guided along the axis of the fiber.

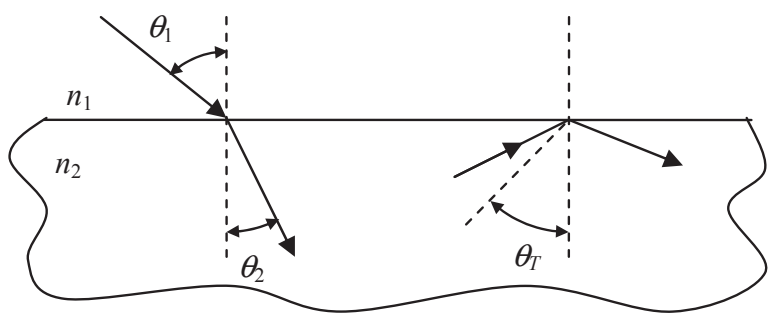

Fig. 1.6. Snell's law of refraction and total internal reflection

This would constitute a *guided mode* that the fiber can support. If the light enters with a large angle with respect to the axis of fiber, at each reflection a portion of energy is radiated out, and therefore after a few reflections the energy inside the

[6] To be more accurate, multi-mode transmission of light can be understood in terms of total internal reflection. However, single-mode transmission requires the wave theory of light. We will discuss this point in Chapter 5.

fiber will attenuate to undetectable levels. This would be an example of a *leaky mode*, i.e., a mode the fiber cannot support.

An optical fiber can normally support several optical modes depending on the thickness of its core. There are two ways to look at wave propagation in a fiber. Based on geometrical optics, we can think of each mode as being a certain path for light rays as they bounce back and forth between the sides of the fiber core while traveling forward. However, a more accurate description requires the wave theory of light, where each mode is like a resonant state in which Maxwell's equations along with the appropriate boundary conditions are satisfied. As the size of the core reduces, so does the number of modes that it can support. Of particular interest is the case where the size of the core is so small that only one mode can propagate in the fiber in parallel with the fiber's axis. This would be a *single-mode fiber*. In a single-mode fiber the loss can be as low as 0.25 dB/km.[7] This means after 1 km, more than 90% of the original optical power remains in the fiber. This is indeed a remarkable degree of transparency. Imagine a piece of glass 1 km thick that only absorbs 10% of the light going through it! In comparison, normal glass starts to become opaque only after a few centimeters. This is why an optical signal can travel as much as 100 km or more in a fiber without any additional amplification and still be detected.

1.6 Advantages of fiber optics

In the beginning of this chapter when we introduced the model of human communication through language, we mentioned the rather obvious fact that in order to transfer more data in a given time, one needs to talk faster. In communication systems, this is equivalent to increasing the modulation speed of signals. In an electrical link, the transmitter and the receiver are connected through a conductive path, for example a pair of twisted wires or a coax cable. However, copper connections suffer from increased attenuation of signals at higher frequencies. The sources of these losses include the skin effect in the conductors, radiation, and loss within the dielectrics and the shields [21].

Optical fibers, on the other hand, provide much higher bandwidths. The frequency bandwidth corresponding to the wavelength range of 850–1600 nm is roughly 100 THz. With advanced techniques, single channel modulation speeds in excess of 50 GHz have been demonstrated and components for 40 Gbps transmission are commercially available [22]. Much higher aggregate data rates can be achieved through wavelength division multiplexing (WDM). As long as the optical power remains low, the fiber acts as a linear medium, which means various wavelengths of light can travel simultaneously within the same fiber without affecting each other. It is this property that is used in WDM systems to increase the link capacity even further. In a WDM system, several wavelengths are launched

[7] Later in this chapter we will discuss the decibel unit of power.

into the fiber in parallel, each modulated with an independent stream of data. At the other side of the link, optical filters separate each wavelength, sending each signal to a separate receiver. In this way the available bandwidth of the fiber is multiplied by the number of wavelengths or channels. For instance, in a course WDM (CWDM) system, four to eight wavelengths are used. In a dense WDM (DWDM) system 80 or more channels may be used. Using advanced techniques, transmission with aggregate data rates well into terra bits per second (Tbps) have been demonstrated [23–25].

We should mention that regardless of the inherent wide bandwidth of the fiber and the multiplexing techniques that utilize that bandwidth, optical signals suffer degradation mainly due to attenuation and dispersion. However, and in spite of these degradations, the bandwidth of an optical fiber far exceeds that of a similar coax cable or electrical conductor. For instance, sending an optical signal at the rate of 10 Gbps in a single mode fiber over a distance of 10 km is well within the capabilities of commercially available parts. If we assume a voice channel takes up 100 Kbps of bandwidth, a single strand of fiber can support around 100,000 simultaneous voice channels. To establish this bandwidth capacity through copper cables requires massive amounts of parallel cabling and a lot of physical space and supporting infrastructure.

The benefits of fiber are not limited to higher bandwidth and lower volume of physical infrastructure. Because fibers carry information in the form of light, they are immune to external electromagnetic interference. A copper cable can act as an antenna, receiving electromagnetic radiations from other manmade or natural radiation sources. Typical manmade sources include the power grid and radio and TV stations. Typical natural sources of noise include lightening or microwave radiation from outer space. None of these sources can interfere with the optical signals in a fiber. This isolation works both ways, i.e., fibers do not radiate energy either. This makes them more suitable for applications where security is a matter of concern. Once a fiber link is established, it is very difficult, if not impossible, to "tap into it" without breaking it. Another advantage of fiber links is that they do not need to carry electrical signals, and more generally, they do not require a conductive path between the transmitter and the receiver. This makes them ideal for applications where the link must pass through an environment where presence of electrical signals poses safety risks or cases where it is desired to keep the two sides of the link electrically isolated.

1.7 Digital and analog systems

So far we have not distinguished between the various ways in which information can be represented. In the physical world, and at least in the macro-level which we deal with in everyday life, most variables are analog. This means that the quantities of interest vary continuously within a range of possible values. In fact we perceive most of the physical quantities that we sense, like light, heat, and pressure,

in an analog manner. Likewise, the most fundamental electrical quantities that represent signals are analog. The voltage of a point in a circuit or the current that flows through an element is an analog variable. However, in spite of the analog nature of these physical quantities, most fiber optic communication links are categorized as digital and function in a digital manner. The reason behind this is that digital signals are more immune to noise, less sensitive to nonlinearities, and easier to produce and detect.

From a historic perspective, the information revolution is inherently tied with computers. Internet, after all, was invented to connect digital computers together. Likewise, all the applications that somehow are related to computers are naturally digital and deal with digital data. This is why when it comes to fiber optics, not only the physical layer, but also most of the backbone data-intensive links are digital, precisely because they have to handle digital data. In spite of this, analog links are still used in some applications, such as video transmission and RF feeds [26]. For this reason, throughout this book, we mostly focus on digital links, although whenever appropriate we also include discussions on analog links.

1.8 Characterization of fiber optic links

So far we have reviewed the building blocks that make up a fiber optic link. However, it is also important to characterize the performance of a particular link in a quantitative way. This is particularly important when a link is being designed. We want to know, for instance, what kind of components are needed in order to achieve our target performance characteristics. Alternatively, we may want to know what kind of *margin* an existing link is operating at. If a link is working with little or no margin, it means a small change in the performance of any component used in the link may cause problems. This is why we need the concept of *link budgeting* in order to quantify the performance of a link.

We should start by discussing the important notion of decibel or dB, which is a logarithmic measure of the ratio of two optical powers. The ratio of two optical powers P_1 and P_2 in decibels is defined as follows:

$$dB = 10\log_{10}\left(\frac{P_1}{P_2}\right) \tag{1.17}$$

For example, if P_1 is ten times larger than P_2, we can say P_1 is 10 dB higher than P_2. Equation (1.17) also provides a way of measuring power as long as we agree on a reference power level. By definition, a power level of 1 mW is used for such a reference. That allows for defining a logarithmic scale, referred to as dB-milliwatt or dBm, to measure optical power:

$$P_{\text{dBm}} = 10\log_{10}(P_{\text{mW}}) \tag{1.18}$$

where P_{dBm} is the power in dBm and P_{mW} is the power in units of mW. Note that Eq. (1.17) is the same as Eq. (1.18) when P_2 is set to 1 mW. It should be noted that in both these equations the logarithm is defined in base 10.

One may question the reason for using a logarithmic scale instead of the more familiar linear scale. In a linear scale, the optical power is usually expressed in units of milliwatt. Likewise, the ratio of two powers (or any other quantity) is just a number, for example 0.5 or 10. So why should we go to the trouble of calculating logarithmic powers and ratios? The answer lies in some of the very useful properties of the logarithm function.

Remember that logarithm is an operator that converts multiplication to addition, and exponents to multiplication. In other words: $\log(ab)=\log(a)+\log(b)$, and $\log(a^b)=b\times\log(a)$. These properties become handy in link budget calculations because the optical power generally decreases exponentially in optical fibers with distance. That is why the attenuation of a particular kind of fiber is usually given in units of decibel per kilometer. For example, a single mode fiber at the wavelength of 1310 nm has a typical attenuation of 0.5 dB/km. This means that for every kilometer of this fiber, we should expect 0.5 dB reduction in power. Another reason that using logarithmic scales is useful is the fact that gain and loss are inherently multiplicative. Therefore, whenever we want to calculate the effects of gains or losses, logarithmic scales allow us to add or subtract instead of multiply or divide. To summarize, once we are used to thinking in terms of decibels, it becomes much easier to understand and visualize the status of an optical signal or an optical link without going through more complex calculations. Table 1.1 summarizes approximate equivalences between linear and logarithmic powers and ratios.

Table 1.1. Linear ratios (powers) and their logarithmic equivalent in dB (dBm)

Amplification/power higher than 1 mW		Attenuation/power lower than 1 mW	
Ratio/mW	dB/dBm	Ratio/mW	dB/dBm
1000	30	1	0
100	20	0.99	−0.05
10	10	0.9	−0.5
5	7	0.8	−1
3	5	0.5	−3
2	3	0.3	−5
1.25	1	0.2	−7
1.1	0.5	0.1	−10
1.01	0.05	0.01	−20
1	0	0.001	−30

It is very useful to remember some of the more common ratios as rules of thumb. For example, doubling the power means adding 3 dB to it, while dividing it by two means subtracting 3 dB from it. Similarly, multiplying by a factor of 10 means adding 10 dB to the signal power, and attenuating a signal by a factor of 10 is equivalent to subtracting 10 dB from its power. Thus, the left side of the table represents amplification of a signal by the given ratios, and the right side represents attenuation of the signal by the given ratios.

This table can also be used to convert back and forth between linear power and decibel-milliwatt. For example, a laser with 2 mW of optical power has an output power of 3 dBm. A receiver with a *sensitivity* of −30 dBm is capable of working properly with optical power levels as low as 0.001 mW. We should note that in this table decibel equivalences for numbers other than exponents of 10 are approximate. However, exact conversions can easily be done through Eqs. (1.17) and (1.18).

We are now in a position to discuss the topic of link budgets. To illustrate the topic, let us consider Fig. 1.7 which illustrates an example of an optical link along with the power levels at each point in the link. The transmitter's output power is assumed to be at −3 dBm. The output of the transmitter has to go through a coupler with a loss factor of 0.5 dB before it enters the fiber. This means the power level at the beginning of the fiber is −3.5 dBm. The fiber has a length of 20 km and is assumed to have a loss of 0.5 dB/km at the operating wavelength. This means for each kilometer of fiber, the power is reduced by 0.5 dB. That adds up to a total loss of 10 dB, which means if the power is −3.5 dBm at the input of the fiber, at the output it is reduced to −13.5 dBm. Then there is another coupler that couples the output of the link to the receiver. Thus, the power level at the receiver is −14 dBm.

Let us now assume that the receiver has a sensitivity of −21 dBm at the data rate and wavelength of interest. In this case, the receiver can work with much lower power levels than it is actually receiving. To be exact, we can say that this link has 7 dB of margin. Referring to Table 1.1 we find that 7 dB corresponds to a factor of 5, and −7 dB corresponds to a factor of 0.2. In other words, in this example the receiver can work with 5 times less power than what it is now receiving. This is indeed a fair amount of margin.

What does it mean to have extra margin in a link? In reality, there are always reasons for component degradation over time. The power of the transmitter may degrade over time because of aging of the laser, just as the sensitivity of the receiver may get worst as the receiver ages. Other causes that influence the behavior of the link include environmental factors such as temperature. If the numbers given in the example are not guaranteed over the entire temperature range in which the link is designed to work, then temperature degradations must also be taken into account. Another cause for additional loss is accumulation of dust particles or lack of perfect mating in optical couplers which could cause further degradation. All these factors must be considered when a link is being designed. As a result, extra margin should always be set aside for these various degradations.

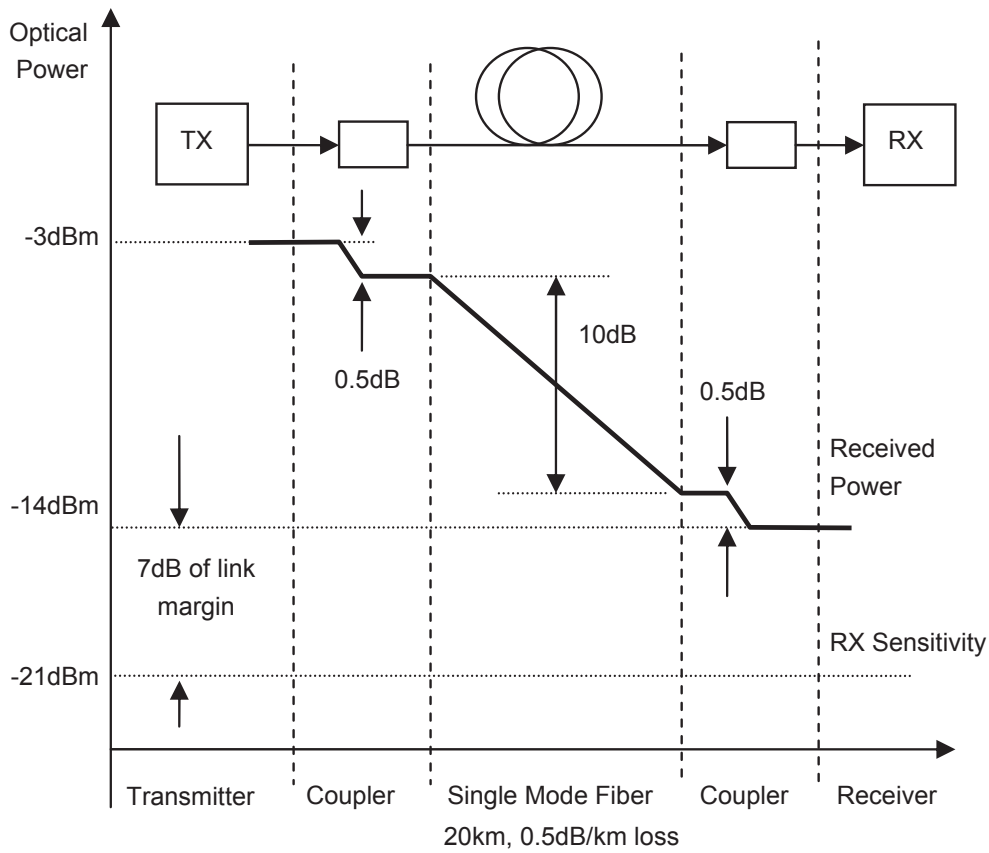

Fig. 1.7. Link budget in a simple optical link

On the other hand, having too much margin could indicate that the link is over-designed, i.e., it is likely that more expensive components have been used where cheaper parts could have worked just as well. Transmitters with higher power or receivers with better sensitivity are generally more expensive.

The choice of link margin is a tradeoff between many factors such as link specifications, cost, robustness, and link quality. Oftentimes standards define the link budget for a specific class of application. Moreover, reliability models based on statistical measures can provide a means of forecasting the performance of individual elements as they go through environmental stress and aging. We will discuss some of these topics as well as the relevant standards in the last chapters of this book.

1.9 Summary

This chapter provided an introduction to fiber optic communication. Through the example of language, we highlighted several basic concepts, including the fact

that communication typically involves some hierarchy, and that at the lowest level, it requires modulation of energy.

The form of energy used in fiber optics is light, and therefore it is important to gain some insight into the nature of light. From the classical perspective, the most complete description of light (and more generally electromagnetic waves) is given by Maxwell's equations. From the viewpoint of modern physics, light is characterized in terms of discrete packets of energy or photons. These two perspectives are not contradictory, but complementary. In fact, as we move from the long wavelength side of the electromagnetic spectrum to the low wavelength extremes, the energy of waves/photons increases, and particle-like behavior becomes more dominant. From a practical and engineering perspective, it is usually sufficient to describe the generation and absorption of light in terms of photons and the propagation of light in terms of classical fields described by Maxwell's equations.

With this basic view of light in mind, we next turned to the description of a generic fiber optic link. Here again we encountered the hierarchical nature of the flow of data, from higher data processing layers to a physical layer and vice versa. In the next chapter we will examine the hierarchical structure of communication networks in more detail. However, in this book we are mainly concerned with the physical layer, which consists of optical sources, optical detectors, and optical fibers. These elements were briefly introduced in this chapter and are examined more carefully in the other chapters of this book.

Fiber optics has become an integral part of the modern telecommunication infrastructure, and for good reasons. We discussed some of the main advantages of fiber optics, which for the most part, can be summarized in terms of higher bandwidths and longer distances. These two elements directly address the bottlenecks that modern telecommunication systems have to deal with, in both digital and analog domains.

We also introduced the concept of link budgeting, which takes the concept of a fiber optic link one step further by quantifying it in terms of link margin. Of particular importance here are the units of decibels and decibel-milliwatt for optical power, which are ubiquitous in the analysis and design of fiber optic links. Getting used to these logarithmic units will be helpful in gaining intuitive insight into a wide range of problems having to do with link budgeting, range, gain, and loss.

References

[1] Aristotle, *On The Soul*, Book II

[2] R. L. Oldershaw, "Democritus-scientific wizard of the 5th century BC," Speculations in Science and Technology, Vol. 21, pp. 37–44, 1988

[3] Plato, *Timaeus*

[4] K. Seyrafi, *Light, The Mystery of Universe*. Electro-Optical Research Company Los Angeles, CA, 1986

[5] R. Loudon, *The Quantum Theory of Light*, 3rd Ed., Oxford University Press, Oxford, 2000

[6] E. Hecht, *Optics*, 4th Ed., Addison-Wesley, Menlo Park, CA, 2002

[7] W. C. Chew, M. S. Tong, and B. Hu, *Recent Advances in Integral Equation Solvers for Electromagnetics*, Morgan and Claypool, San Rafael, CA, 2009

[8] J. Z. Buchwald, "Oliver Heaviside, Maxwell's apostle and Maxwellian apostate," Centaurus, Vol. 28, pp. 288–330, 1985

[9] G. Keiser, *Optical Fiber Communications*, McGraw-Hill, New York 1999

[10] V. Jacques, "Experimental realization of Wheeler's delayed-choice gedanken experiment," Science, Vol. 315, pp. 966–968, 2007

[11] V. V. Dodonov, "Nonclassical states in quantum optics: a 'squeezed' review of the first 75 years," Journal of Optics. B, Quantum and Semiclassical Optics, Vol. 4, pp. R1–R33, 2002

[12] M. M. deSouza, "Classical electrodynamics and the quantum nature of light," Journal of Physics A-Mathematical and General, Vol. 30, pp. 6565–6585, 1997

[13] D. F. Walls, "Evidence for the quantum nature of light," Nature, Vol. 280, pp. 451–454, 1979

[14] B Lounis and M. Orrit, "Single-photon sources," Reports on Progress in Physics, Vol. 68, pp. 1129–1179, 2005

[15] S Gasiorowicz, *Quantum Physics*, John Wiley & Sons, Hoboken; NJ, 1974

[16] T. L. Dimitrova and A. Weis, "The wave-particle duality of light: a demonstration experiment," American Journal of Physics, Vol. 76, pp. 137–142, 2008

[17] K. Camilleri, "Heisenberg and the wave-particle duality," Studies in History and Philosophy of Modern Physics, Vol. 37, pp. 298–315, 2006

[18] S. S. Afshar et al., "Paradox in wave-particle duality," Foundations of Physics, Vol. 37, pp. 295–305, 2007

[19] D. W. Ball, "The electromagnetic spectrum: a history," Spectroscopy, Vol. 22, pp. 14–20, 2007

[20] D. Supper et al. "Long-wavelength VCSEL with integrated relief for control of singlemode emission," Optics Communications, Vol. 267, pp. 447–450, 2006

[21] C.C. Xu and S. A. Boggs, "High frequency loss from neutral wire-shield interaction of shielded power cable," IEEE Transactions on Power Delivery, Vol. 23, pp. 531–536, 2008

[22] Y. D. Chung et al., "A 60-GHz-band analog optical system-on-package transmitter for fiber-radio communications," Journal of Lightwave Technology, Vol. 25, pp. 3407–3412, 2007

[23] K. Gnauck et al., "25.6-Tb/s WDM transmission of polarization-multiplexed RZ-DQPSK Signals," Journal of Lightwave Technology, Vol. 26, pp. 79–84, 2008

[24] H. Suzuki et al., "12.5 GHz spaced 1.28 Tb/s (512-channel×2.5 Gb/s) super-dense WDM transmission over 320 km SMF using multiwavelength generation technique," IEEE Photonics Technology Letters, Vol. 14, pp. 405–407, 2002

[25] J. X. Cai et al., "Long-haul 40 Gb/s DWDM transmission with aggregate capacities exceeding 1 Tb/s," Journal of Lightwave Technology, Vol. 20, pp. 2247–2258, 2002

[26] C. H. Cox, *Analog Optical Links*, Cambridge University Press, Cambridge, 2004

Chapter 2
Communication Networks

2.1 Introduction

In the previous section we reviewed the basic blocks that make up a fiber optic system. We divided a link vertically into two domains: the processing layers and the physical layer. The function of the physical layer is to convert an electrical signal into an optical signal, transmit the optical signal through the fiber, and convert it back to an electrical signal at the receiver. The performance of a physical layer system can be measured by the signal fidelity at the receiver end, i.e., by comparing the regenerated electrical signal at the receiver to the original electrical signal at the transmitter.

The processing layers, on the other hand, include the hardware and the software that handle tasks such as switching, routing, error detection, multiplexing, demultiplexing, as well as a wide range of other data processing tasks. Oftentimes these higher layers are generically called networking layers. In this book we are mainly concerned with fiber optics at the physical layer. However, before getting into the details of the blocks that make up the physical layer, we need to gain a better understanding about the networking layers.

The networking layers make up the backbone of the infrastructure that connects the users around the world together. These layers fill the space between the end users and the physical layer systems that handle the optical or electrical signals in cables and fibers. In a way, the networking layers use the services of the physical layer in order to move information contents around in communication networks.

Within the last few decades, networking has been a dynamic, fast-evolving area, going through continuous and at times fundamental changes in terms of technology, data rates, nature of traffic, scale, governing standards, and economy [1–2]. In this chapter we will go over some of the fundamental concepts in networking, with an eye on the role of fiber optics. In this way the contents of the data processing layers in Fig. 1.3 would become more meaningful, and we will get a better context for the topics of the remaining chapters.

2.2 Network topologies

In the first chapter we discussed the concept of a fiber optic link, which basically consists of an optical source, a fiber, and an optical receiver. We can think of an optical link as the basic building block for constructing more complex networks. One way to study networks from a more abstract perspective is by modeling them as a graph. A graph can be thought of as a set of connected points. The points, or *nodes*, are connected to each other via *links*. Thus, a graph is a set of nodes and

M. Azadeh, *Fiber Optics Engineering*, Optical Networks,
DOI 10.1007/978-1-4419-0304-4_2, © Springer Science+Business Media, LLC 2009

links. Moreover, by assigning a failure probability to the links in a graph, the performance and reliability of real networks can be modeled [3–6].

Let us consider the example of language again. When two individuals talk, we have a case of a simple network: each individual is a node, and the communication channel between them is the link (Fig. 2.1a). However, we can easily expand this example: we can think of three people sitting around a table, all talking the same language. Here we have a case of a network with three nodes and three links, because each individual can talk to the other two (Fig. 2.1b). On the other hand, consider the case where one individual, A, speaks only English, and the second, B, only French, while only the third, C, speaks both English and French. Here we have a case of a network with three nodes and two links. The individual who speaks both languages is in the middle, separately linked to the other two nodes (Fig. 2.1c). If C leaves the room, the link between A and B is broken, i.e., they cannot talk to each other because they can only communicate through this third node (Fig. 2.1d).

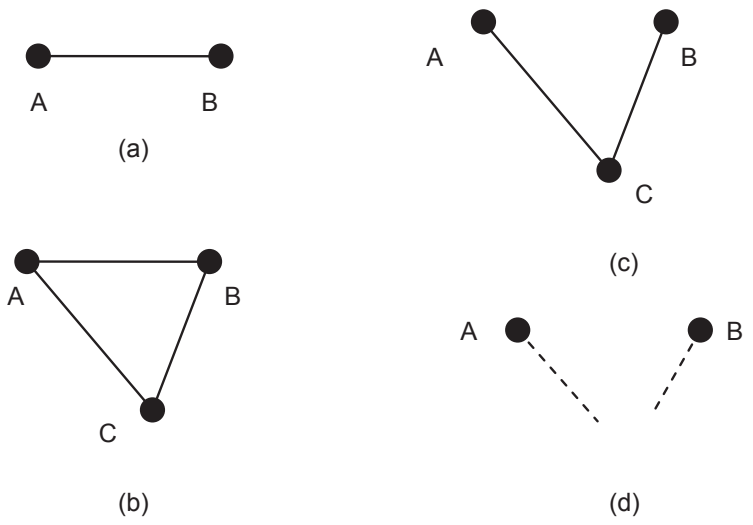

Fig. 2.1. Logical structure of a simple network

This simple example serves well to demonstrate the logical structure, or *topology,* of a network, as represented by a graph. The graph is a schematic representation of the flow of information between various nodes and how nodes are positioned with respect to each other. Each node can only talk directly to the nodes to which it is connected. The number of links is a measure of the interconnectedness of the network. Moreover, a more interconnected network is (naturally) more complex, but more resistant against disruption. For example, Fig. 2.1b represents a more robust network compared to Fig. 2.1c, because if any of the individuals in Fig. 2.1b leaves the room the other two can continue to talk, whereas in Fig. 2.1c, C is a *critical node*, because A and B are linked through C, and if C leaves the room, the communication between C and B is broken.

Communication networks work very much under the same principles. Each node is a point in the network that is capable of sending and receiving information, and the links are channels that they use to communicate to each other. The performance of a network is strongly affected by its topology [7]. Naturally, two nodes can communicate as long as they "talk the same language." Networking languages are defined in networking standards such as SONET and Ethernet. We will cover some of the main standards in fiber optics in Chapter 12.

We can now examine some of the common network topologies and their properties. Some of these common structures are shown in Fig. 2.2. One of the most straightforward ways to connect a certain number of nodes together is the linear or bus topology. Although connection details may vary, in general in a bus topology all the nodes have access to the same cable or channel.

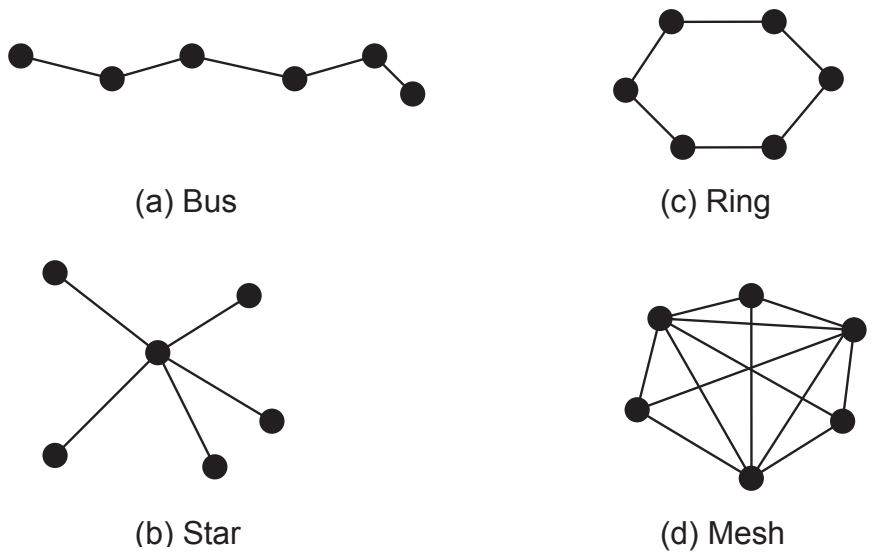

Fig. 2.2. Logical network topologies

If a node wants to talk, it broadcasts a message on the bus, but only the destination node responds to the message. The bus topology has the advantage of simplicity, and in terms of physical cabling requirement it needs the minimum amount of cables. It is also very scalable: adding one node requires adding one link. However, there are several shortcomings. For one thing, the *diameter* of the network is large. We can think of the diameter as the maximum length a signal must travel between any two nodes. If the two nodes at the edges of the bus want to talk, the signals must travel the entire network back and forth. This also exposes another weakness in the bus architecture: a single failure anywhere in the link brings down the whole network. Thus, bus topologies are less reliable compared to other topologies.

The star topology shown in Fig. 2.2b resolves some of these limitations. It consists of a central node, called a *hub*. All the other nodes have to talk through the hub. Thus, the hub's load is disproportionably heavier than the rest of the nodes,

because it must handle the traffic from all nodes. Compared to the bus architecture, the star topology is more complex. In practice, it also requires more cabling. Consider the case where two of the nodes are close to each other but physically far from the hub. We need to have two separate links running side by side to the hub. However, the star topology has several advantages. In a star network, if any of the links fails, only the node connected to that link will fail, and the rest of the network can function normally (the exception is the central node or the hub, which if failed, brings down the entire network). The diameter of the network is also much smaller. It takes a maximum of two links for any two nodes to talk to each other. The star topology is also easily scalable. Adding one node requires adding one link from that node to the hub.

Another common topology is the ring topology shown in Fig. 2.2c. Here the load is more evenly distributed, resulting in a more organized network. Bi-directional rings also are advantageous in terms of fault tolerance. A single failure in the network still leaves the other path for the nodes to communicate. Effectively, a single failure reduces a ring to a bus-type topology. The diameter of the network is somewhere between those of the bus and star structures. If the two nodes happen to be at the opposite ends of the ring, the number of links required for them to talk to each other is approximately half the number of nodes. A disadvantage of ring topology is that any change in the network may cause disruption in the entire network. Also, a ring topology is not easily scalable. If a new node is to be added, it has to go in between two other nodes, and generally two additional cables are needed.

The mesh topology, shown in Fig. 2.2d is the most interconnected network topology. In a fully connected mesh, all nodes are connected to all other nodes. The obvious advantage is that the network becomes very *fault tolerant*. Each node is connected to the rest of the network via multiple links; and therefore single or even multiple failures tend to remain contained. A fully connected mesh also provides the minimum diameter possible: each node is only one link away from any other node. But these advantages come at a price. A mesh network requires the maximum amount of physical cabling. Moreover, unlike the other topologies, the signals have more than one path to go from one node to another. This means that switching or routing decisions are needed. Obviously, then, a mesh network can become very complex if the number of nets gets large.

In practice, the choice of network topology depends on the specific situation and is dictated by many parameters. For example, for small networks within a house or a small office, a bus or star network may be an optimal solution. However, for larger networks, it becomes increasingly difficult to justify a single topology, and therefore a mixture of topologies must be used. Still larger networks may not even belong to a single company and could consist of many patches developed separately for various applications and purposes. This is especially true with the telecommunication infrastructure, which expands whole cities or continents. For example, individual users may be connected to a local central hub in a star configuration, the local hubs connected to each other through a ring topology

that circles a city, and the ring topologies in several cities connected to each other in the form of a mesh, either partially or fully connected.

In fiber optic networks, the nodes in a particular topology consist of optical transmitters and receivers. We shall discuss these from a hardware perspective in more detail in Chapters 4 and 6. The links themselves are made of optical fibers that are connected to the nodes (and to each other) through optical couplers and other passive components. These components are the subjects of Chapters 5 and 7.

2.3 Telecommunication networks

Fiber optics is an integral part of the telecommunication infrastructure that has come to be such a dominant feature of modern life. Thus, in order to gain a better understanding of fiber optic networks, it is necessary to gain a general perspective on the telecommunication infrastructure [8–10]. This perspective clarifies the position of fiber optics within the broader context of telecommunication industry and provides insight into the forces behind the evolution of the fiber optic industry.

The oldest and perhaps the most familiar telecommunication network is the telephone system. Currently, most major points on earth are interconnected via the telephone networks. Each user has direct access to this network via a dedicated line, which is typically a real pair of copper wires that connect the user to the rest of the system. Voice signals do not require much of a bandwidth, typically only a few kilohertzs of bandwidth is sufficient for a telephone conversation. This bandwidth is provided to each user separately. Telephone networks are usually *circuit switching* networks, which means the system, at least on a conceptual level, provides a direct path between the two users at the two sides of a phone conversation. This path is provided by the switching systems that use the information in the phone number to connect the source with the intended destination, and it is established as long as the phone connection is continuing, regardless of the actual signal contents. In other words, a channel with the required bandwidth is established between the two sides, regardless of the two sides using it or leaving it idle. The technical details are hidden from the users: they can be in two rooms next to each other or at opposite sides of the world. As far as they are concerned, it is as if a pair of wires is connecting the two sides together.

In reality, the signals from individual users go to local telephone centers, where they are converted to digital signals and aggregated through time division multiplexing with signals from other users. Thus, a hierarchy is created: lower data rates are combined together to form higher rates. In case of long-distance calls, these higher data rate signals are typically carried on optical fibers to their destination, where they are de-multiplexed into separate signals, and ultimately switched to their individual destination.

The other major telecommunication system is the Internet. Compared to the telephone network, the roots of Internet are much more recent. What is currently

called Internet originated in military applications in the 1960s and 1970s, with various academic institutions joining later. The idea was to connect various computers together through a network based on *packet switching*, as opposed to circuit switching, which was the basis of the telephone system. By mid 1990s the exponential growth of Internet had already started, and the rest is history.

Unlike the telephone system which is designed for the transfer of real-time audio signals, the Internet is designed with the purpose of transfer of digital data. Individual users are connected to the Internet through a variety of ways. These typically include traditional phone lines, digital subscriber line (DSL), and services provided by cable companies. Larger users such as companies or universities have direct access to the Internet. The principle of signal aggregation is valid for Internet too. For long-distance transmission, several low data rate signals are combined into fewer high data rate signals through time division multiplexing. Once data is aggregated into these high data rate streams, it is often mixed with other data formats and carried on the same backbone infrastructure that carries high-speed switching-based signals. At the destination, the signals are unpacked and then routed toward their individual destinations.

Another major component of the telecommunication infrastructure is the radio and TV broadcasting systems. Unlike the telephone and Internet networks, the radio and TV systems were primarily designed for one-way transmission: the same data stream flows from a central TV radio station to all the users within the reach of the signal. Cable TV companies are more or less based on the same model: they transmit the same signal to all the users that are connected to their services, although they may scramble certain prime channels and only provide the descrambler equipment to those subscribers who are paying for those prime channels.

In general, TV and radio broadcasting have their roots in analog transmission formats, based on frequency multiplexing of different channels with different carriers. This means broadcast signals generally cannot be baseband, otherwise all the sources will have to use the same frequency band for their signals, which causes significant interference and renders all the signals useless. Frequency multiplexing thus allows each user to tune in to a specific signal. A clear example is radio: all the stations broadcast their signals on individual carriers at various frequencies that span from approximately 500 to 1600 kHz for AM bands and from 88 to 108 MHz for FM bands. Thus, frequency multiplexing provides a solution for broadcast communications. However, as far as the analog nature of broadcasting is concerned, things are changing. This is especially evident in TV broadcasting. Many TV stations and cable TV companies are already broadcasting in high-definition digital formats (HDTV), and the other stations are following.

Telephone, data, and broadcasting applications are, of course, by no means separate or isolated. Many individual subscribers are connecting to Internet via telephone or cable TV lines, just as many broadcast stations receive or transmit their contents between themselves and other major content sources through backbone Internet structures. In general, the farther we get from the individual users, the more convergence we encounter in terms of data formats, speeds, and standards. This is a requirement based as much on technological reasons as on economic

grounds. If a specialized content is to be transmitted across the globe, and if it is going to use the existing long-haul telecommunication structures such as the submarine fiber optic cables, it has to conform to the standards and formats of digital optical signals. As a result, sooner or later it has to be transformed into a digital format and multiplexed within the rest of the traffic that the fiber optic link will ultimately carry. Fortunately, the hierarchical structure of the networks hides these complications from users. The content is converted into the right format by the intermediate processing layers and converted back to its original form on the other side. The end users do not have to worry about these conversions. On the other hand, as far as the fiber optic link (physical layer) is concerned, everything is 1s and 0s: it receives a digital stream at the source side from the higher processing layers, and it must deliver the same signal at the other side of the link to the higher processing layers.

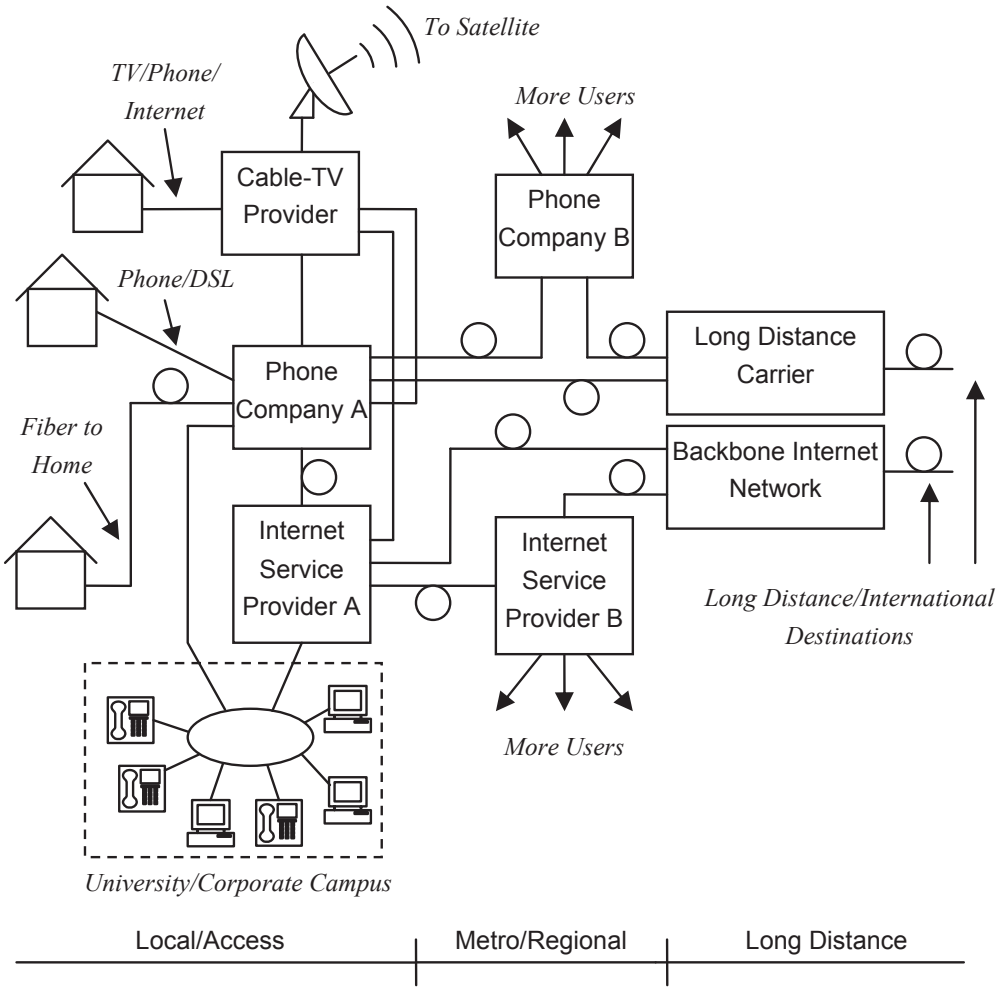

Fig. 2.3. A simple illustration of telecommunication infrastructure

Figure 2.3 is a simple illustration of the telecommunication infrastructure. It should be noted that what is depicted in this figure is merely an illustration since each particular network differs from others in many details. Thus, this figure is

like a representative map that only serves to clarify and illustrate the concepts. In general, this figure is arranged in terms of physical distance from left to right. Local or *access* networks that are limited to smaller geographical areas are in the left side. These could be the networks within a corporate or university campus or the infrastructure that connects individual homes to the main networks. The reach of access networks is typically up to a few kilometers. Regional or *metro* networks are the next level. These are networks that connect local telephone companies or Internet providers within a larger geographical area such as a city. The reach of these networks can be up to a few hundred kilometers. The next level comprises the long-distance carriers. These carriers transfer the aggregated data from regional phone companies and Internet providers and transfer them across long distances, for instance across a country or from one continent to another. An example of long-distance networks is the submarine fiber optic cables that transfer the traffic across the Atlantic or Pacific oceans.

We can see that individual homes are connected to the network via regular phone lines, DSL lines, or cable connections to the cable companies. It used to be the case that various companies offered distinct services, for example cable companies offered TV programming, while phone companies offered standard phone services. But we see an increasing merger of these services. With the expansion of Internet, phone lines were also used to carry data signals, although in slow rates, while digital subscriber line (DSL) modems offered faster data rates. However, increasingly, cable companies offer both phone and internet services over their copper cables. A new development, as far as fiber optics is concerned, is the emergence of fiber to the home (FTTH) technology, which satisfies all user data requirements with a single strand of fiber while providing very high data rates by taking advantage of the high bandwidth of the optical fibers.

There are also campus-wide networks (usually called LANs or local area networks) that connect the computers within a university or company campus. These networks then usually have a direct access to the main data networks through Internet service providers. Typically, they also have an internal phone network, and thus they require dedicated connections to the local phone companies as well.

The next level up in the network is the regional or metro networks, which typically include local phone companies and various Internet providers. If a call originates from a subscriber, it typically goes to the local switching board where it is handled by a local phone service company. If the destination is another local subscriber within the same area, the signal stays within the same area. If the destination is a subscriber in the same region, the call may be switched from one regional office to another, or even it may be handled by more than one phone company. If the destination is farther away, for example in an international call, the signal is passed on to a long-distance carrier, where it is transferred on specialized high bandwidth links to other geographical areas.

Long-distance carriers typically function on a different business model. They do not interact with individual users directly. Instead, they are specialized in receiving relatively high-speed signals from regional phone companies and Internet providers and multiplexing them in yet higher data rates, in formats suitable for

long-distance transmission usually in fiber optic links. Before the growth of fiber optics, these long-distance communications had to be handled by microwave links or specialized copper-based systems. However, with the advantages offered by fiber optics, long-distance communication is now almost exclusively the realm of fiber optics.

Figure 2.3 also illustrates the role of fiber optics in the telecommunication infrastructure. In this figure the links that are likely to be fiber optic based are marked with a circle on the connections between the nodes, while copper connections are just simple lines. It can be seen that in general, the farther the distances, and the higher the data rates, the more likely it is for a link to be fiber based. This is of course a direct consequence of the characteristics of fiber optics that makes it the main technology of choice for long-distance communications. For example, submarine fiber optic cables transfer data streams at speeds of many gigabits per second under the oceans for thousands of kilometers. The only practical alternative to long rang fiber-based links are satellite links. For example, as shown in Fig. 2.3, a cable TV company may receive some of its contents through a satellite link from a main feed. Satellite links are also useful for areas where the fiber-based infrastructure does not exist.

However, as we get closer to the individual users, where the distances are shorter and the data rates are lower, more competing technologies exist. Depending on the special circumstances and the existence of legacy infrastructure, local phone companies or switching offices may be interconnected to each other through either copper cables or fiber links. Likewise, depending on their sizes and complexities, local area networks can be copper or fiber based, although in many cases new installations tend to be fiber.

Individual homes, on the other hand, are still connected mostly through copper wires and cables. On one side, this is a result of legacy: copper cables and telephone wires have a long history, and telephone and cable companies are reluctant to incur cost and modify existing (and working) systems. On the other hand, expanding fiber toward individual subscribers must result in an economically viable system, and for that to happen, the cost of the fiber optic box that goes in each subscriber's house along with the cables that connect each house to a central office should be low enough to make the resulting service a consumer product. In spite of these challenges, the fiber is finding its way to individual homes. This is an area at its beginning phases, but is expanding fast. In practice, this means in many cases the fiber does not stop at a local cable or phone company office, but comes closer to individual users, perhaps to an apartment complex, or in many cases even to individual homes. The fiber to the premise (FTTP) or fiber to the curb (FTTC) or fiber to the home (FTTH) infrastructures are all variations on the same idea and are sometimes categorized together under the more generic name of FTTX. We will examine this subject again later in this chapter.

2.4 Networking spans

Topology is just one attribute that can be used to describe and categorize networks. Another useful way to categorize networks is based on their span or range. From this perspective, networks are usually divided into local area networks (LAN), metropolitan area networks (MAN), and wide area networks (WAN). Although these categories are already evident in Fig. 2.3, we should discuss them in some more detail for future reference.

2.4.1 Local area networks (LANs)

A LAN is typically a small network that connects computers and other resources in a small geographical distribution. Examples include university campuses, hospitals, and corporation or factory buildings. As a result, usually LANs are maintained by and belong to the same organization. The maximum distance a LAN covers is determined by the mode of communication. Fiber optic links can extend to tens of kilometers, whereas copper cables are more limited. The reach of a WiFi LAN is limited by the power of the transceivers, and it could extend to tens of meters.

The important feature of a LAN is that it allows the connected resources to share information with each other. For example, in a company, all computers, several printers, and several servers may be connected together through a LAN. A very popular protocol for LANs is the Ethernet, specified in various flavors of the IEEE 802.3 standard, which in its original form defined a 10 Mbps protocol on a coaxial bus topology [11–12].

However, the bus topology is not very successful for fiber optic links. The reason is that without using optical to electrical converters, tapping a link off an optical bus requires a coupler, and using couplers causes loss of power, which in turn limits the number of nodes that can be connected to a LAN.

In a LAN, each node can access the network at any time, and this may cause access collisions. Thus, it is the responsibility of protocols to resolve any potential conflicts arising from the use of the same resources by many users. Usually LANs are not isolated, rather, they are connected to larger and more extended networks.

2.4.2 Metropolitan area networks (MANs)

The next level of span in a network is when the network range extends beyond that of a LAN, perhaps to several blocks, multiple neighborhoods or districts, or an entire city. Such a network is usually known as a metropolitan area network or a MAN.

Unlike LANs, MANs are usually not managed or used only by a single individual or organization. For instance, a group of service providers may share the ownership of a MAN in a city, although it is possible for a company to lease

bandwidth on a MAN from another company. Generally speaking, individual users do not have access to a MAN directly. Instead, MANs are used to interconnect other networks such as individual LANs. Therefore, individual users can connect to a MAN either through a LAN or through an access network, such as community access TV (CATV) network or a DSL line.[1]

Because MANs typically cover longer distances and higher data rates compared to LANs, they are more frequently based on fiber optic links. A MAN can have a physical reach of up to 100 km and may carry data rates as high as 40 Gbps. Moreover, because of the pressure toward larger bandwidths and because the costs can be shared between a larger number of end users, MANs afford to use more sophisticated technologies and devices. Routers and switches, and bandwidth-enhancing techniques such as WDM, can be used in a MAN. A typical protocol for MANs is asynchronous transfer mode (ATM), although migration toward Ethernet standards such as 10 Gbps Ethernet is taking place as well [13].

2.4.3 Wide area networks (WANs)

Networks that extend to many cities, countries, or even continents are designated as wide area networks or WANs. Because WANs have the longest physical range among the networks, they are usually not owned by a single entity. Instead various entities may own and operate various segments of WANs. Like the case of metropolitan area networks, individual users do not have direct access to WANs. Instead, WANs are used to connect smaller MANs and LANs.

WANs handle large amounts of data at high data rates and over large distances. As a result, fiber optics plays a dominant role in WANs. Moreover, advanced multiplexing techniques, especially WDM techniques, are more likely to be used in WANs, because they increase the effective bandwidth of existing fiber. A clear example is the submarine cables that carry the traffic between the continents. It is much more expensive to deploy a submarine cable than it is to add sophisticated equipment at the two sides of an existing link to improve the bandwidth efficiency.

Because of the large reach of WANs and the fact that they typically connect many other types of networks with a vast array of devices and technologies, WANs tend to use a mixture of different topologies. WANs, like other kinds of networks, have to follow well-defined specific protocols. Common protocols include ATM and SONET (synchronous optical network).

[1] That is why sometimes access networks are also considered a form of LAN.

2.5 Hierarchical structure of networks

So far we have described the networks from a "horizontal" perspective. Describing a network from a topological viewpoint, for instance, is essentially a description of how various nodes are connected to each other and communicate with each other. Likewise, categorizing networks as LANs, MANs, or WANs is a horizontal categorization, because it divides the networks based on their geographical reach. However, a complete description of networks requires some level of hierarchical or vertical description as well. A vertical description refers to the fact that information must go through several layers of processing before it can be converted to a physical signal and transferred in a physical medium. In the example of human language in Chapter 1, this hierarchy refers to the processes that take place in the brain in the form of structuring an idea in terms of language, organizing and refining the language into a set of letters, words, and sentences, changing these linguistic units into neural signals, and finally modulating the air through the vocal system driven by those signals. Once the resulting sound waves travel through the medium, the opposite processes must take place to convert the physical signal back into ideas.

While this vertical or hieratical process can be immensely complex and "fuzzy" in case of human conversation, for communication networks it must be well defined and logically clear. All the required processes must be well thought of and be divided into manageable, clear steps. These steps or *layers* must be defined in a way that each has a clear and separate function, so that the whole task of communication between two separate points can be divided into small, manageable, logically separate chunks. Moreover, each layer should ideally interact only with the layers above and below it.

Such a division of labor is clearly advantageous for many reasons. It simplifies the task of each layer and makes improvements easier, because the changes in each layer do not have to affect other layers. Thus, each layer can specialize and excel in a particular task, resulting in networks with improved design, better efficiency, and more reliability.

2.5.1 Open System Interconnect (OSI) model

One of the most common models for this hierarchical approach is the Open System Interconnect (OSI) model developed in 1984 by International Organization for Standardization (ISO) [14–16].[2] The OSI model itself is not a protocol or standard, and it is not even followed in all practical systems. However, it defines a conceptual framework that can be used as a reference for other hieratical structures and, as such, is a widely used and referenced model. Figure 2.4 shows an overview of the OSI model.

[2] For more information on ISO and its activities see Chapter 12.

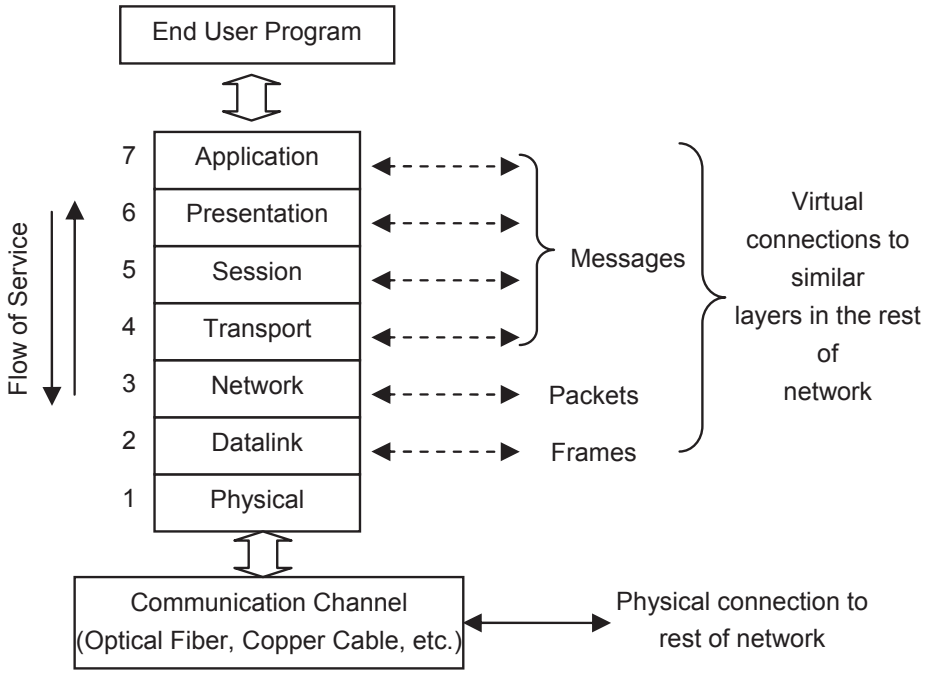

Fig. 2.4. The OSI hierarchical network model

The higher layers in this model are closer to the end users, while the lower layers deal with more hardware-oriented tasks such as addressing, switching, and multiplexing functions. The physical layer, which is the lowest layer, deals with physical signals such as electrical voltages and optical signals. It receives the information from the datalink layer and sends it over the physical channel, which can be fiber, cable, free space, etc. The physical layer on the other side of the link receives the signal and sends it up to the datalink layer. The highest layer, or the application layer, is the closest layer to the end users. Although in the OSI model it does not interact directly with the end users, both the end user and the application layer interact with various programs that have some form of communication function, for example web browsers, and email programs.

Figure 2.4 also illustrates another feature of the OSI model. In general, each layer communicates both vertically and horizontally. Vertically, it communicates with the layers below or above itself. A higher layer requests some kind of service from the layer immediately below itself, and the lower layer provides that service to the higher layer. The interface between the two layers is called a service access point (SAP). Therefore, the relationship between the layers is a *client–server* relationship.

Horizontally, each layer interacts with its *peer* layer on the other side of the link. For example, the network layer is receiving data from the transport layer and passing it down to the datalink layer. At the same time, it is interacting with its peer network layer on the other side of the link, although it does not care about the

details of how that communication has actually been achieved through the lower layers. That is why in Fig. 2.4 all the horizontal connections are designated as virtual: only the physical layer provides physical connection between the nodes.

As noted before, the major focus of this book is fiber optics communication from the physical layer perspective. The physical layer interacts directly with the physical medium. In case of fiber optics, this includes the modulation of light in the form of optical signals and sending and receiving it through the optical fiber. We will cover these topics in detail throughout this book. However, we should have some insight into the higher layers, especially those that are closer to the physical layer. Therefore, we will have a brief review of these higher layers, with an emphasis on those closer to the physical layer.

2.5.2 Datalink layer

The datalink layer (layer 2) is the closet layer to the physical layer. In the OSI model, the datalink layer is responsible for reliable transfer of data between adjacent nodes. In other words, the datalink layer does not concern itself with sending data to its ultimate destination; it is just concerned with sending data out to the next available node(s).

The unit of data in the network layer is usually called a frame. A frame at the datalink layer consists of data from layers above encapsulated in headers and trailers added by the datalink layer. These headers and trailers are intended for use by the datalink layer in the destination and are transparent to higher layers on both sides. Functions of the datalink layer generally include such tasks as encapsulating data into frames, ensuring the frames are transmitted and received without errors, and controlling the flow of frames.

Following the IEEE 802 standard, sometimes the datalink layer itself is divided into two sublayers: the logical link control (LLC) and media access control (MAC) sublayers, where the LLC layer is located above the MAC layer. In 802.2, the LLC is responsible for communication between devices located on a single link. This is done by dedication of various overhead fields in the frame such that higher layers can use and share the same physical link through these fields. The MAC layer is responsible for organizing access to the physical layer. For example, in many cases the same medium is shared between several nodes, such as in a bus topology where all the nodes use the same link. Therefore, a set of rules must be established to avoid conflict. IEEE 802 also defines physical MAC addresses that uniquely identify each device on the network (not to be confused with the IP address, which is a higher layer concept). Note that the datalink layer is the lowest level in which data frames and various fields within a frame are meaningful, because the lower physical layer does not distinguish between these various fields and simply treats data as a serial digital stream.

Some of the most popular protocols that define the datalink layer include IEEE 802.3 Ethernet (carrier-sense multiple access with collision detection LAN) and IEEE 802.5 (token ring LAN).

2.5.3 Network layer

The network layer (layer 3) is the next layer up in the OSI model. The main task of network layer is the transfer of information from source to the final destination, which in general involves routing functions. The unit of information in the network layer is sometimes called a packet. The network layer makes packets by receiving data from higher layers and encapsulating them with its own headers and trailers. These headers and trailers are intended for network layer operation and are transparent to layers below and above. Thus, the datalink layer below treats a packet as data and adds its own headers and trailers, as explained before.

2.5.4 Higher layers

The first three layers in the OSI model (physical, datalink, and network layers) are sometimes grouped together as interface layers. These layers are directly involved in the telecommunication functions. On the other hand, the higher layers (transport, session, presentation, and application) are involved with end-to-end (also called peer-to-peer) tasks. The functions of these layers include establishment and termination of sessions, data compression and decompression, interactions with operating systems, translating data into standard formats, user interface services, etc. Moreover, these higher layers are very "fluid" and are highly dependent on evolving technologies and protocols [17]. We do not go into the details of these layers, because for the most part, these layers are not directly involved in the telecommunication aspects of networking.

As noted before, the OSI model is a useful tool for understanding network concepts. In reality, not all networks conform to this model. In practical networks some of these layers may be omitted, expanded, or mixed. It is also possible that some technologies define their functions in a way that does not clearly correspond to the standard OSI layers. For instance, SONET has provisions to deal with end to end connections defined in the network layer. However, SONET has other functions too that do not correspond to the network layer. Despite these conformance details, the OSI model provides a useful conceptual tool for mapping other protocols, standards, and technologies.

2.6 Circuit switching and packet switching networks

Communication services generally fall in two categories: connection oriented and connectionless. These types of services are closely related to the ways networks (and related protocols that operate over the networks) function.

Generally, there are three different approaches to digital communication. In a *broadcast* network, all the messages that are sent from any of the nodes are received by all other nodes. For example, consider an Ethernet LAN, where all

computers are connected to a single bus. When one computer sends a message, all other computers on the network receive that message. It is only the content of certain fields within the message that marks the destination. Obviously things can get more complicated, for instance, when two computers send their messages simultaneously and a collision happens. The network protocol must therefore have provisions to resolve the contention. A broadcast network is thus like a gathering at the dinner table, where everybody can hear what everybody else says. It can be seen that a broadcast approach is not suitable for large networks with multiple nodes. The alternative to a broadcast strategy is to implement some form of switching or routing, so that data from one node can be directed only to its destination node. This gives rise to two alternative approaches: circuit switching and packet switching [18–19].

2.6.1 Circuit switching

The circuit switching approach has its origins in the telephone network. In a phone conversation, the two sides are connected together through a real circuit. In other words, there must be a channel that connects the two sides for the duration of conversation. As long as the two sides keep the connection (even if they do not talk) they have the channel only to themselves. This is because voice is a *time-sensitive* application. If I talk to my friend, I like my friend to hear my words in the same order that I speak them. Indeed, all real-time audio and video applications are time sensitive. The receiver must receive the signals in the same order and at the same rate that the transmitter is sending them. This could automatically be obtained in a single point-to-point link: after all, the signal sent through a fiber link reaches the other end of the link at the same rate and in the same order.

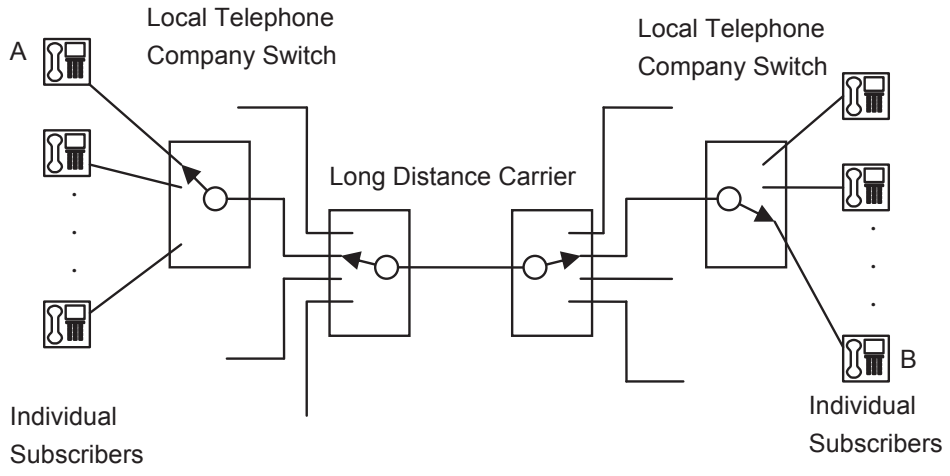

Fig. 2.5. Schematic of a circuit switching network

However, in a network where signals must be multiplexed and routed, the issues can be more complex and require special attention. This is why telephone networks are based on the concept of circuit switching. As shown in Fig. 2.5, when two individuals talk over the phone, various switches in the network provide a direct path between them, so that during their conversation they are connected in real time and have the entire bandwidth of a standard telephone line to themselves.

Of course in reality the signal will be multiplexed into successively higher rates and then demultiplexed back as it approaches its destination. Nevertheless, these details are transparent to the two subscribers at the two sides. A prominent example of circuit switching architectures is SONET. We will talk more about SONET later in this chapter.

2.6.2 Packet switching

Unlike voice and video applications, data applications are not so sensitive to time delays. Moreover, data traffic tends to come in bursts. At one point in time the application may produce a large amount of data that needs to be transmitted, while in the next moment it may not need to transmit any data. Obviously dedicating an entire channel to this kind of traffic is not efficient, because the channel may have to be idle most of the time. The solution is to break down the data into packages and send each packet independently over the network. Here a technique called *statistical multiplexing* can be used: packets from various nodes will be transmitted through the same channel as they arrive. If the number of packets exceeds the available bandwidth of the channel, they must be buffered or stacked and sent later once the bandwidth is freed up. In this way, digital content can be broken down into several pieces and transmitted separately, even over different paths. Once all the pieces arrive at the destination, the original content can be reconstructed.

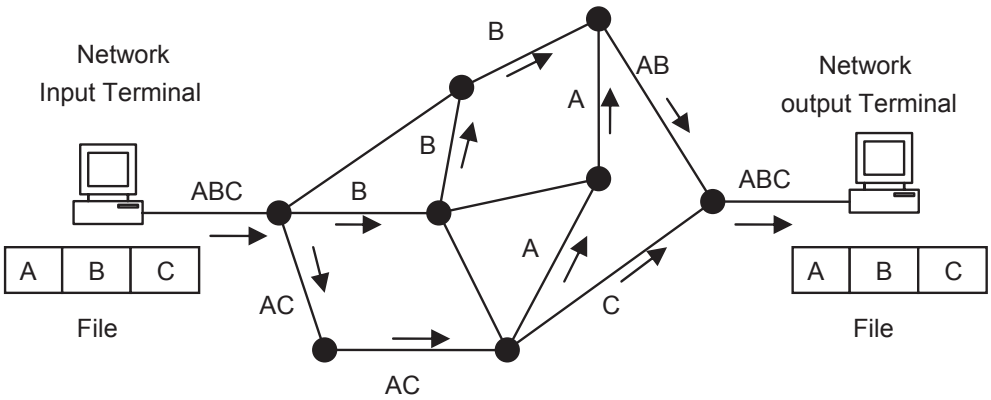

Fig. 2.6. Packet switching network

This process is schematically shown in Fig. 2.6. Here some digital content needs to be sent across the network. Instead of having a direct dedicated channel, the file is divided into several pieces or packets. Next, the packet is embedded with additional data called *headers*. The header includes information such as the address of the destination to which the file must be sent. The packets are then sent across the network. Various routers and switches in the way decide on the path that each packet will take by looking at the destination address in the header, as well as by considering the load on each available link. If no link is available at the time, the packet may be buffered and sent in a later time where a link becomes available. As a result, depending on the particular protocol in use, not all pieces may travel along the same path, and they may not even arrive at the destination in the same order that they were sent (more on this later). Once all the packets arrive, the original content will be reassembled based on the additional information in each packet.

As noted above, the advantage of packet switching is that available bandwidth is used much more efficiently. However, the disadvantage is that time-sensitive applications may suffer, because unless explicit provisions are made, no real-time relationship between the source and the destination exists. Examples of packet switching technology include IP, ATM, and MPLS.

Naturally, one may ask if there is any way to accommodate time-sensitive applications with packet switching. Conceivably, a direct connection may be "emulated" by a packet switching network through provisions that ensure packets arrive at their destination in the same order and time frame that they were sent. As a result of such considerations, packet switching networks can be divided into two further categories: connectionless and connection oriented.

In a *connection-oriented* network, a virtual path between the source and the destination must be established first. In the next phase, all the packets are sent through this same path. Finally, once all the packets are transferred, the connection will be terminated. A connection-oriented network is different from a circuit switching network in that several virtual paths may share the same bandwidth, i.e., the channel is not solely dedicated to any single virtual path. Primary examples of connection-oriented networks include ATM and MPLS.

In a *connectionless* packet switching network, on the other hand, no virtual path between the source and the destination is established and each packet is treated individually. The fact that no prior connection has been established means that the information may be sent without verifying if the receiver has a valid address or is actually capable of receiving the information. A primary example of connectionless networks is the Internet protocol (IP).

2.7 SONET/SDH

As noted above, a prominent example of circuit switching is SONET, which is an acronym for synchronous optical network. Closely associated with SONET is its

international counterpart, the synchronous digital hierarchy or SDH. SONET and SDH are well-established robust technologies in wide use, especially in backbone networks. Since their introduction in the late 1980s and early 1990s, SONET/SDH have had a fundamental role in optical networks by defining a clear set of optical signals, hierarchical multiplexing schemes, bit rates, and signal management rules [20–23].

The origins of SONET and SDH go back to the telephone network and bandwidth requirements for voice. Although humans can be sensitive to audio frequencies as high as 20 kHz, for a normal conversation only a fraction of this bandwidth is sufficient. Therefore, a standard telephone channel assigns 4 kHz of bandwidth for a normal conversation. The audio signal is then sampled at twice this rate or 8000 times per second. This assures that later on it can be faithfully reproduced when it is converted back to analog format. Moreover, each audio sample is digitized with a resolution of a byte, which represents 128 possible values. This brings the number of bits per second for a digitized signal to 8000×8=64,000. Thus, we can think of the rate of 64 Kbps as a "unit" of bandwidth.

To utilize the available bandwidth of communication channels more effectively, several voice channels can be multiplexed. The DS standard, adopted in North America, defines such a multiplexing scheme for higher data rates. The next level of multiplexing defined by the DS standard is when 24 voice channels are combined, which results in a bandwidth of 24×64 Kbps=1.544 Mbps. Such a signal is known as a digital signal 1 (DS-1), or T1. Higher data rates can be achieved by multiplexing more voice channels. Of particular interest to our discussion is the DS3 (also known as T3) signal which consists of 672 voice channels and corresponds to a data rate of 44.736 Mbps.

Outside North America, the same hierarchy was defined slightly differently through the ITU-T standards. ITU-T defines the first level of multiplexing with 30 voice channels. The result is called an E1 signal and corresponds to a data rate of 30×64 Kbps=2.048 Mbps. Multiplexing more voice channels results in successively higher data rate signals, for instance, the E3 signal consists of 480 voice channels and corresponds to a data rate of 34.368 Mbps.

The SONET standard originally proposed by Telcordia (formerly known as Bellcore) developed out of the need to multiplex several DS signals such as the T1 signal and transmit them optically. Because SONET was not exactly conforming to ITU's digital hierarchy signal structure, ITU defined a similar set of standards to SONET called synchronous digital hierarchy (SDH), which was suitable for multiplexing ITU's digital signals. We should note, however, that the SDH standard was defined in a way to accommodate SONET signals too, and therefore all SONET standards comply with SDH.

Both SONET and SDH handle data in the form of frames. Figure 2.7 shows the basic structure of a SONET frame. The SONET frame can be represented as a two-dimensional array of bytes with 90 columns and 9 rows.

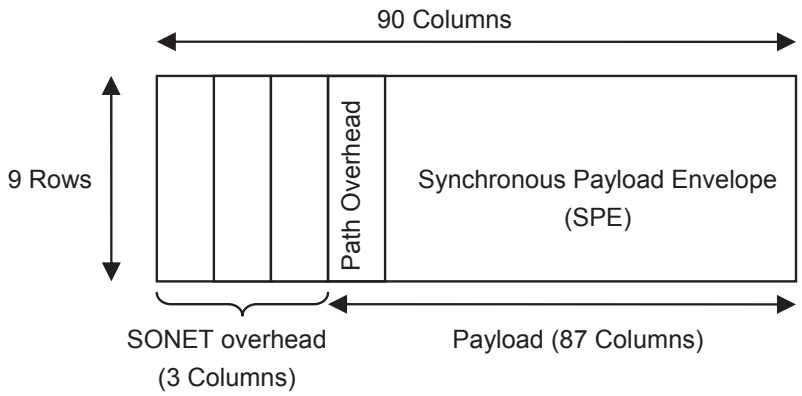

Fig. 2.7. Structure of the basic SONET frame

Data is transmitted serially starting from left to right and from top to bottom, with the most significant bit in each byte transmitted first. The first three columns are set aside as transport overhead and are used for SONET's internal functions. The remaining 87 columns are called synchronous payload envelope. Also, another column is set aside as path overhead, which leaves 86 columns for users' data.

Each SONET frame is defined to be 125 μs long, corresponding to 8000 frames per second. Since each frame has $9\times90=810$ bytes, and each byte is 8 bits, the base bit rate in a SONET signal would be: $8000\times810\times8=51.84$ Mbps. This base rate is defined as a STS-1 signal. Note that not all this bandwidth is user data. From Fig. 2.7 it can be seen that the payload is 86 columns out of the total 90 columns, or 774 bytes out of the 810 bytes. This means a SONET signal can support user payloads of up to $8000\times774\times8=49.536$ Mbps. Looking back at the DS standard, we see that this bandwidth is sufficient for a DS-3 signal.

The STS-1 signal described above is a *logical* signal. By itself, this signal may not be very easy to transmit over the fiber, because it may include long sequences of identical bits, a situation that can cause problems for an optical transmitter. To avoid these problems, and also to make clock recovery easier on the receiver side, the STS-1 signal is scrambled before being converted to an optical signal. The resulting optical signal, after this scrambling, is called an OC-1 signal, where OC stands for optical carrier.

SONET allows multiplexing several STS-1 signals into a higher level STS-N signal. This is done by dividing each column in the SONET frame into N columns. Obviously, then, an STS-N frame contains N times as many bytes as an STS-1 signal and involves a physical layer data rate N times higher. The corresponding optical signal would then be called an OC-N signal. According to existing SONET standards, values of N equal to 3, 12, 24, 48, 192, and 768 are allowed. These values correspond to approximate data rates of 155 Mbps, 622 Mbps, 1.25 Gbps, 2.5 Gbps, 10 Gbps, and 40 Gbps, respectively. In practice, the electrical STS-N and optical OC-N signals are not usually differentiated, and both signals are referred to under the generic OC-N designation.

Table 2.1 SONET and SDH signal levels and physical layer data rates

SONET		SDH	
Logical signal	Optical signal	SDH signal	Data rate (Mbps)
STS-1	OC-1	Not defined	51.84
STS-3	OC-3	STM-1	155.52
STS-12	OC-12	STM-4	622.08
STS-24	OC-24	STM-8	1,244.16
STS-48	OC-48	STM-16	2,488.32
STS-192	OC-192	STM-64	9,953.28
STS-768	OC-768	STM-256	39,813.12

The SDH signal structures are very similar to SONET, with the difference that SDH does not have an equivalent for STS-1. Instead, it starts from the equivalent of STS-3, or 155.52 Mbps. In SDH, such a signal would be called a synchronous transport module level 1, or STM-1 signal. The designation for higher data rates is STM-N, where N=1, 4, 16, 64, and 256. Thus, in order to obtain the equivalent SONET signal we should multiply N by 3. For example, a STM-16 signal in SDH is equivalent to an OC-48 signal in SONET. Moreover, in SDH no distinction is made between the logical and optical signals, and they are both represented under the same STM level signal designation. Table 2.1 includes a summary of the SONET and STM signals.

The above discussion gives an overview of SONET and SDH. For the most part, from a physical layer perspective the most important parameter of a SONET/SDH signal is its data rate. There are many details about the internal structure of a SONET/SDH frame and methods of multiplexing lower signal levels into higher levels. But such details fall outside the scope of this book [22,24].

2.8 WDM networks

Throughout the evolution of telecommunications, one of the few constants in the otherwise evolving scene has been the need for higher data rates. This has been a direct consequence of the replacement of text-oriented traffic with bandwidth-intensive graphics and video-rich applications. As a result, the need to squeeze more bandwidth out of a channel has always been present and is likely to continue for the foreseeable future. One solution to this problem has come from the time division multiplexing schemes like SONET which allow for packaging data in higher and higher data rates.

Time division multiplexing is a very effective tool, but it has its own limitations. In particular, it becomes increasingly difficult to handle higher speed signals at the physical layer. It is more difficult to modulate light at higher data rates, and higher data rate optical signals suffer more from effects such as dispersion. As a result, the physical reach of an optical link decreases as the data rate increases.

A different approach to use the wide bandwidth potential of optical fibers is the wavelength division multiplexing (WDM) technology [25–28]. In this technology, multiple wavelengths of light are modulated separately and sent into the fiber simultaneously. As long as the power within each signal is not too high, the fiber acts as a linear medium, the interaction of different wavelengths on each other will be negligible, and each wavelength propagates in the fiber independent of the others. [29]

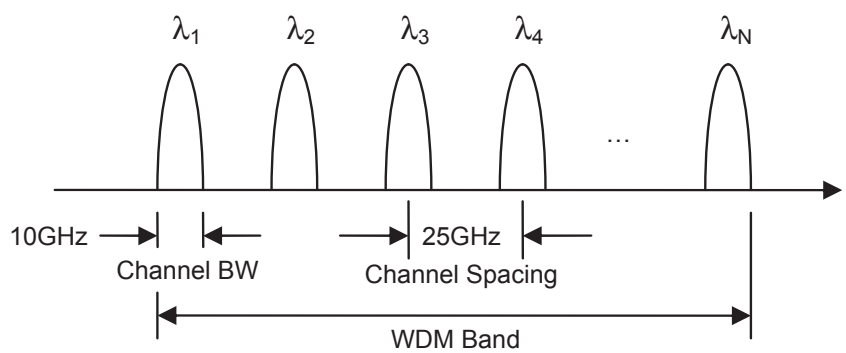

Fig. 2.8. WDM multiplexing

As a result, the effective usable bandwidth of the channel will be multiplied by the number of wavelengths in use. Figure 2.8 illustrates an example of WDM multiplexing. Here *N* wavelengths have been used, each modulated with 10 GHz of bandwidth. The wavelengths are spaced 0.2 nm apart in wavelength, which corresponds to a frequency spacing of 25 GHz. All these signals can travel together in a fiber, as long as they all fall within a low-loss transmission window. For instance, the C band is defined between 1530 and 1565 nm, which covers approximately 4.3 THz of bandwidth.[3] Given a 25 GHz channel spacing, we can fit approximately 170 WDM channels in the C band, while the electronic circuits that modulate and demodulate light operate at the reasonable speed of no more than 10 GHz. In practice, the wavelengths used in WDM systems are set by standards [30]. The recommended value for channel spacing is 50 GHz (0.4 nm) and 100 GHz (0.8 nm), referenced to a frequency of 193.1 THz (1552.524 nm). In a practical WDM link, as many as 80 wavelengths can be used, which would result in a tremendous increase in bandwidth otherwise unachievable.

[3] For a discussion of optical spectral band designations and transmission windows, see Chapter 5.

WDM technology has also opened new possibilities in the area of optical networking. The different types of networks we have discussed so far use a mixture of optical and electrical technologies. As we noted, the common trend is that as the range of a network and/or the amount of traffic a network carries increase, optical fibers become more attractive and are more widely used. However, the role of optics in these networks is limited mainly to data transmission. Higher level functions are realized in the domain of electronics.

What makes it hard for optics to play a more active role at the higher layers is the difficulty of processing signals in the optical domain. Thus, we have to use optical-electrical-optical (OEO) converters, where signals must be converted from the optical domain to electrical domain, processed in the electrical domain, and then converted back into optical domain and sent into fibers toward their destination [31]. Such a network is sometimes called a *non-transparent network*, meaning there is no direct optical path between the source and the destination.

A disadvantage of an OEO link is that at higher data rates the burden on the electronics in the routers and switches increases, causing higher cost and complexity. Thus, it would be useful to find ways to keep optical signals within the optical domain, and only convert them to the electrical domain at the destination. WDM technology provides one of the few practical ways of achieving this goal. In principle, this is doable because WDM allows access to components of an aggregate signal in the optical domain. Without WDM, all the various streams multiplexed within a signal are bundled into one optical signal. Therefore, any data processing on any of the individual signals requires a conversion to electronic format. In other words, the optical signal must be terminated, and after processing, regenerated.

This is what happens, for instance, in a SONET Add/Drop Multiplexer (ADM), where a higher level SONET signal can be decomposed into its constituent lower level SONET signals. Some of those signals whose destination is different can be extracted (or dropped), and other SONET streams whose destination is the same can be added into the stream from. The regenerated SONET signal can then be sent back into the fiber to continue its path toward its destination. In this way the ADM is very much like a bus station, and the individual lower rate signals like the passengers in a bus. The station provides an opportunity for individual passengers to get off the bus, while allowing other passengers to get on. The bus can then continue its trip toward its destination.

WDM allows a similar process to take place only in optical domain. The reason is that the individual data streams that have made up the multiplexed signal can be distinguished optically. This optical "marker" is the wavelengths of individual streams, which provides granularity at optical level. This allows for certain routing functions to be carried out completely in optical domain. In this way, we get closer to the realizations of all optical networks.

WDM technology depends on advancements in several fronts. From the standpoint of networks, the appearance of wavelength switching and routing devices has been a key enabling development. Two examples of these devices are *Optical Add/Drop Multiplexers* (OADMs), and *Optical Crossconnects* (OXCs).

An OADM is a device that operates on a WDM signal and allows extracting or adding some wavelengths at a node while allowing other wavelengths to pass through without interruption. For example, the signals whose destinations are the local node can be terminated and extracted from the WDM signal, while the signals that the local node needs to add to the stream can be added. Other wavelengths whose destinations are elsewhere continue through the OADM.

An OADM can be implemented in a variety of ways, but Fig. 2.9 shows a typical block diagram. Let us consider one direction of traffic flow, say, from left to right. A WDM signal comprising N wavelengths λ_1 to λ_N enters a WDM demultiplexer which separates the signal into individual wavelengths. Each wavelength goes through a 2×2 switch, which can be in one of the two states: it can either directly connect its input and output ports or it can cross connect the inputs and outputs to the local ports. In the former case, the wavelength passes through the switch undisturbed. In the latter case, the wavelength is dropped from the traffic to be used locally. If needed, a different signal can be added to the traffic on the same wavelength. The outputs of the switches are then multiplexed again and sent out on the other nodes.

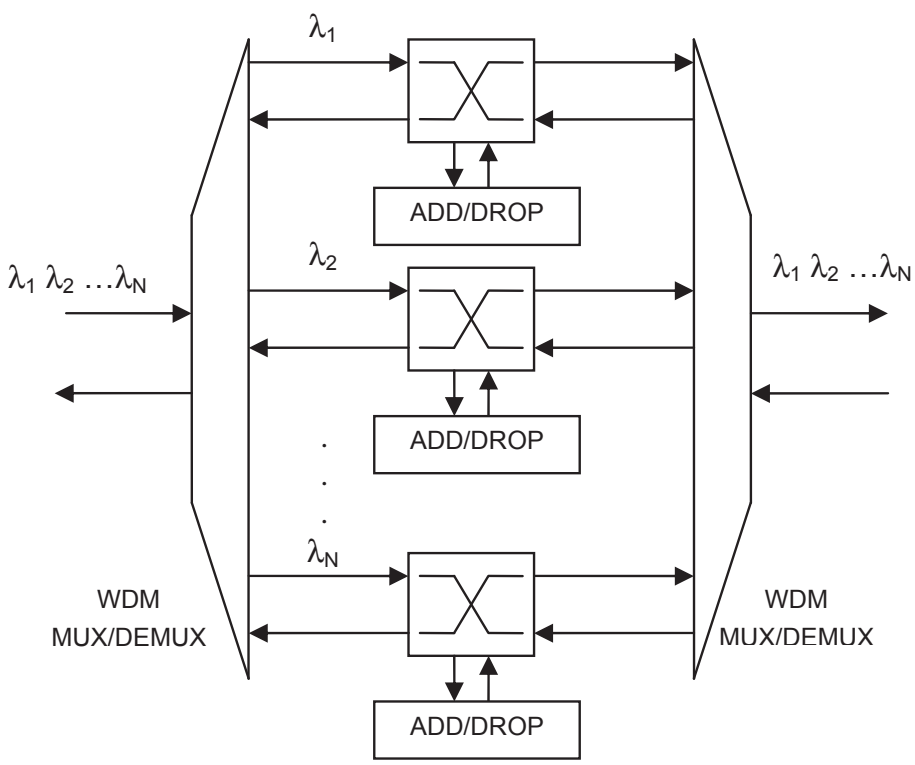

Fig. 2.9. An optical add/drop multiplexer (OADM)

An OADM operates on a single fiber and is a useful device for networks with simple topologies. However, for more complex topologies, a multi-port device is needed which would be capable of arbitrary routing of wavelengths between sev-

eral fibers. This is a function an OXC can fulfill. Figure 2.10 shows the concept of an OXC. Each optical port is connected to a fiber which carries *m* wavelengths. Each fiber goes to a demultiplexer which decomposes the signal into various wavelengths. There are also *m* $N \times N$ optical switches. Each switch is dedicated to a particular wavelength and can connect one of its inputs to one of its outputs. The outputs from all the switches are then multiplexed, and are coupled back into the *N* output ports. It can be seen that an OXC is a very flexible device, as it is capable of selectively switching and routing different wavelengths between various fibers.

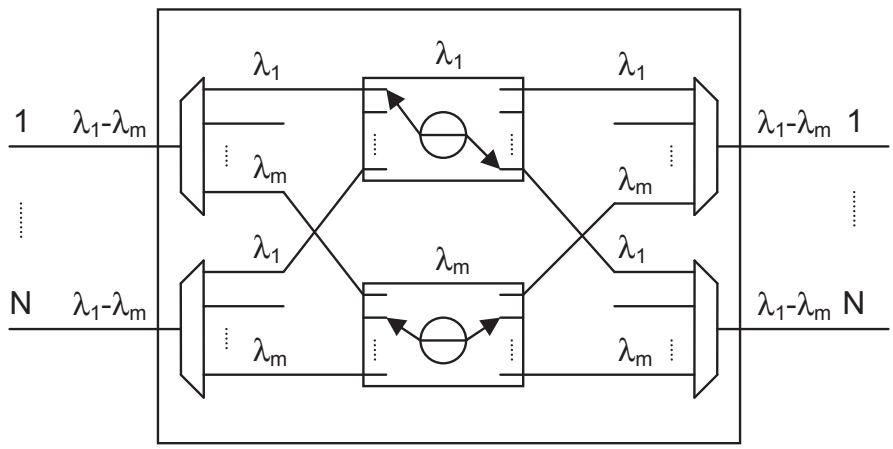

Fig. 2.10. An optical cross-connect (OXC)

To illustrate the application of these devices in an optical WDM network, let us consider Fig. 2.11 which depicts a network with five nodes, A–E, and four wavelengths, λ_1–λ_4. In this network node A transmits signals on three wavelengths λ_1–λ_3. These wavelengths go through an OADM which drops λ_1 to node B. Therefore, node A communicates to node B through λ_1. The other two wavelengths λ_2 and λ_3 pass through the OADM and arrive at the OXC. The OXC has two input and two output ports. One of the input ports receives λ_2 and λ_3 from the OADM. The other input port receives λ_4 from node C. The OXC is configured to separate λ_2 and λ_3 and combine λ_3 with λ_4 and route them to node E, while sending λ_2 to node D. Thus, A is connected to B via λ_1, A is connected to D via λ_2, A is connected to E via λ_3, and C is connected to E via λ_4.

This example can be used to highlight a few important concepts. First, note that connections are established between nodes via *lightpaths*. A lightpath is a continuous uninterrupted optical link, akin to a virtual fiber, along which signals travel without being converted back and forth between electrical and optical domains. An example of a lightpath is the dashed line shown in Fig 2.11 representing λ_2 that originates from A, passes through the OADM, and is routed to D via the OXC. This makes such a WDM network an example of a circuit switching net-

work, because the lightpath acts as a real-time connection that establishes a dedicated channel between two nodes.

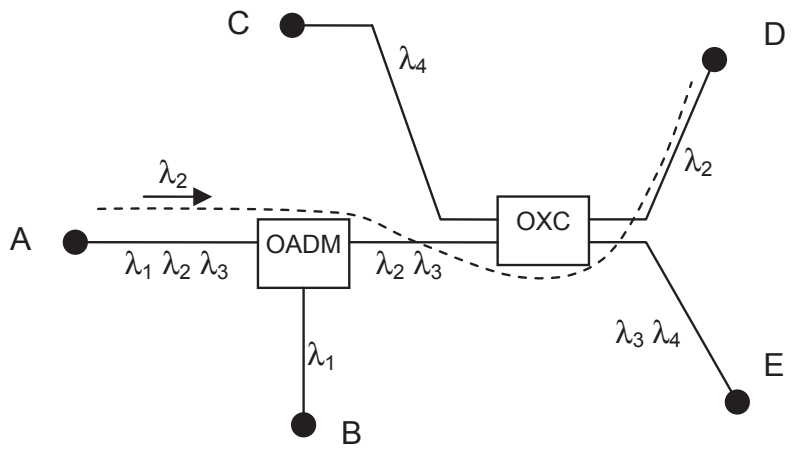

Fig. 2.11. Wavelength routing in a WDM network

Moreover, because each wavelength is routed and switched in optical domain, the lightpaths are *protocol agnostic*. Another way to state this is to say that several protocols can simultaneously use the networks without interfering with each other. This is a very important feature, because it makes the network very flexible. In practice, for instance, a wavelength can be leased to a customer, and the customer chooses the protocol and the data rate of the leased channel without having to worry about compatibility issues or interference with other traffic that use the same fibers.

Note also that the addition of the OXC adds considerable flexibility to the network. Obviously, as a router it can control the flow of traffic. For instance, it can swap λ_2 and λ_4 by routing λ_4 to D and λ_2 to E. In that case C would be connected to D and two channels would be established between A and E. This *dynamic reconfigurability* allows for the topology of the network to be changed based on real-time parameters such as the status of the traffic or load on the nodes.

From this simple example it can be seen that WDM networks offer great potentials for wideband optical networking. This is further illustrated by the fact that the standards bodies have recognized the potentials in this technology and are already moving toward providing encompassing standards and protocols for taking advantage of these capabilities. Emerging technologies and protocols such as optical transport networks (OTN) are establishing the basis upon which next-generation optical networks are expected to be based [32–33].

2.9 Passive optical networks (PONs)

The need for wider bandwidths is not limited to large companies. Increasingly, smaller customers such as residential complexes and individual homes require wider bandwidths. This need is driven by data-intensive applications such as high-definition TV (HDTV), bandwidth-intensive Internet applications, as well as other voice and video services. As a result, network architectures known as fiber to the curb (FTTC), fiber to the home (FTTH), fiber to the premise (FTTP), or more generically fiber to X (FTTX) have been developed [34–36].

In an FTTX network some or all of copper cables that are normally used to establish the communication link between end users and the larger communication infrastructure are replaced by fibers. The basis for most FTTP architectures is a passive optical network (PON), shown in Fig. 2.12.

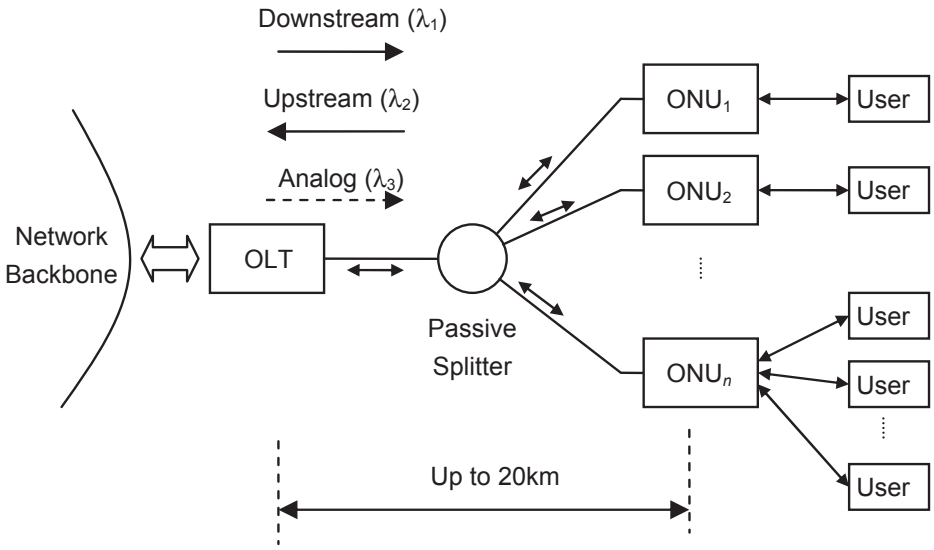

Fig. 2.12. Typical passive optical network (PON) architecture

A PON system consists of an optical line terminal (OLT) and a number of optical network units (ONUs). The ONU is also called optical network terminal or ONT. Thus, the terms ONU and ONT are often used interchangeably. The OLT is located at a central office, and the ONUs are located close to the end users. The connection between the OLT and the ONUs is established by optical fibers and through a passive splitter. The OLT provides the connection of the PON to the network backbone. On the other hand, the ONUs establish the connection to the individual users. The individual user can be a single subscriber, or a number of subscribers, like in the case of a residential complex.

In a PON architecture, the traffic flow from the OLT to the ONUs is usually referred to as *downstream traffic*, while the direction from the ONUs to the OLT is referred to as *upstream traffic*. The downstream and upstream traffics are typically

carried over different wavelengths. Some PON deployments also use a third wavelength in the downstream direction for an analog overlay, usually for a video signal. Therefore, a PON deployment is an example of WDM architecture. Moreover, PONs can be symmetric or asymmetric in terms of data rate, as the transmission rates in the downstream and upstream directions are usually not equal. Most PON systems are *asymmetric* in this respect, i.e., the downstream data rate is higher than upstream data rate. This is because the end user's receiving bandwidth needs are typically higher than their transmitting bandwidth needs.

As can be seen from Fig. 2.12, the central piece in a PON is a passive splitter/coupler. We will discuss passive devices such as couplers in Chapter 7 in more detail. The transmitted signal from the OLT is divided equally between the ONUs through this splitter. Thus, if the output power of the OLT is P, the power received by each ONU is roughly P/n, where n is the splitting number, also known as the *splitting ratio*. A 50:50 split represents a 3 dB loss, thus, a splitting ratio of n represents a loss of $3 \times \log_2(n)$. For example, a splitting ratio of 32 represents a splitting loss of 15 dB. Obviously, there is a trade-off between the splitting ratio and the physical distance the PON can support. Most PONs are implemented with a splitting ratio of 16 or 32, with a typical distance of 20 km. Another constraining factor is data rate, as higher data rates imply reduced distance.

Note that in this architecture, in the downstream direction, all ONUs receive the signal transmitted by the OLT. As such, a PON is an example of a broadcast network. Consequently, if the OLT intends to address only a specific ONU, it must add additional addressing information to the data to specify which ONU is the target of the message. On the other hand, in the upstream direction, the signals sent by individual ONUs are added together by the coupler and received simultaneously by the OLT. To prevent the mixing of signals, a time division multiplexing scheme is used that ensures only one ONU transmits at any given time. This scheme is shown in Fig. 2.13. Each ONU is assigned a particular time slot during which it can transmit its data, while it should stay off in the remaining time slots. In this way, the OLT can distinguish between the signals from various ONUs.

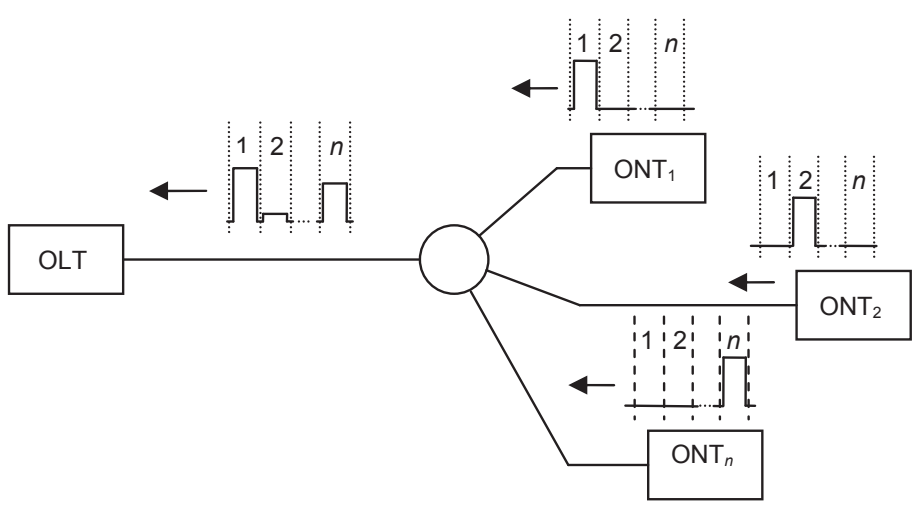

Fig. 2.13. Upstream burst mode traffic in a PON

The fact that the ONUs can be at different physical distances from the OLT complicates the structuring of traffic in the upstream direction. When the PON is being set up, through a process known as *ranging*, the OLT interrogates each ONU separately to determine the round trip delay associated with each branch. Only after knowing each individual delay exactly can the OLT assign time slots to the ONUs; otherwise the packets issued by the ONUs can collide. Because these delays are not completely deterministic and because the ONUs need some time to turn on and off, a *guard time* is inserted between packets to reduce the collision probability.

Another consequence of the difference in physical distances is that the optical powers of signals originated from different ONUs are different. The signals from closer ONUs are stronger than signals from farther ONUs. As a result, once the signals are added, the average power in the resulting stream may vary from one packet to the next. The *burst mode* nature of traffic in the upstream direction creates special challenges both for ONU transmitters and OLT receivers. We will discuss some of these challenges in Chapters 8 and 9.

From the above discussion it becomes clear that a particular PON implementation involves several parameters, among which are the upstream and downstream data rates and wavelengths, the splitting ratio, and the distance. Several standards exist that determine these parameters. Currently the most popular standards are broadband PON (BPON), Ethernet PON (EPON), and Gigabit PON (GPON). Work has also started in defining 10G-based PONs. We will revisit these standards again in Chapter 12.

2.10 Summary

Fiber optics is a crucial part of the communication networks. Therefore, a full understanding of fiber optics is not possible without having a general perspective on some main networking concepts and the role that fiber optics plays within the telecommunication infrastructure. Thus, this chapter presented an overview of networking concepts with an emphasis on the role of fiber optics communications.

We started first by examining networks from a topological perspective and reviewed common topologies such as star, ring, and mesh. The topology of a network does not say anything about its physical properties. Instead, it describes the logical structure of a network.

From the perspective of rang, networks are generally divided into LANs, MANs, and WANs, in the order of physical area they cover. Although the distinction between these categories can sometimes be fuzzy, this classification is still a useful framework to keep in mind, especially because a network's management requirements oftentimes depend on the geographical area it is covering.

It is also useful to look at networks from both a "horizontal" and "vertical" perspective. Aspects such as topology or physical distance can be considered horizontal, as they deal with connectivity between various nodes in a network or how far a

network is extended. However, equally important are issues related to the hierarchy within each individual node, i.e., the way data is organized and handed from higher levels down to the physical layer and vice versa. A popular model for this purpose is the seven-layer OSI model. Not all networks can be mapped directly into the OSI model; nevertheless, the OSI model gives a useful framework for understanding and categorizing networking hierarchies. In the OSI model each layer acts as an interface between the layer above and the layer below it and hides the unnecessary details of each from the other.

Another way to categorize networks is according to the way they transfer information. From this point of view, networks can be divided into packet switching and circuit switching. In packet switching networks, data is divided into separate packets, and each packet is individually routed by the network. Thus, in principle, various packets belonging to the same set of data can travel through different routes and arrive at their destinations at different times. The advantage of this approach is bandwidth efficiency. However, time-sensitive applications such as voice or video may suffer. On the other hand, in circuit switching networks a direct connection in real time is established between the source and the destination. This may give rise to inefficient use of bandwidth, because the actual amount of data that needs to be transferred may not need a full dedicated channel. However, time-sensitive applications can run more reliably because a certain level of quality of service is guaranteed by the nature of the link. A well-established example of circuit switching protocols is SONET, which is in wide usage in fiber optic communications.

Traditional networks utilize optical links at the physical layer, i.e., mainly for transfer of information from one node to another. The need to conduct all higher level networking functions in the electronic domain is a big burden and a limiting factor in achieving higher speeds. It is expected that transferring more complex tasks into the optical domain will increase the speed and efficiency of networks. Wavelength division multiplexing (WDM) technology is the primary candidate for realization of this goal. In WDM networks, the ability to dedicate various wavelengths to different and independent data streams has created exciting new opportunities, yielding to concepts such as wavelength switching and wavelength routing. WDM technology has also allowed for much more efficient use of fiber bandwidth, because each wavelength can be modulated separately without much interference from other wavelengths present in the fiber.

Fiber optics is also expanding its reach toward individual users. An exciting new development in this area involves fiber to the home (FTTH) networks. These networks provide very wide bandwidths to the end users through fiber. FTTX networks are mainly based on passive optical network (PON) architecture, where several users are connected to a central office through optical fibers passively coupled together. Traffic is bidirectional, and separate wavelengths carry the upstream and downstream data. These networks provide another example of the potentials of fiber optics and the key roles it can play in modern telecommunication infrastructure.

References

[1] J. Berthold, A. A. M. Saleh, L. Blair, and J. M. Simmons, "Optical networking: past, present, and future," Journal of Lightwave Technology, Vol. 26, pp. 1104–1118, 2008

[2] Recommendation G.Sup42, "Guide on the use of the ITU-T, Recommendations related to optical technology," ITU-T, 2008

[3] H. Masahiro and T. Abe, "Evaluating reliability of telecommunications networks using traffic path information," IEEE Transactions on Reliability, Vol. 57, pp. 283–294, 2008

[4] S. Y. Kuo, F. M. Yeh, and H. Y. Lin, "Efficient and exact reliability evaluation for networks with imperfect vertices," IEEE Transactions on Reliability," Vol. 56, pp. 288–300, 2007

[5] S. Soh and S. Rai, "An efficient cutset approach for evaluating communication network reliability with heterogeneous link-capacities," IEEE Transactions on Reliability, Vol. 54, pp. 133–144, 2005

[6] Y. Chan, E. Yim, and A. Marsh, "Exact & approximate improvement to the throughput of a stochastic network," IEEE Transactions on Reliability, Vol. 46, pp. 473–486, 1997

[7] N. F. Maxemchuk, I. Ouveysi, and M. Zukerman, "A quantitative measure for telecommunications networks topology design," IEEE/ACM Transactions on Networking, Vol 13, pp. 731– 742, 2005

[8] A. Valdar, *Understanding Telecommunications Networks*, Institution of Engineering and Technology, London, 2006

[9] T. Anttalainen, *Introduction to Telecommunications Network Engineering*, Artech House, London, 2003

[10] M. El-Sayed and J. Jaffe, "A view of telecommunications network evolution," IEEE Communications Magazine, Vol. 40, pp. 74–78, 2002

[11] IEEE802.3, 2005, available from www.ieee.org

[12] M. Huynha and P. Mohapatra, "Metropolitan Ethernet network: a move from LAN to MAN," Computer Networks, Vol. 51. pp. 4867–4894, 2007

[13] Javvin Technologies, *Network Protocols Handbook*, Javvin Technologies, 2005

[14] Recommendation X.200 [ISO/IEC 7498-1.1994] "Information technology-open systems interconnection-basic reference model: the basic model," ITU-T, 1994

[15] T Tuma, et al., "A hands-on approach to teaching the basic OSI reference model," International Journal of Electrical Engineering Education, Vol. 37, pp. 157–166, 2000

[16] L. Raman, "OSI systems and network management," IEEE Communications Magazine, Vol. 36, pp. 45–53, 1998

[17] D. Wetteroth, *OSI Reference Model for Telecommunications*, McGraw-Hill, New York, 2002

[18] W. Stallings, *Data and Computer Communications*, Prentice Hall, Englewood Cliffs, NJ, 2007

[19] R. A. Thompson, "Operational domains for circuit- and packet-switching," IEEE Journal on Selected Areas in Communications, Vol. 14, pp. 293–297, 1996

[20] GR-253, "SONET Transport Systems: common Criteria," Telecordia, 2005

[21] G.957, "Optical interfaces for equipments and systems relating to the synchronous digital hierarchy," ITU-T, 2008

[22] H. G. Perros, *Connection-Oriented Networks, SONET/SDH,ATM, MPLS, and Optical Networks*, John Wiley & Sons, Hoboken, NJ, 2005

[23] V. Alwayn, *Optical Network Design and Implementation*, Cisco Press, Indianapolis, IN, 2004

[24] U. Black, *Optical Networks, Third Generation Transport Systems*, Prentice Hall, Englewood Cliffs, NJ, 2002

[25] B. Mukherjee, *Optical WDM Networks*, Springer, New York, 2006

[26] B. Mukherjee, "WDM optical communication networks: progress and challenges," IEEE Journal on Selected Areas in Communications, Vol. 18, pp. 1810–1824, 2000

[27] J. Zhang and B. Mukherjee, "A review of fault management in WDM mesh networks: basic concepts and research challenges," IEEE Networks, Vol. 18, pp. 41–48, 2004

[28] G. Keiser, "A review of WDM technology and applications", Optical Fiber Technology, Vol. 5, pp. 3–39, 1999

[29] A. R. Chraplyvy, ''Limitations on lightwave communications imposed by optical-fiber nonlinearities,'' Journal of Lightwave Technology, Vol. 8, pp. 1548–1557, 1990

[30] G.692, "Characteristics of optical components and sub-systems," ITU-T, 1998

[31] J. M. Simmons, *Optical Network Design and Planning*, Springer, New York 2008

[32] ITU-T, "Optical Transport Networks & Technologies Standardization Work Plan," 2007. Available from http://www.itu.int/itudoc/itu-t/com15/otn

[33] G.709, "Interfaces for the Optical Transport Network (OTN)" ITU-T, 2003. Available from www.itu.int.

[34] L. Hutcheson "FTTx: current status and the future," IEEE Communications Magazine, Vol. 46, pp. 90–95, 2008

[35] M. Nakamura et al., "Proposal of networking by PON technologies for full and Ethernet services in FTTx," Journal of Lightwave Technology, Vol. 22, pp. 2631–2640, 2004

[36] G. Keiser, *FTTX Concepts and Applications*, John Wiley & Sons, Hoboken, NJ, 2006

Chapter 6
PIN and APD Detectors

6.1 Introduction

As mentioned in Chapter 1, in a fiber optic link an optical source, such as semiconductor laser, converts an electrical signal to an optical signal. The optical signal, once coupled properly into an optical fiber, can travel as a guided wave for relatively long distances. At destination, the optical signal must be converted back from the optical domain to the electrical domain. This conversion is accomplished by using a photodetector, which is a light-sensitive device that converts the received photons into electrons.

There are a wide variety of photodetectors that can be used for different purposes. In fiber optics, two types of photodetectors are of primary interest: PIN diodes and APD diodes. Almost all practical fiber optic receivers use one of these two devices at their front end. Therefore, this chapter is dedicated to these two detector structures. Later in Chapter 9, we discuss complete optical receiver circuits, in which the electrons generated by the detector are converted into a useful electrical signal that represents the original data sent from the transmitter.

6.2 The PIN diode and photon–electron conversion

As noted in Chapter 1, the generation and detection of light is a phenomenon that is most properly described by quantum physics [1]. The photoelectric effect provides a clear example where the absorption of photons results in the release of free electrons. If high-energy photons hit a metal, they knock out electrons from the surface of the metal. If an external electric field is applied, these electrons can be collected and generate a photocurrent. This is the operating principle behind vacuum tube devices such as vacuum photodiodes and vacuum photomultipliers. However, such devices are not very useful for fiber optic applications, because they are bulky, require high voltage, and hard to operate. Semiconductor photodetectors provide several advantages in terms of size, cost, operating voltage, responsivity, reliability, and integration with other optoelectronic devices. That is why virtually all photodetectors used in fiber optic receivers are semiconductor based.

In many ways, the process of conversion of photons to electrons in a semiconductor is the opposite of the process of electron to photon conversion that takes place in a semiconductor laser. We start by examining the process of photodetection in PIN diodes.

M. Azadeh, *Fiber Optics Engineering*, Optical Networks,
DOI 10.1007/978-1-4419-0304-4_6, © Springer Science+Business Media, LLC 2009

6.2.1 PIN diode, static characteristics

The most widely used semiconductor detector is a revered biased p-i-n (PIN) junction [2–10]. Figure 6.1 outlines the basic operating principles of a PIN photodetector. As can be seen from the figure, the PIN diode consists of a p-doped and an n-doped semiconductor separated by an intrinsic material. The junction is reverse biased by an external source. The bias voltage, V_s, appears almost entirely across the intrinsic section in the middle. This is because the intrinsic section is almost devoid of free charges, and therefore is much more resistive compared to the *p* and *n* sections. As a result, a strong electric field is formed in the intrinsic section.

When photons hit this section, they cause valence band electrons to jump into the conduction band, leaving a positive charge, or hole, behind. Thus, a population of photo-generated carriers is created in the intrinsic region. These carriers drift out of the intrinsic region because of the present electric field. The electrons move toward the *n*-region, and the holes move toward the *p*-region. This causes a photocurrent to flow in the circuit.

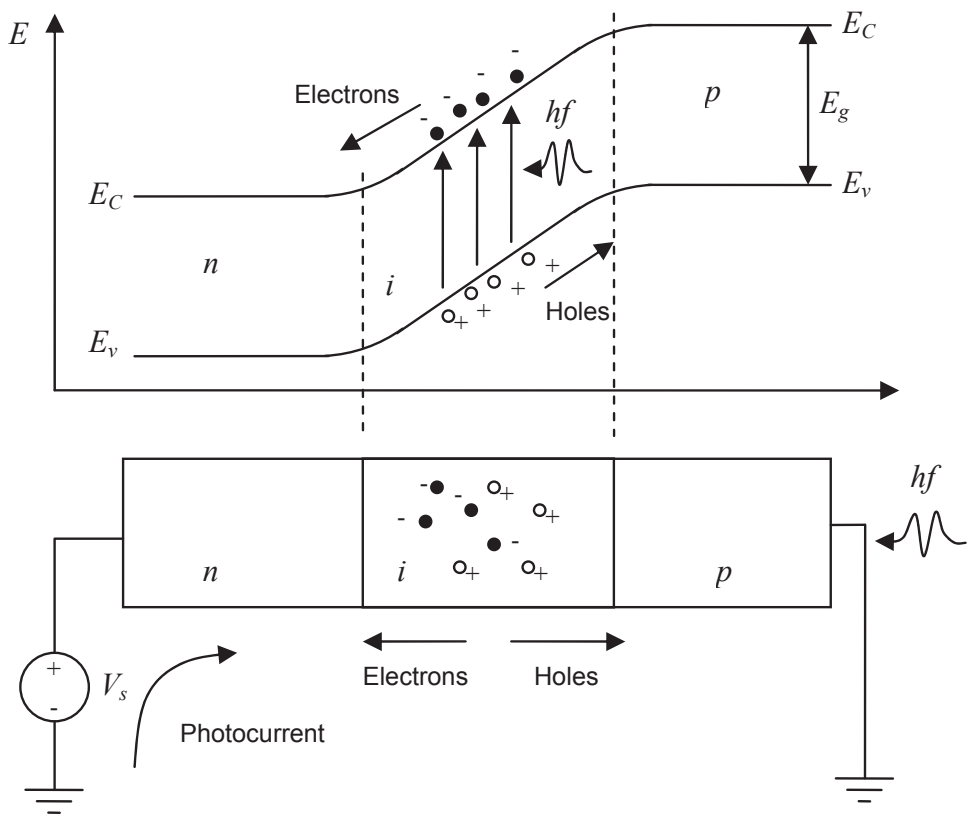

Fig. 6.1. Energy band diagram and schematic of a reverse bias PIN photodetector

From this description, it is obvious that a photon can knock an electron out of the valence band and into the conduction band if its energy exceeds the bandgap energy of the semiconductor:

$$hf \geq E_g \tag{6.1}$$

where $E_g = E_c - E_v$ is the bandgap energy, f is the photon's optical frequency, and h is Planck's constant. The efficiency of a photodetector is characterized by its *responsivity*, which is defined as

$$\rho = \frac{I}{P} \tag{6.2}$$

where I is the photocurrent, P is the optical power incident on the detector, and ρ is responsivity in units of A/W.

Equation (6.2) provides the definition of responsivity in terms of measurable quantities. It is also instructive to express Eq. (6.2) in terms of quantum mechanical parameters. This is rather straight forward. If P is the optical power and $E=hf$ is the energy of a photon, then P/E or P/hf is the number of photons that hit the photodetector. Because not every absorbed photon generates an electron/hole pair, we need to define a *quantum efficiency*, η, which represents the fraction of photons that generate an electron–hole pair and thus contribute to the photocurrent. Therefore, the number of generated electrons (or holes) in a second is given by $\eta P/hf$. Multiplying by e, the charge of an electron, we get the value of photocurrent I as $e\eta P/hf$. Finally, using the definition given by Eq. (6.2), we get

$$\rho = \frac{e\eta}{hf} \tag{6.3}$$

which is an expression relating responsivity (which is a measurable parameter) to the quantum mechanical parameters and frequency.

From a circuit point of view, a PIN detector can be thought of as a light-controlled current source. Figure 6.2 shows the typical IV characteristic of a PIN diode. As long as the junction is reverse biased, application of optical power generates a photocurrent proportional to the optical power.[1] It can also be seen that even in the absence of light application of voltage results in a small current. This is because the intrinsic region also includes thermally generated free carriers which contribute to a current even in the absence of light. This is known as the dark current.

[1] The diode can also generate photocurrent when it is forward biased. This is called the photovoltaic mode of operation, where light is converted to forward bias current. Photocells work in the photovoltaic mode. However, PIN detectors work in the reverse-biased region, also known as photo-conductive mode.

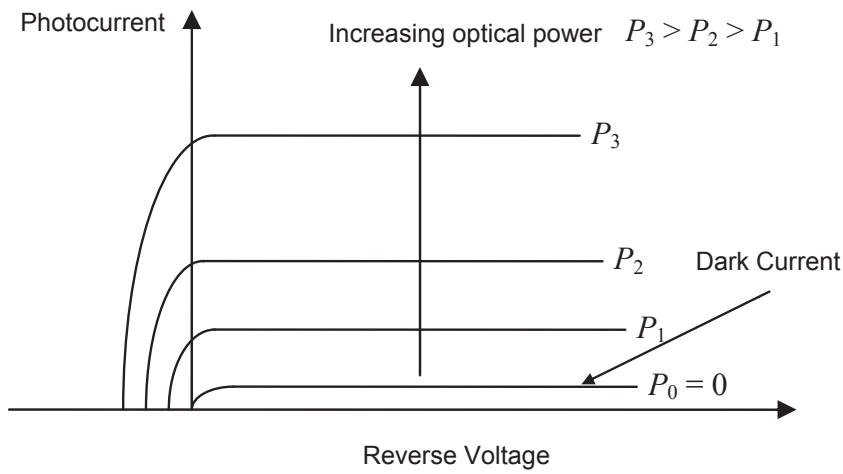

Fig. 6.2. Current vs. voltage characteristics of a reverse-biased PIN diode

The characteristic of a current source is that its voltage must be determined by other elements in the circuit. Thus, as shown in Fig. 6.3, if the diode is connected to a load resistance of R_L, the output voltage of the circuit is given by

$$V = P\rho R_L \tag{6.4}$$

This equation is valid as long as $V < V_s$. Figure 6.3 shows the simplest way to bias a PIN diode to convert the optical signal into a voltage signal. However, as a practical circuit, this scheme suffers from certain shortcomings, which we will discuss further in Chapter 9.

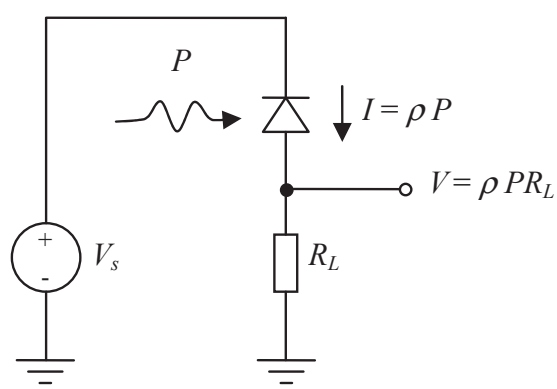

Fig. 6.3. Biasing a PIN photodetector

6.2.2 PIN diode, dynamic characteristics

The static response of a PIN diode provides a sufficient description when the input optical signal varies slowly compared to the internal time constants of the device. However, when the signal has high-frequency components, it is also important to take into account the dynamic response of a PIN diode [11].

The dynamic response of a PIN diode is determined by several factors, some of which are a result of the structure of the diode and others result from the external circuits. In general, carriers can move as a result of diffusion and drift. Drift is the main mechanism behind carrier movement in the intrinsic depletion region, because it is where the electric field is strong. On the other hand, as a result of the high conductivity in the *p* and *n* regions, the voltage drop and thus the electric field is small, and carrier movement takes place as a result of diffusion. Among the limiting factors, the most important are

- Transit time: The drift of carriers across the intrinsic region is not instantaneous, and this is a major factor that limits the response time of the detector. The transit time is a function of the speed of the carriers and the width of the depletion region. To reduce the transit time, either the speed of carriers will have to be increased by increasing the reverse voltage or the width of the intrinsic region must be reduced. Increasing the reverse voltage is effective as long as long as carrier velocities do not saturate. On the other hand, the intrinsic region cannot be made very thin because then the photons may pass through it without being absorbed [11]. Carrier speeds can be in the order of thousands of meters per second, and the width of depletion region can be around a few microns. This can result in sub-nanosecond transits times.
- Diffusion time: Diffusion is a slower process compared to drift, and therefore to reduce the response time, it is desirable to minimize the amount of carriers that are generated in the *p* and *n* regions. One way to achieve this is to reduce the length of the *p* or *n* regions. For instance, in Fig. 6.1 the photons have to pass through the *p* region before they can reach the depletion region. Therefore, by making the *p* region narrower, we can minimize the number of carriers that are generated there.
- Capacitance: Another limiting factor is the capacitance of the junction as well as the parasitic capacitance of the package. The reverse-biased PIN junction acts like a capacitor because charge can be stored in the *p* and *n* sides with the depletion layer acting as a dielectric. Notice that although reducing the length of the depletion region helps transit time, it cannot be made very narrow, because it increases the junction capacitance. The package capacitance includes the capacitance of the pads provided for wirebonding to the detector. These capacitances, in conjunction with the external load connected to the detector, form a low-pass filter.

To analyze the operation of a photodetector, it is useful to develop models that can predict its behavior based on the detector's parameters [12–15]. Figure 6.4 shows a simple linear ac equivalent circuit of a PIN diode along with the load resistor attached to it.

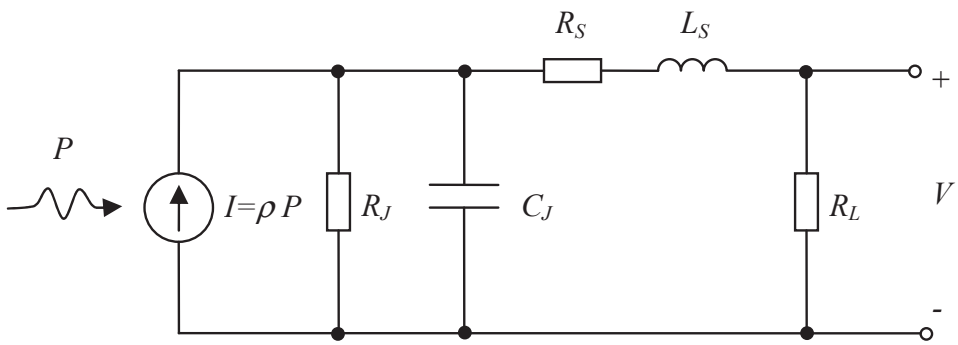

Fig. 6.4. ac model of a PIN detector

The photodetector is represented by a current source, with the junction capacitance C_J and the internal resistance of the depletion layer, R_J parallel to it. Typically there is some extra series resistance associated with the p and n regions, which is shown as R_S. The inductance of the wire bonding used to connect the diode to the load is shown by L_S, and the load itself is shown by R_L.

It is obvious that this circuit acts as a low-pass filter, with a possible resonance created by L_S and by C_J. Normally, the resistance of the depletion layer R_J is high, while the series resistance of the p and n regions is relatively low. This means R_J and R_S may be ignored for approximate analysis. Also, assuming very short wire-bondings, L_S can be ignored. With these simplifications, the circuit's behavior can be approximated as a regular low-pass RC filter, with the 3dB frequency given by

$$f_{3dB} = \frac{1}{2\pi R_L C_J} \tag{6.5}$$

Thus, in order to maximize the frequency response of the diode, both its capacitance and the resistance of the load it is attached to must be minimized. Minimizing the junction capacitance can be achieved by careful design of the diode structure, and practical highspeed diodes can have capacitances in the order of sub-pF. Minimizing the load resistance is not a function of the diode itself, but is related to the receiver's electrical design. This is a subject we will consider in Chapter 9.

6.3 Avalanche photodiode (APD)

The PIN diode described in the previous section is the most commonly used detector in fiber optics. However, in some cases the responsivity of a PIN diode is not sufficient to achieve the needed performance, in which case an avalanche photo-

diode structure (APD) is typically used. The main difference between an APD and a regular PIN detector is that an APD provides inherent current gain. This is a clear advantage, because in a receiver, if the signal is amplified in an earlier stage, the signal-to-noise ratio will be improved. In an APD, the reverse bias voltage applied to the junction is much higher. As a result, once a photon generates an electron–hole pair, the intense electric field present in the depletion region greatly accelerates the pair. These high-energy carriers then collide with other atoms. As a result of these collisions the atoms are ionized, releasing additional pairs of electron holes. This process is known as *impact ionization* or *avalanche multiplication*. Therefore, a single photo-generated carrier pair can ultimately generate many more carriers, resulting in a very high current gain.

Figure 6.5 shows a popular APD structure known as a *reach-through* APD [16,17]. Also shown are the electric field and energy diagrams along the junction under reverse bias. Note that the band gap voltage is small compared to the reverse bias.

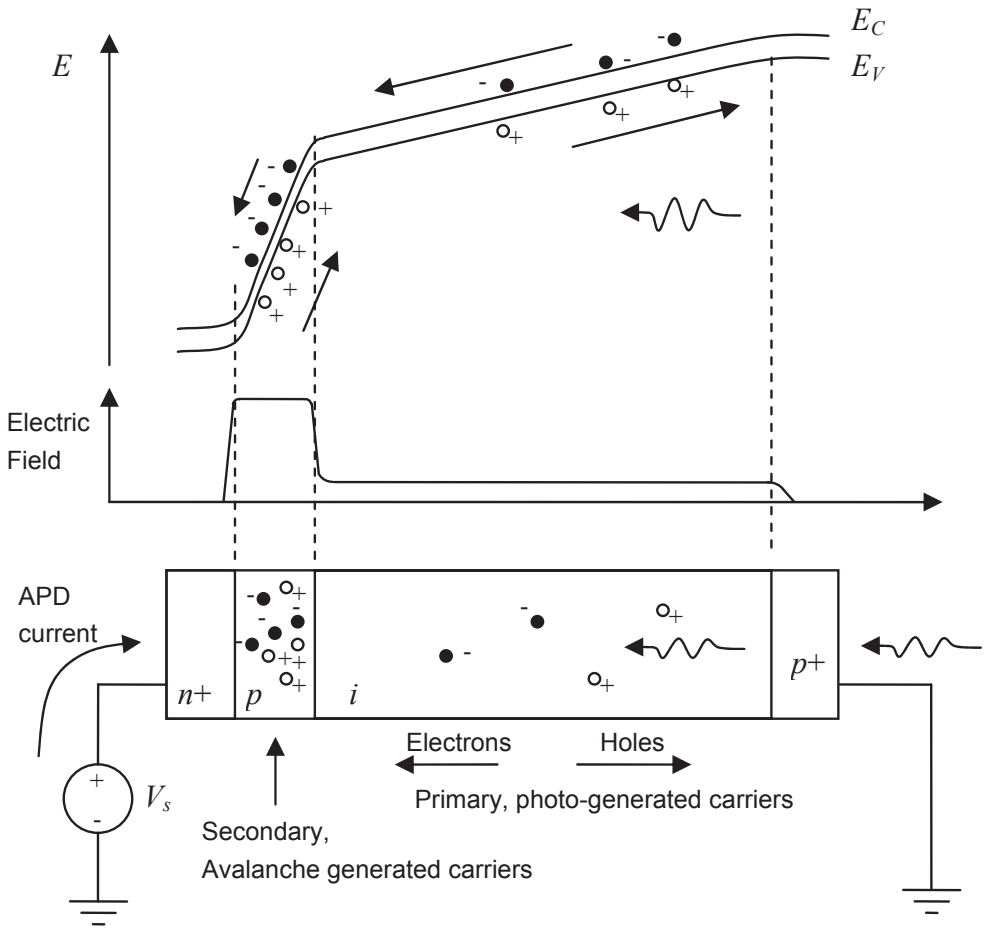

Fig. 6.5. The structure of a reach-through APD as well as the energy and electric field diagrams along the junction

The structure consists of highly doped *p*+ and *n*+ regions on either side of a lightly doped, almost intrinsic region in the middle. The intrinsic region is sometimes referred to as the π region. There is also an additional *p* layer sandwiched between the *n*+ and *p* regions. When the structure is reverse biased, most of the voltage drop appears across the *n*+*p* junction. The photons enter from the *p*+ side, and are absorbed in the intrinsic region where they generate electrons and holes. These are primary photo-generated carriers. The holes drift back toward the *p*+ side. But the electrons drift toward the reverse bias junction where the electric filed is intense. In a reverse bias junction, the width of the depletion region on either side of the junction is inversely proportional to the doping level. Therefore, in this structure, the depletion region extends in the *p* side, and with enough reverse bias, "reaches through" to the intrinsic region. Once the electrons enter the high electric field region, they accelerate and cause avalanche multiplication through impact ionization. These *secondary carriers* are the main contributors to the overall APD current. Note that in this device only electrons migrate toward the high field area and initiate the avalanche multiplication. This is beneficial for noise purposes, because the noise performance of devices that rely on one primary carrier is better than the performance of those relying on both carriers.

The current gain of an APD, denoted by *M*, is a function of the reverse bias voltage. As the reverse bias voltage approaches the *breakdown voltage*, V_{BR}, the gain starts to increase. An empirical description of this behavior can be given by the following formula [18]:

$$M = \frac{1}{1-\left(V / V_{BR}\right)^n} \tag{6.6}$$

where *V* is the reverse bias voltage and *n* is an empirical parameter that varies between 2 and 7. For long wavelength detectors based on InGaAs, typical breakdown voltages can range from 30 to 100 V, and gains of 10–30 are achievable. Gains of 100 and higher can be achieved by silicon-based APDs, but the breakdown voltage is higher, and they cannot be used for long-wavelength applications. We will discuss the wavelength response of different semiconductors later in this chapter.

It may seem from Eq. (6.6) that it would be better to bias the APD as close to the breakdown voltage as possible to maximize the APD gain. In practice, however, there is a relatively narrow optimal range to bias an APD. Beyond that range, the additional gain of the APD is quickly offset by excess noise.

Figure 6.6 shows a typical plot of *M* and signal-to-noise ratio as a function of bias voltage. The optimal bias range may lie up to few volts below the breakdown voltage. Biasing the APD beyond that point will result in a lower signal-to-noise ratio, thus deteriorating the sensitivity of the receiver.

We should also note that the gain of an APD decreases at higher temperatures. This is because at higher temperatures the mean free path between carrier-atom impacts decreases.

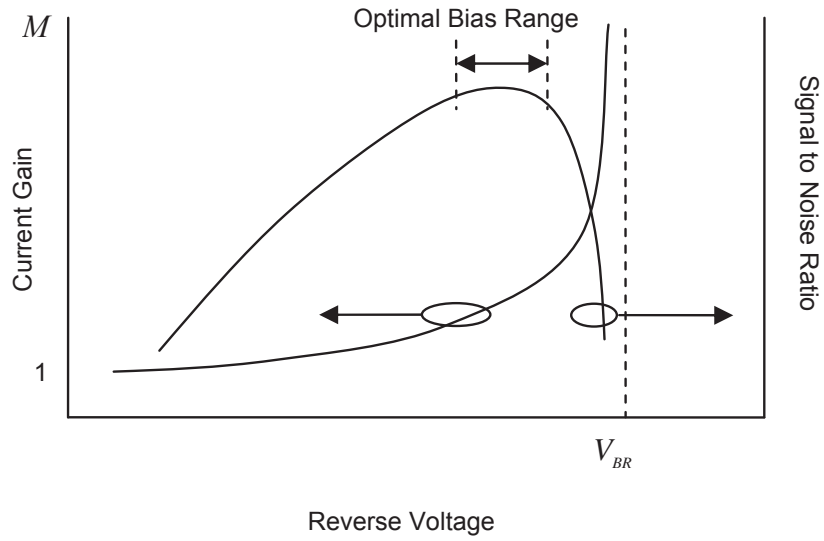

Fig. 6.6. Signal-to-noise ratio and current gain vs. reverse bias voltage

In other words, as a result of more thermal vibration, the carriers are scattered more frequently and do not reach high-enough velocities needed to initiate secondary ionization. As a result, the breakdown voltage generally increases with temperature. To first order, this dependence is linear, and can be described by

$$V_{BR}(T) = V_{BR}(T_0) + K_{BR}(T - T_0) \tag{6.7}$$

where T and T_0 are two temperatures, $V_{BR}(T)$ and $V_{BR}(T_0)$ are the breakdown voltages at those two temperatures, and K_{BR} is the breakdown voltage temperature coefficient in volts/°C. Typical values for K_{BR} are in the range of 0.1–0.3 volts/°C.

Another phenomenon that one needs to consider when using APDs is *gain saturation*. The current gain expression given by Eq. (6.6) holds at low optical powers. As the received optical power increases, the differential current gain reduces. Thus, an optical power–photocurrent curve starts linearly at low powers but becomes sub-linear at high optical power. This behavior can be attributed to a field screening effect of very high concentrations of free carriers, so that at very high powers, the effective electric field inside the device is reduced [19].

The dynamic performance of the APD is also crucial for highspeed applications. Here the common figure of merit is APD's *gain-bandwidth product*. We expect that biasing an APD at a higher gain should result in a reduction of its bandwidth. In reality, the gain-bandwidth product is a complex parameter and is a function of many factors, including quantum efficiency, material, device geometry, and even the gain itself. Bandwidths in the range of 20–30 GHz, and gain-bandwidth products as high as 150 GHz, have been achieved [20].

6.4 Noise in photodetectors

In discussing the gain of an APD we alluded to the fact that merely increasing the current gain in an APD does not necessarily improve the performance of a receiver. Beyond a certain point, the noise generated by the APD starts to dominate, causing the detected signal to deteriorate. Noise always accompanies physical signals, and anytime a signal is detected or amplified, the accompanying noise is also amplified. Moreover, any electronic device adds some additional noise of its own to the signal. As a result, it is common to keep track of the signal-to-noise ratio (SNR) as a signal propagates through a system. To reconstruct a signal at the receiver faithfully, the noise power must be small compared to the signal level. In other words, the SNR is a figure of merit in characterizing a receiver circuit.

In a photodetector, noise can be attributed primarily to two factors. The *shot noise*, also known as *quantum noise*, originates from the statistical nature of photon to electron conversion. *Thermal noise*, on the other hand, is an intrinsic property of any electrical circuit that is connected to the photodetector.

6.4.1 Shot noise

Photon to electron conversion is fundamentally a quantum mechanical process. When a photon is absorbed, a pair of electron–holes is generated. Therefore, the photo-generated current is not truly continuous, but has a discrete nature. It fluctuates around some average value as a result of the discrete charge of the carriers that contribute to it.

Because of the random nature of the current fluctuations, the noise current must be characterized in a statistical manner. It is common to describe the noise current by its mean square value. For a PIN detector, the mean square value of the shot noise is [18]

$$\left\langle i_N^2 \right\rangle = 2eI_P B \tag{6.8}$$

where I_P is the photocurrent, e is the electron charge, and B is the bandwidth within which the noise is being measured. Equation (6.8) implies that shot noise has a constant spectral density, an assumption that holds for all frequencies of interest. Normally, B is set by the bandwidth of the receiver. This shows that one way to minimize the effects of shot noise is to keep the bandwidth of circuit as narrow as possible.

The current flowing through a PIN diode is not just photo-generated. Any reverse bias junction has a leakage current. For photodetectors the leakage current is called dark current, because it exists even when these is no optical power. As a result, the mean square value of the total shot noise is given by

$$\left\langle i_N^2 \right\rangle = 2e\left(I_P + I_D\right)B \tag{6.9}$$

For APDs, both the photocurrent and the dark current are amplified by the inherent current gain of the device. It can be shown that the mean square value of the total shot noise for an APD is given by

$$\left\langle i_N^2 \right\rangle = 2e\left(I_P + I_D\right)BM^2F(M) \tag{6.10}$$

where $F(M)$ is the excess noise factor, and describes the statistical nature of the avalanche multiplication [21]. An expression for $F(M)$ is given by [22]

$$F(M) = M - (1-k)M^3\left(M-1\right)^2 \tag{6.11}$$

In Eq. (6.11), k is the ratio of the ionization coefficient of the holes to that of the electrons, assuming that the electrons initiate the multiplication.

Figure 6.7 shows an equivalent circuit for analysis of noise in a photodetector. This figure is a simplification of Fig. 6.4 when $R_J >> R_L$, $R_s << R_L$, and $L_S \cong 0$.

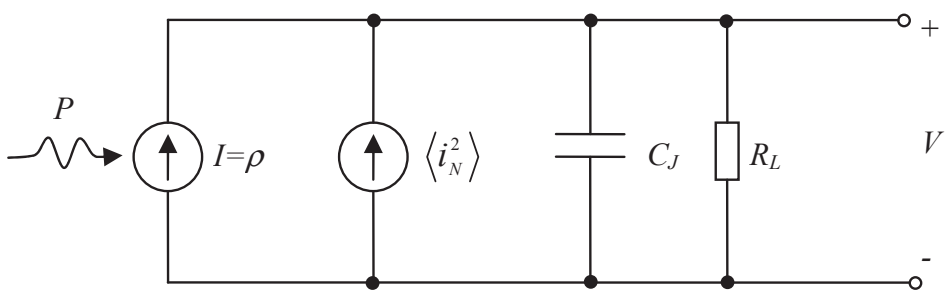

Fig. 6.7. Modeling shot noise as an extra current source

As noted earlier, the junction capacitance and the load resistance form a low-pass filter. As a result, the noise bandwidth term B in Eqs. (6.9) and (6.10) is determined by the bandwidth of this low-pass filter as defined by Eq. (6.5).

6.4.2 Thermal noise

Shot noise is a consequence of the quantum nature of light detection. Therefore, it is a fundamental property of the photodetector and sets a maximum limit on the value of SNR. In such a case, the SNR is said to be *quantum limited*. In reality, however, almost always there are other sources of noise present. Chief among

these is *thermal noise*, also known as *Johnson noise*, associated with the electric circuits connected to the detector.

The source of this noise is the thermal motion of electrons in the load resistor R_L. The mean squared of the thermal noise in the load resistor is given by

$$\left\langle i_T^2 \right\rangle = \frac{4kTB}{R_L} \tag{6.12}$$

where k is Boltzmann's constant, T is the absolute temperature, and B is bandwidth. Like shot noise, thermal noise has a constant spectral density. This is another reason to keep the bandwidth of a receiver as low as possible, i.e., just sufficient to pass the signals of interest. We can incorporate the thermal noise as an additional current source into our circuit model, with the result shown in Fig. 6.8. This figure is similar to Fig. 6.7 with the difference that all band-limiting effects including those of the junction capacitance are lumped together in the form of a low-pass filter (LPF) with bandwidth B.

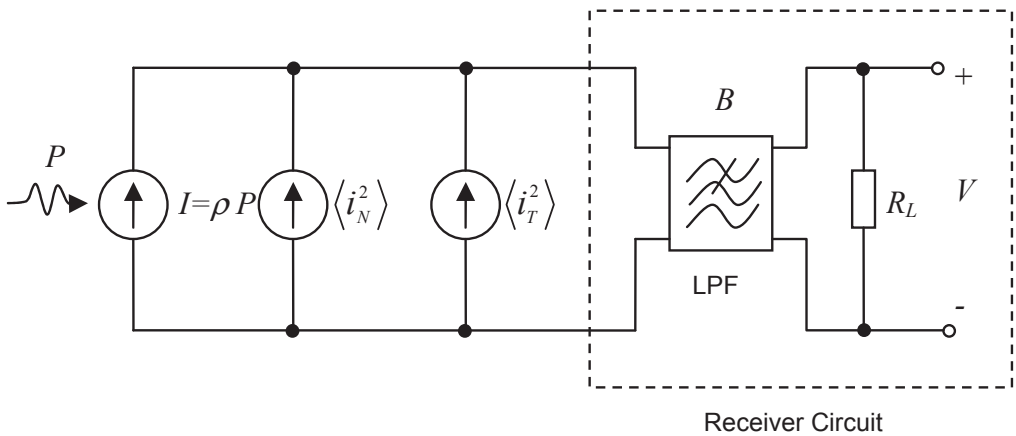

Fig. 6.8. Circuit schematic of the thermal and shot noise

If the photocurrent drives an amplifier instead of a resistor, B will be determined by the bandwidth of the amplifier. In either case, the total noise current at the input of an optical receiver is limited by the effective bandwidth of the receiver. This is why receivers with a higher bandwidth generally have a lower sensitivity.

6.4.3 Signal-to-noise ratio (SNR)

Once we have characterized the noise level at the input of a receiver, it is possible to analyze the SNR. The SNR is an important parameter because it determines the performance of a receiver. In digital receivers, SNR can be related to the bit error rate. In analog receivers, SNR is the main figure of merit and characterizes the quality of the analog link.

Let us assume that the optical power is given by a signal $P(t)$. Assuming the responsivity is given by Eq. (6.3), the detector current is given by

$$I(t) = M\frac{e\eta}{hf}P(t) \tag{6.13}$$

Note that a gain current gain term M is added to account for APD detectors. For a PIN detector, M=1. The electrical signal power associated with this current is given by

$$P_E(t) = R_L I(t)^2 = R_L M^2\left(\frac{e\eta}{hf}\right)^2 P(t)^2 \tag{6.14}$$

where $P_E(t)$ is the *electrical* power delivered to the load.

On the other hand, the total noise power delivered to the load, consisting of shut and thermal noise, can be expressed as

$$P_N = \left[\left\langle i_N^2\right\rangle + \left\langle i_T^2\right\rangle\right]R_L = 2e\left[\frac{e\eta}{hf}P(t) + I_D\right]BM^2\,F(M)R_L + 4kTB \tag{6.15}$$

where P_N is the total noise power delivered to the load. Again, the two terms M and $F(M)$ relate to APD detectors, and for a PIN diode they are both 1. From Eqs. (6.14) and (6.15) we can calculate SNR, defined as $P_E(t)/P_N$. After simplification, we obtain

$$SNR = \frac{\eta}{2Bhf}\times\frac{P(t)^2}{\left[P(t) + \frac{hf}{e\eta}I_D\right]F(M) + \frac{2hfkT}{\eta e^2 M^2 R_L}} \tag{6.16}$$

This equation provides several insights into the SNR behavior of a receiver. Notice that as expected, increasing the optical power $P(t)$ increases the SNR. On the other hand, increasing the bandwidth of the receiver, B, reduces the SNR. The denominator of Eq. (6.16) shows the contribution of thermal noise and shot noise to SNR. An interesting point is that as R_L increases, the effects of thermal noise decrease. In a practical circuit, however, R_L cannot be increased too much because it will reduce the bias headroom of the photodetector diode (refer to Fig. 6.3). As we will see in Chapter 9, using a transimpedance amplifier removes this limitation. Another point to notice is the effect of current gain, M. Higher M values reduce the effects of thermal noise, and therefore improve the SNR. That is why an APD detector provides a higher sensitivity compared to a PIN diode. However, beyond a point, the $F(M)$ term in the denominator starts to take over and will cause SNR degradation. This means that when an APD is used, there is an optimal

biasing point for best SNR. We discussed this point qualitatively in a previous section, where we noted that usually the best point to bias the APD is a few volts below the breakdown voltage (see Fig. 6.6).

Equation (6.16) can be simplified in several cases. For example, we can consider a case where the optical power is not very low, in which case the dark current I_D can be ignored. Now, consider a case where thermal noise power is much less than the shut noise. In this case, known as the *quantum* or *shut noise limit*, Eq. (6.16) simplifies to

$$SNR = \frac{\eta}{2hfBF(M)} P(t) \tag{6.17}$$

which represents the best possible SNR attainable from a detector. Note that in this case using an APD does not provide a clear advantage, except possibly in reducing the effects of thermal noise by diminishing the thermal noise term in the denominator of Eq. (6.15).

On the other hand, we can consider the case where thermal noise dominates over shut noise. In this case, Eq. (6.16) simplifies to

$$SNR = \left(\frac{e\eta M}{hf}\right)^2 \frac{R_L}{4B} P(t)^2 \tag{6.18}$$

Note that in this case, which usually corresponds to low optical powers, the SNR is more sensitive to both optical power and the current gain M. As expected, this implies that using an APD is very effective in improving the SNR at low optical powers.

6.5 Photodetector materials and structures

So far we have concentrated on the operational principles behind PIN and APD diodes. In this section we briefly review the different materials and structures that are used in building these detectors.

6.5.1 Photodetector materials

We mentioned before that in order for a photon to be absorbed and generate carriers, its energy must be larger than the bandgap of the semiconductor material. As a result, the responsivity of a detector to light is a function of the material and wavelength. This dependence is characterized by the quantum efficiency, η. This makes responsivity a function of wavelength and material.

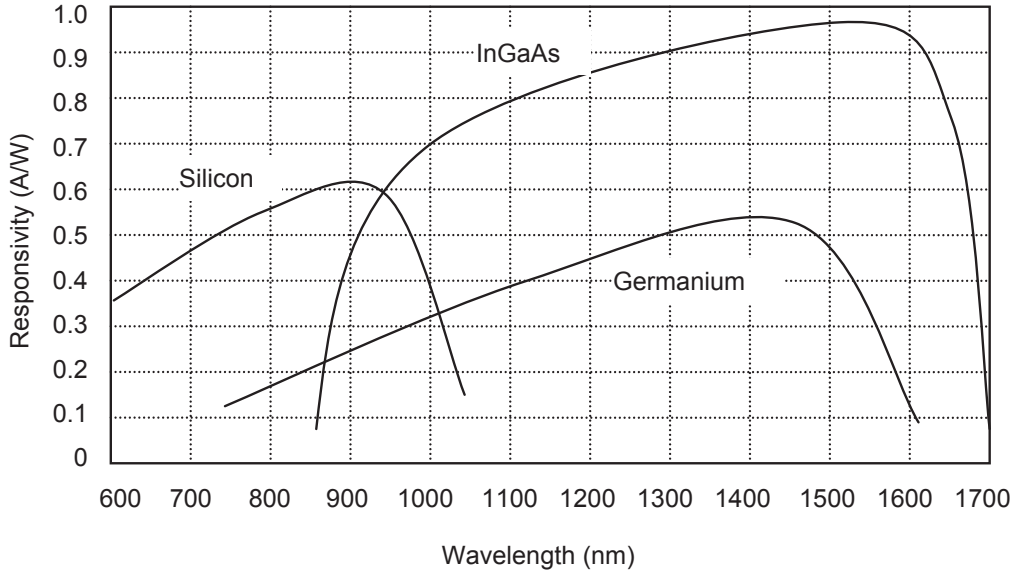

Fig. 6.9. Typical responsivity curves for silicon, InGaAs, and germanium

Figure 6.9 shows the responsivity of the common semiconductors used in photodetectors as a function of wavelength. It can be seen that the responsivity of a detector varies widely with wavelength and is maximum at a particular wavelength for each material. At longer wavelengths responsivity drops sharply because the photons do not have enough energy to knock electrons from the valence band to the conduction band. At lower wavelengths the responsivity also reduces because of increased absorption. For long wavelength applications at 1300 and 1550 nm, InGaAs is an excellent choice, as its response extends beyond the 1550 nm window. Moreover, it provides the highest responsivity, which approached 1A/W at 1550 nm. Figure 6.9 shows the inherent responsivity of the materials, without any additional current gain. In an APD, as a result of the avalanche gain M, the overall device responsivity can be much higher.

Table 6.1. Typical parameters for silicon, germanium, and InGaAs detectors

	Wavelength range (nm)	Responsivity (A/W)	Dark current (nA)	Avalanche gain
Silicon	400–1100	0.4–0.6	0.1–10	20–400
Germanium	800–1600	0.4–0.5	50–500	50–200
InGaAs	1100–1700	0.7–1	0.5–50	10–50

Table 6.1 provides a summary of typical values for the critical receiver parameters for these three materials.

6.5.2 PIN diode structures

Practical APD and PIN diode structures are based on the same principles we discussed before. Figure 6.10 shows an example of a practical InGaAs PIN diode structure. The light enters from the top and goes through an aperture surrounded by a metal contact. The aperture is anti-reflection coated to minimize light reflection. Because of the high index of refraction of semiconductors, without anti-reflection coating a large percentage of incident light will be reflected, causing a substantial reduction in the efficiency. Typical anti-reflection coatings may consist of quarter-wavelength single or multi-layers.

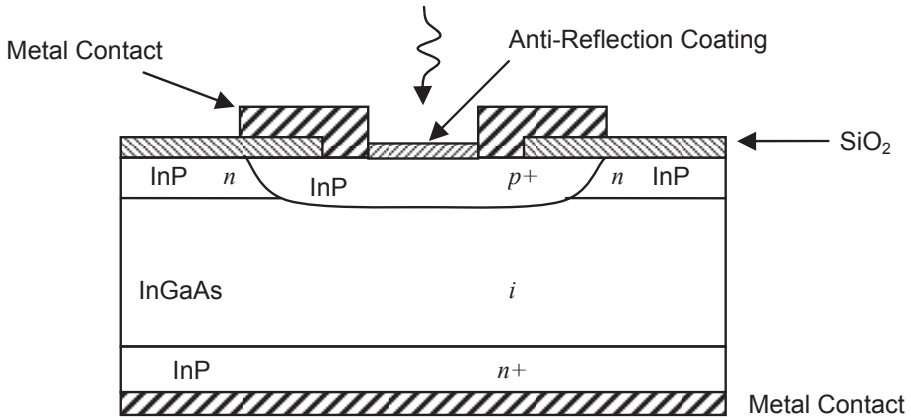

Fig. 6.10. Structure of a practical InGaAs PIN diode

The diode itself consists of a lightly doped, almost intrinsic InGaAs surrounded by two *n*-doped InP layers. On the top side, an additional diffused *p* layer makes up the *p* side of the junction. The top InP layer has a wider bandgap and therefore allows photons to pass through and reach the internal InGaAs layer. Therefore, the wavelength response of the detector is determined by the intrinsic layer, as photons are mostly absorbed in this region. Metal contacts on both sides provide the electrical connections to the junction.

6.5.3 APD structures

A practical APD structure is in many ways similar to a PIN diode. Figure 6.11 shows an example of an InGaAs APD structure similar to the PIN structure of Fig. 6.10. The difference is the addition of a *p* layer between the intrinsic region and the bottom *n* contact. When the APD is reverse biased, most of the voltage is

dropped across the *n*+*p* junction. Because of the higher level of doping in the *n*+ layer, the depletion region extends into the *p* layer, and with enough reverse bias reaches the intrinsic region. This is the basis of the "reach-through" APD structure that we discussed previously. Like the PIN structure, the InP has a wider bandgap and is almost transparent for the long wavelengths that pass through it and are absorbed by the intrinsic InGaAs layer. These photons generate the primary carriers. However, the avalanche gain occurs in the *p*-doped InP layer, which is where the electric field is much higher because of the reverse voltage drop in this layer.

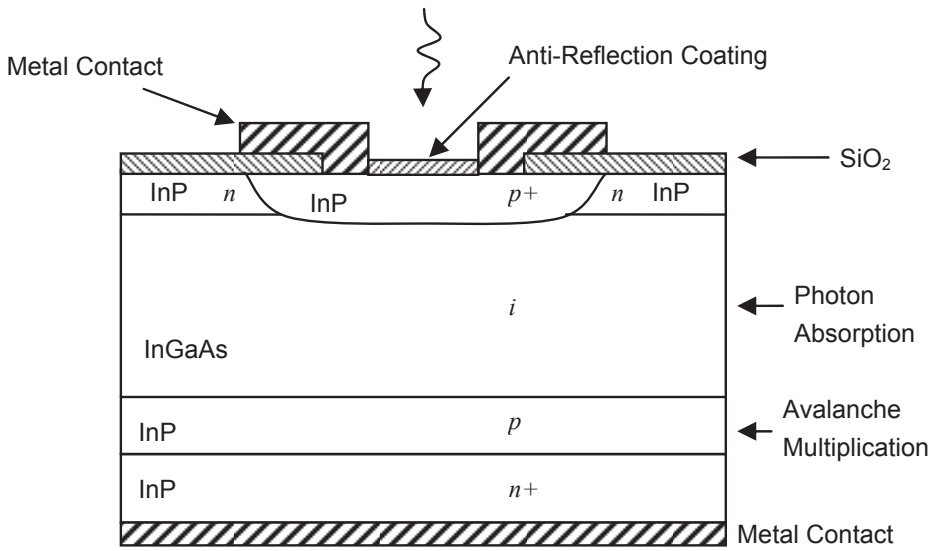

Fig. 6.11. Structure of a practical InGaAs PIN diode

In effect, absorption and avalanche multiplication are separated in this structure. This is advantageous, because each layer can be optimized for a different purpose. For instance, the InP layer is more suitable for impact ionization, because as a result of its higher bandgap tunneling current is reduced, and therefore higher electric fields can be supported.

6.6 Summary

As the front end of optical receivers, photodetectors convert photons to electrons. Although a variety of devices can accomplish this function, for fiber optic communications almost exclusively PIN and APD diodes are used. These devices provide very good responsivity in the wavelengths used in fiber optics. Moreover, they are robust, cheap, small in size, and mechanically compatible with other optoelectronic devices and optical subassemblies.

A PIN diode consists of an intrinsic semiconductor layer sandwiched between *p*- and *n*-doped materials. Typically, the device is reverse biased as a photodetector. The photons are absorbed in the intrinsic region, as long as the photon's energy is larger than the bandgap of the material. Photons are absorbed in the intrinsic region and generate carrier pairs. Because of the electric field in the intrinsic region, these carriers are swept out of the junction, making up the output current of the device. This process is characterized by the quantum efficiency, defined as the number of generated electrons for every 100 photons absorbed. Related to this is the responsivity of a detector, defined as the ratio of photocurrent to the optical power. Responsivity is a parameter that can easily be measured and is a key factor in the performance of a receiver. Typical responsivities vary from 0.5 to 1 A/W.

An APD detector's primary advantage over a PIN detector is internal current gain, which can substantially improve the sensitivity of an optical receiver. In an APD, the junction is reverse biased at a much higher voltage. Therefore, photogenerated carriers are accelerated to very high speeds and, as a result of impact ionization, generate additional secondary carriers. These secondary carriers themselves accelerate to high speeds and generate more carriers. This is known as avalanche multiplication. In the way very high current gains can be achieved. Usually, the optimum biasing point for an APD is a few volts below the breakdown voltage of the junction.

Noise is a key parameter to consider in photodetectors. There are two primary mechanisms for noise: shot noise or quantum noise, and thermal or Johnson noise. While there are some ways to reduce the effects of thermal noise, shot noise is always present and sets a limit on the maximum achievable signal-to-noise ratio. Simple mathematical results allow for the study of the effects of these two noise sources.

The most common materials used in building PIN and APD detectors are Si, Ge, and InGaAs. These materials have different operating wavelengths. For long-wavelength applications, InGaAs provides an excellent choice both in terms of responsivity and in terms of optical bandwidth. Its operating wavelength spans from almost 1000–1700 nm and its responsivity peaks around the 1550 nm. Consequently, in practical PIN and APD structures it is common to use InGaAs as the absorption layer.

References

[1] H. Zenk, "Ionization by quantized electromagnetic fields: the photoelectric effect," Review in Mathematical Physics, Vol. 20, pp. 367–406, 2008

[2] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd Ed., John Wiley & Sons, Hoboken, NJ, 2007

[3] A. J. Seeds and K. J. Williams, "Microwave photonics," Journal of Lightwave Technology, Vol. 24, pp. 4628–4641, 2006

[4] G. Wang et al., "Highly reliable high performance waveguide-integrated InP/InGaAs pin photodiodes for 40 Gbit/s fibre-optical communication application," Electronics Letters, Vol. 39, pp. 1147–1149, 2003

[5] L. Y. Lin et al., "High-power high-speed photodetectors-design, analysis, and experimental demonstration," IEEE Transactions on Microwave Theory and Techniques, Vol. 45, pp. 1320–1331, 1997

[6] Y. J. Chiu et al., "GaAs-based, 1.55 μm high speed, high saturation power, low-temperature grown GaAs p-i-n photodetector," Electronics Letters, Vol. 34, pp. 1253–1255, 1998

[7] Y. J. Chiu et al., "Ultrafast (370GHz bandwidth) p-i-n traveling wave photodetector using low-temperature-grown GaAs," Applied Physics Letters, Vol. 71, pp. 2508–2510, 1997

[8] Y. G. Wey et al., "110-GHz GaInAs/InP double heterostructure PIN photodetectors," Journal of Lightwave Technology, Vol. 13, pp. 1490–1499, 1995

[9] A. R. Williams, A. L. Kellner, and P. K. L. Yu, "Dynamic range performance of a high speed, high saturation InGaAs/InP pin waveguide photodetector," Electronics Letters, Vol. 31, pp. 548–549, 1995

[10] J. E. Bowers, C. A. Burrus, and R. J. McCoy, "InGaAs PIN photodetectors with modulation response to millimeter wavelengths," Electronics Letters, Vol. 21, pp. 812–814, 1985

[11] A. Bandyopadhy and M. J. Deen, "Photodetector for optical fiber communications," in *Photodetectors and Fiber Optics*, Edited by H. S. Nalwa, pp. 307–368, Academic Press, New York, 2001

[12] M. Lazovic, P. Matavulj P, and J. Radunovic, "The few SPICE models of ultra fast P-i-N photodiode," Journal of Optoelectronics and Advanced Materials, Vol. 9, pp. 2445–2448, 2007

[13] G. Wang et al., "A time-delay equivalent-circuit model of ultrafast p-i-n photodiodes," IEEE Transactions on Microwave Theory and Techniques, Vol. 51, pp. 1227–1233, 2003

[14] J. J. Jou et al., "Time-delay circuit model of high-speed p-i-n photodiodes," IEEE Photonics Technology Letters, Vol.14, pp. 525–527, 2002

[15] K. J. Williams, R. D. Esman, and M Dagenais, "Nonlinearities in p-i-n microwave photodetectors," Journal of Lightwave Technology, Vol. 14, pp. 84–96, 1996

[16] T. F. Refaat, H. E. Elsayed-Ali, and R. J. DeYoung, "Drift-diffusion model for reach-through avalanche photodiodes," Optical Engineering, Vol. 40, pp. 1928–1935, 2001

[17] H. Ando, Y. Yamauchi, and N. Susa, "Reach-through type planar InGaAs/InP avalanche photodiode fabricated by continuous vapor phase epitaxy," IEEE Journal of Quantum Electronics, Vol. 20, pp. 256–264, 1984

[18] E. Garmire, "Sources, modulation, and detectors for fiber-optic communication systems," in *Fiber Optics Handbook*, Edited by M. Bass, pp. 4.1–4.78, McGraw-Hill, New York, 2002

[19] J. W. Parks et al., "Theoretical study of device sensitivity and gain saturation of separate absorption, grading, charge, and multiplication InP/InGaAs avalanche photodiodes," IEEE Transactions on Electron Devices," Vol. 43, pp. 2113–2121, 1996

[20] J. Campbell et al., "Recent advances in avalanche photodiodes," IEEE Journal of Quantum Electronics, Vol. 10, pp. 777–787, 2004

[21] R. J. McIntyre, "The distribution of gains in uniformly multiplying avalanche photodiodes: Theory," IEEE Transactions on Electron Devices, Vol. 19, pp. 703–713, 1972

[22] C. Yeh, *Applied Photonics*, Academic Press, New York, 1994