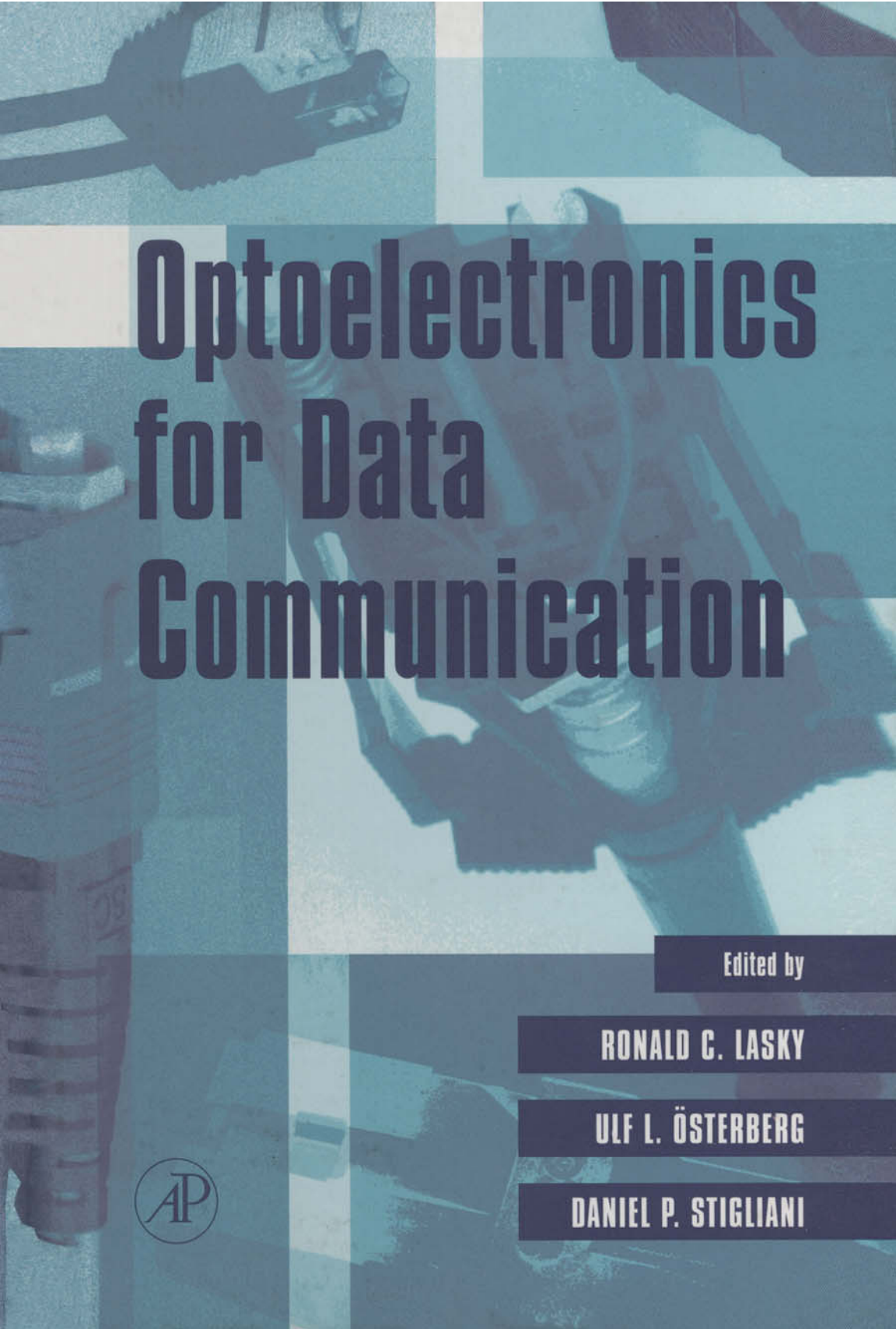


Optoelectronics for Data Communication

Edited by

RONALD C. LASKY

ULF L. ÖSTERBERG

DANIEL P. STIGLIANI

Optoelectronics for Data Communication

This Page Intentionally Left Blank

Optoelectronics for Data Communication

Edited by

Ronald C. Lasky
Universal Instruments Corporation
Binghamton, New York

Ulf L. Österberg
Dartmouth College
Hanover, New Hampshire

Daniel P. Stigliani
IBM Corporation
Poughkeepsie, New York

Academic Press
San Diego New York Boston London Sydney Tokyo Toronto

This book is printed on acid-free paper. ♾

Academic Press, Inc.
A Division of Harcourt Brace & Company
525 B Street, Suite 1900, San Diego, California 92101-4495

United Kingdom Edition published by
Academic Press Limited
24-28 Oval Road, London NW1 7DX

Library of Congress Cataloging-in-Publication Data

Optoelectronics for data communication / edited by Ronald Lasky, Ulf Lennart Österberg, Daniel Stigliani.
p. cm.
Includes bibliographical references and index.
ISBN 0-12-437160-4 (alk. paper)
1. Optical data processing. 2. Data transmission systems.
3. Optoelectronics I. Lasky, Ronald. II. Österberg, Ulf Lennart
III. Stigliani, Daniel.
TA1632.O73 1995
621.39'81--dc20 95-30715
CIP

PRINTED IN THE UNITED STATES OF AMERICA
95 96 97 98 99 00 EB 9 8 7 6 5 4 3 2 1

Contents

7 Bringing a Product to Market
Thomas B. Kellerman

8 The Future of Information Technology
George Cybenko

This Page Intentionally Left Blank

Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

A. Alex Behfar (77), IBM Corporation, Microelectronics Division, Hopewell Junction, New York 12533

Lawrence Brehm (165), IBM Corporation, Microelectronics Division, Endicott, New York 13760

Darrin P. Clement (165), Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755

George Cybenko (311), Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755

Casimer DeCusatis (219), IBM Corporation, Poughkeepsie, New York 12601

Thomas B. Kellerman (285), Optical Link Products, IBM Corporation, Endicott, New York 13760

Ronald C. Lasky (1, 127, 165), Universal Instruments Corporation, Binghamton, New York 13902

Janet L. Mackay (77), IBM Corporation, Microelectronics Division, Essex Junction, Vermont 05452

Ulf L. Österberg (1, 7), Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755

Carolyn Paddock (127), IBM Corporation, Essex Junction, Vermont 05452

Jerry Radcliffe (127). BroadBand Technologies, Research Triangle Park, North Carolina 27709

Daniel P. Stigliani (1), IBM Corporation, Poughkeepsie, New York 12601

James R. Webb (7), IBM Corporation, Microelectronics Division, Endicott, New York 13760

This Page Intentionally Left Blank

Preface

Optoelectronics for telecommunication is a well-established field. As a result, numerous textbooks have been written on this topic. With the advent of optoelectronics for data communication it became obvious that a new text addressing the specific needs of optical data communication was needed. This book is a result of this need. The idea for the book came after two of the editors ventured to form a seminar at Dartmouth College. With the encouragement of Academic Press, the outline for this course was used as an outline for this text. The book was written by industrial and academic workers in the field of optoelectronics for data communication from the middle of 1993 through the end of 1994. The book can be used as a reference for practicing engineers and scientists or as a text for an advanced undergraduate or graduate course.

R. C. Lasky

This Page Intentionally Left Blank

Chapter 1

Introduction

Ronald C. Lasky
Universal Instruments Corporation
Binghamton, New York

Ulf L. Österberg
Dartmouth College
Hanover, New Hampshire

Daniel J. Stigliani
IBM Corporation
Poughkeepsie, New York

1.1 Introduction

The advantages of lightwave technology for telecommunications became apparent in the early stages of its development. The potential of very high bandwidth transmission over long distances at a reasonable cost spurred investment by the telecommunications industry. In fact, in the early stages, all the fundamental research and application development was directed toward low cost, long distance applications. This application is fiber cost-intensive, as opposed to short distance applications, for which transceiver cost, size, and reliability are the most critical parametrics.

The tremendous advantages of the technology were also apparent to the information technology industry. With the maturity of optoelectronic technology in terms of level of integration, cost, and reliability, development activity began in the early 1980s to apply lightwave

Optoelectronics for Data Communication

technology to data interconnection. The first major application of lightwave technology in the information technology industry was Enterprise System Connection Architecture (ESCON) by IBM.[1] This application initially used multimode (MM) fiber and light emitting diode (LED) technology to achieve a data rate performance of 17 Mbyte/sec at a distance of 3 km. Subsequently, a 20-km distance option was offered by using single-mode (SM) fiber and 1300-nm lasers. ESCON is currently the primary interconnection to the IBM System/390 large processor family of products. ESCON was a significant improvement over the standard copper interconnection.

Also in the 1980s, the American National Standards Institute (ANSI) defined and adopted the first local area network (LAN) using lightwave technology as the primary physical medium. Fiber distributed data interface (FDDI) operates at 100 Mbit/sec and uses both MM/LED and SM/laser media. It also uses a dual counterrotating token ring embodiment with an optical bypass switch at each station.

This book is primarily concerned with the optoelectronic technology important for data communication between computers in the local and wide area networks. Because of optical fiber's small size and low cost, in addition to the properties just mentioned, optical fiber data communication promises to revolutionize the office environment. The two preceding applications are only the precursors of the use of this technology in the information technology industry. This book is intended to convey an appreciation of the unique requirements of the data communications application (versus telecommunications) and of how these requirements have been satisfied. Design concepts, trade-offs, and potential future applications are presented. However important computers may seem to us today, the history of using lightwave technology is, nevertheless, initimately related to the history of telecommunications and significant synergy exists within these two broad applications.

1.2 Historical Perspective

Over the last 25 years there has been tremendous development in the use of light waves for communication. It is possible to send, simultaneously, in one optical glass fiber, telephone conversations, data, and video signals, adding up to bandwidths in excess of 10^9 bps (this corresponds to sending an entire encyclopedia of information in 1 s). Because of the very low attenuation of optical fibers, it is possible to maintain these large bandwidths for hundreds of miles with a minimum of repeaters. Even though the bandwidths that can be utilized today with optical fibers are impressive, it should be remembered that optical fibers are inherently capable of providing bandwidths in excess of

[1] ESCON is a trademark of the International Business Machines Corporation.

tens of Tbit/s ($=10^{12}$ bps). To gain access to these large bandwidths, new techniques must be developed for multiplexing and demultiplexing at these high speeds.

The use of light for long distance communication (i.e. semaphore type communication) enjoyed an almost complete monopoly until the early nineteenth century when the telegraph was invented. The use of electronics for communication was further established toward the end of the nineteenth century with the invention of the telephone. Simultaneously, major theoretical and experimental breakthroughs were being reported by Maxwell and Hertz, dramatically improving our understanding of electromagnetic waves. Much of the first half of the twentieth century was devoted to the utilization of microwaves for radio and telephone communication. This work brought with it an increased understanding of modulation of signals, noise, filtering, multiplexing, and digitization of analog signals. With much of the research in communication concerned with basic properties of waves, it was realized early on that the shorter the wavelength of the electromagnetic waves (the higher the frequency), the more available bandwidth. This immediately meant that researchers once again started to look to light as a preferred source for transmitting information, because of its very high bandwidth.

The quest for an optical communication system can be divided into two separate research efforts: (1) to find a low cost and reliable light source, and (2) to find a low cost, reliable, and weakly attenuating medium for transmitting the lightwaves. Air would obviously be cheap but could not be considered particularly reliable or having low attenuation. The realization of both a reliable light source and a light-transmitting medium is a tribute to material scientists around the world. The ultimate problem in finding both the material for the light source and the light-transmitting medium was in reducing the number of defects and impurities in semiconductor and glass materials. In the case of glass, the amount of impurities has to be less that 1 ppb (parts per billion) for the glass to be practically useful from a signal attenuation perspective in fiber applications. Figure 1.1 illustrates the attenuation of fiber glass compared with window glass.

Two obvious candidates for the light source are the semiconductor laser and the LED, both originating from the pn-junction in doped semiconductors. Physically realized in the early 1960s by several independent groups, it was not until 10 years later that the lifetime of these lasers was long enough to make them economically viable in an optical communication system. The major technological innovation that allowed these lasers to last from a few hours in the laboratory, originally, to more than hundreds of thousands of hours in more realistic environments was the ability to prevent defects from migrating into the active region of the semiconductor laser and stoping it from lasing. Since then, a tremendous evolution of semiconductor lasers has taken place; it is now possible to make semiconductor lasers with spectral bandwidths less than 100 kHz or pulse lengths shorter than 1 ps (10^{-12} s). By

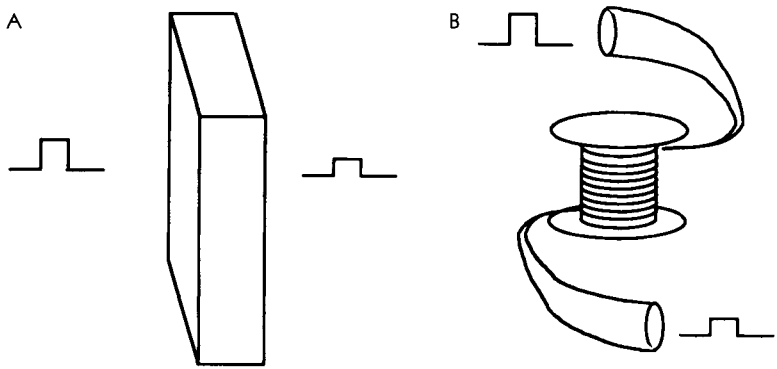


FIGURE 1.1 Purity of various glasses. A) Window glass: 50% attenuation after 3 cm. B) Fiber glass: 50% attenuation after 10 km.

using novel mode-locking methods, it is also possible to generate pulses with repetition rates in excess of 500 GHz. In addition to improvements of the light source itself, the field of nonlinear optics has produced several devices for manipulating amplitude, frequency, and phase of the light outside the laser cavity. This has led to the development of new multiplexing schemes to increase further the effective bandwidth of optical fiber systems.

1.3 Recent Efforts in Optical Data Communication

In 1987, IBM formed a small group in Endicott, New York, called the Optoelectronic Enterprise. This group's mission was to develop and manufacture world-class optoelectronic data communication technology at the component (cable, transceiver) level. By 1990, this group had developed and manufactured a 200-Mbit/s SM transceiver module that was offered as the ESCON XDF optical data communications feature in the IBM System/390 large processor family. The IBM ESCON Architecture was developed by Large Scale Computing Division (formerly Enterprise Systems) engineers in Poughkeepsie and Kingston, New York.

This book has been written, in part, by some of the men and women who were the leaders and designers in both the MM and SM ESCON technologies. The differences in technology between optical data communication and optical telecommunication drove this team to develop many new packaging, testing, and manufacturing concepts for the successful implementation of optical data communications. These differences, which at first were thought to be minor, cannot be overstated; it is because of these differences that this book was written. Hence we consider it a first of its kind.

The main differences between optical data communication and optical telecommunication are summarized here.

- Data communication will require shorter links, in general, and many of them will favor less expensive MM technology (contrary to the strong SM trends in telecommunication). Single-mode links will not dominate in data communication applications until required serial data rates exceed 2 Gbits/sec.
- In data communication, one user per link is usually the norm. The optimum use of the fiber bandwidth is not the driving factor, but rather the cost of electronic adapters and transducers. Usual link physical topologies are point-to-point, whereas the logic topology may be switched star, ring, and so on. Telecommunication has many users per link (to minimize cabling cost) and is most clearly illustrated by the trans-Atlantic fiber optic cable.
- The tremendous performance growth of data communication (processing capability doubles approximately every 18 months) has fostered a dynamic data processing environment. Customers are continually reconfiguring their data processing complexes and interconnection networks to take advantage of the improved capability. Light-wave technology must be robust, user friendly, and foolproof for reconfiguring by personnel unskilled in lightwave technology. For example, connection to a fiber optic transducer is done by a keyed duplex connector to assure proper mechanical mating.
- All optical links should be worldwide Class 1 optical radiation certified. This allows data processing personnel and cable installers to reconfigure optical links at the customer's premises without the expense associated with laser safety training, certification, and periodic verification. This has become the data communications industry norm.

None of the preceding issues are significant in telecommunications applications. In addition, most of the equipment in the telecommunications industry is owned by the telephone company, usually indicating a common supplier for optical technology. In data communications, the many intermatings of products from different vendors will require interoperability standards. All of these issues are addressed in this book.

The following list provides a brief synopsis of the remaining chapters.

- Chapter 2 discusses the fundamental aspects of light coupling into MM and SM fiber and fiber characteristics. The sections on cable design, and especially connector design issues, are new in that they are written mostly from the perspective of unique data communications environmental requirements.
- Chapter 3 covers material related to light sources and detectors. This chapter has an extensive discussion on semiconductor lasers and gives a qualitative comparison of the positive and negative features of the different common designs.

- Chapter 4 discusses the packaging issues related to the data communication transceiver module as well as the significant issues in the electrical, mechanical, thermal, and optical design of the transceiver module and link. Criteria are defined in the sub-component selection process from cable selection to encode/decode circuity. The process of building a link is described, and the chapter concludes with industry examples and future trends.
- Chapter 5 discusses a design descipline developed for the connector/module interface that was used in optimizing the coupling of light in IBM's ESCON SM and MM links. Mechanical issues such as the consistent ability to plug the connector and optical issues such as coupled optical power consistent with an acceptable link loss budget and obedience to laser safety requirements, all of which are considerable data communications challenges, are analyzed in this section.
- Chapter 6 starts with an overview of data processing systems and ties in the requirements for an optical link design. Optical link design philosophy, concepts, and implementation are reviewed with an emphasis on ESCON. The different present and emerging industry standards for optical communication are also reviewed, including FDDI, Fiber Channel Series (FCS), and asynchronous transfer mode (ATM)/synchronous optical network (SONET). The chapter concludes with a review of potential future intraconnection and interconnection data communications applications.
- Most technical texts do not discuss the business aspects of the technology. This text is an exception in that all of the technologists were involved in the "selling" of the technology and recognize the importance of a good business plan. Hence Chapter 7, entitled "Bringing an Idea to Market" discusses the business issues related to making a technical product that not only works but is marketable.
- This book ends with a look at the future of the new "Information Imperative". Chapter 8 reviews some of the issues in this new communications era and, in particular, reviews the work at Dartmouth College to develop the "library of the future."

Chapter 2

Fiber, Cable, and Coupling

James R. Webb
IBM Corporation
Microelectronics Division
Endicott, New York

Ulf L. Österberg
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire

2.1 Introduction

Optical glass fibers have been around for a little over 40 years. The early fibers were used with high-pressure arc lamps for medical applications, the so-called fiberscopes. These early fibers guided light in a glass core that was surrounded by air. They had attenuation losses on the order of 500 dB/km and they suffered from large dispersion. Obviously, these fibers were unacceptable as a long distance, high bandwidth communication medium. It was the concept of the core-cladding glass fiber in conjunction with improved fabrication methods of glass fibers at the end of the 1960s that laid the groundwork for the use of optical glass fibers in communication systems.

The core-cladding fiber is depicted in Figure 2.1. It consists of a core that is made out of fused silica doped with a small amount of germanium to give it a slightly higher refractive index than cladding that primarily consists of fused silica.

Optoelectronics for Data Communication

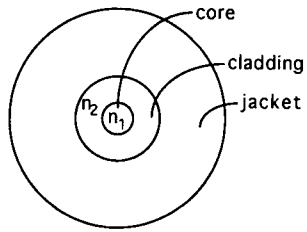


FIGURE 2.1. Core-cladding geometry in a standard optical fiber.

The refractive index difference between the core and the cladding ($n_{\text{core}} - n_{\text{cladding}}$) is in the order of 0.003 to 0.008 for communication fibers. These fibers are referred to as weakly guiding fibers [1], and have superior dispersion properties compared with the early air-cladding fibers. As a matter of fact, in an optical fiber, the dispersion can be made zero at wavelengths around 1.3 μm [2]. The improved fabrication methods resulted in fibers with an attenuation of 20 dB/km, this attenuation was reduced to less than 1 dB/km with a few years, and today's technology has reduced losses to as low as 0.14 dB/km at 1.55 μm [3].

It is the intent of this chapter to describe the basic ideas of how light propagates in an optical fiber and to familiarize the reader with some of the more important concepts in optical fiber wave propagation.

2.2 Ray and Electromagnetic Mode Theory

The basic equations that govern propagation of light in optical fibers are Maxwell's equations:

$$\begin{aligned} \nabla \cdot \mathbf{D} &= 0 \\ \nabla \cdot \mathbf{E} &= 0 \\ \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} \end{aligned} \tag{2.1}$$

Because glass is a dielectricum, there are no free charges or free currents in the medium, this is manifested through the divergence-free displacement field **D** and the omission of the conduction current for the magnetic field **H**. The **E** and **D** fields and the **B** and **H** fields are related, respectively, through the constitutive equations:

$$\begin{aligned} \mathbf{D} &= \varepsilon \mathbf{E} \\ \mathbf{B} &= \mu \mathbf{H} \end{aligned} \tag{2.2}$$

For most communication applications the light that is sent down the fiber is in the form of optical pulses, these are by necessity made up of several frequencies. Because we are working only with linear interactions, we know from basic system theory [4] that we can analyze each frequency component individually as it travels down the fiber and collect the different frequencies at the end using Fourier analysis. It is therefore sufficient for us to describe the propagation of one individual frequency component of the light, and this is most easily done using the formalism of phasors. We therefore write

$$\begin{aligned} \mathbf{E}(\mathbf{r}, t) &= Re[\mathbf{E}(\mathbf{r})e^{j\omega t}] \\ \mathbf{H}(\mathbf{r}, t) &= Re[\mathbf{H}(\mathbf{r})e^{j\omega t}] \end{aligned} \tag{2.3}$$

where the expression within the bracket is the phasor. Consequently, the time derivatives in Eq. (2.1) are simply transformed to multiplication with $j\omega t$, hence

$$\begin{aligned} \nabla \times \mathbf{E} &= -j\omega t\mathbf{B} \\ \nabla \times \mathbf{H} &= j\omega t\mathbf{D} \end{aligned} \tag{2.4}$$

So far, we have not made any approximations. To deal with the concept of rays propagating in our medium, however, we have to limit ourselves to light beam diameters much larger than the wavelength of light (a consequence of this is that the ray analysis is only practical for describing light propagation in multimode (MM) fibers). Nevertheless, the ray picture is important, as it helps to give an intuitive picture of how light can be confined within the core of an optical fiber. The equation for the ray can be written [5]

$$|\nabla S|^2 = n^2(\mathbf{r}) \tag{2.5}$$

where $n(\mathbf{r}) = \sqrt{\mu_r \varepsilon_r}$ and S is defined from

$$\mathbf{E}(\mathbf{r}) = \mathbf{e}(\mathbf{r})e^{-jk_0 S(\mathbf{r})} \tag{2.6}$$

where $k_0 = \omega\sqrt{\mu_0 \varepsilon_0}$, and k_0 is the free-space wave vector. From Eq. (2.5), S can be interpreted as refractive index *times* distance; so in the case of S being real, $k_0 S$ can be interpreted as a phase where the gradient of S gives the direction of the ray. S is referred to as the eikonal or optical path [6]. For example, it is shown that when absorption is negligible, **e** and **h** are perpendicular to each other and the direction of propagation, that is, the ray, behaves like a plane wave. Equation (2.5) can now be used to trace a ray through an inhomogeneous medium (i.e., the refractive index is a function of space), as shown in Figure 2.2.

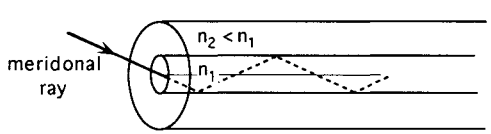


FIGURE 2.2 Ray tracing using the eikonal equation.

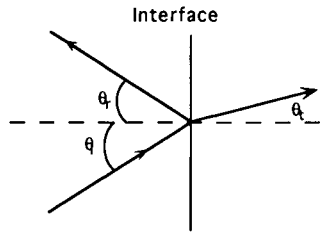


FIGURE 2.3 Reflection and transmission at a boundary.

If we follow a ray as it is incident on a boundary between two media described through the refractive indices n_1 and n_2, respectively (Fig. 2.3), we can, through the empirical law of Snell, describe what happens with the ray at the boundary. If we define θ_i, θ_r, and θ_t as the incoming, reflected, and transmitted rays, respectively, we have

$$\begin{aligned} \theta_i &= \theta_r \\ n_1 \sin\theta_i &= m_2 \sin\theta_t \qquad \text{(Snell's law)} \end{aligned} \tag{2.7}$$

From Eq. (2.7), we see that for incoming angles larger than

$$\theta_c > \sin^{-1}\frac{n_2}{n_1} \tag{2.8}$$

sine is not defined, if $n_2 < n_1$. This is the critical angle. Incoming rays at an angle larger than the critical angle are undergoing total internal reflection at the boundary.

For rays to be guided by the optical fiber, it is obvious that they have to undergo repeated total internal reflections in the fiber. This requirement leads to the concept of an acceptance angle, as shown in Figure 2.4. In other words, if an incoming ray within the cone is defined by the acceptance angle, it is guaranteed to undergo total internal reflection at the core-cladding boundary and therefore be guided by the fiber.

From the geometry of Figures 2.3 and 2.4, and from Eqs. (2.7) and (2.8), we find that [7]

$$n_{\text{air}} \sin\theta_a = \sqrt{n_1^2 - n_2^2} \tag{2.9}$$

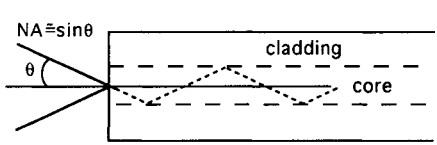


FIGURE 2.4 Acceptance angle for an optical fiber.

where n_1 and n_2 are the refractive indices of the core and cladding, respectively. The left-hand side of Eq. (2.9) is defined as the numerical aperture,

$$NA = n_{\text{air}} \sin \theta_a \tag{2.10}$$

The numerical aperture (NA) is an important parameter for characterizing an optical fiber. If we set $n_{\text{air}} \approx 1$ we find that for a single-mode (SM) fiber with typical values of $n_1 \approx 1.46$ and $\Delta n \approx 0.008$, the NA of the fiber is ≈ 0.18.

We can also define the relative refractive index difference, Δ, another important parameter used for characterizing optical fibers. Δ is defined as

$$\Delta = \frac{n_1^2 - n_2^2}{2n_1^2} \approx \frac{n_1 - n_2}{n_1} \tag{2.11}$$

where the last approximation is good for fibers that are weakly guiding, that is, $\Delta \ll 1$. Combining Eqs. (2.10) and (2.11) we obtain

$$NA \approx n_1 \sqrt{2\Delta} \tag{2.12}$$

The dimensionless parameters Δ and NA are extremely useful in comparing different fiber characteristics; in particular, it should be noted that Δ and NA do not depend on the radius of the fiber. This is a fact of ray approximation, that is, the wavelength is much smaller than the fiber dimensions. This approximation, however, breaks down as the fiber diameter gets smaller, as it does for a SM fiber. In this case, it is necessary to use the wave description of light to analyze completely the properties of the optical fiber.

To derive the equation that determines how electromagnetic waves are propagating in an arbitrary medium, it is necessary to return to Maxwell's Eqs. (2.1) and (2.2). From these, the wave equation can be written in Cartesian coordinates as

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + [n^2 k^2 - \beta^2] E_z = 0 \tag{2.13}$$

where E_z is the electric field component in the z-direction, nk is the phase constant for a plane wave (in an infinite medium), and β is the phase constant for the wave in the optical fiber. A few things should be noted about Eq. (2.13): (1) E_z is a function of both x and y, the transverse components, and so is the refractive index n; (2) an identical equation to Eq. (2.13) can be written for the H_z component; (3) solving for the electric and magnetic z-components of the field is sufficient to determine uniquely all components, that is, E_x, E_y, H_x, and H_y, of the electric and magnetic fields (this is from the curl relationships in Eq. (2.1) [6]); and (4) $[n^2 k^2 - \beta^2]$ is sometimes referred to as the transverse phase constant. Because the geometry of the optical fiber is a cylinder, it is easier to solve the wave equation in cylindrical coordinates.

Equation (2.13) is then transformed into

$$\frac{\partial^2 E_z}{\partial r^2} + \frac{1}{r}\frac{\partial E_z}{\partial r} + \frac{1}{r^2}\frac{\partial^2 E_z}{\partial \varphi^2} + \beta_t^2 E_z = 0 \tag{2.14}$$

By assuming that the refractive index is only a function of the radial variable r and that the function E_z can be separated into $E_z = R_z(r)\Theta_z(\varphi)$, one finds that [6]

$$\Theta_z(\varphi) = C_n \cos(n\varphi) + D_n \sin(n\varphi) \tag{2.15}$$

$$R_z(r) = \begin{cases} A_n J_n(\beta_t r) + B_n N_n(\beta_t r) & \text{(core)} \\ \tilde{A}_n K_n(|\beta_t| r) + \tilde{B}_n I_n(|\beta_t| r) & \text{(cladding)} \end{cases} \tag{2.16}$$

where A_n, B_n, C_n, D_n, $\tilde{A}_n$, and $\tilde{B}_n$ are arbitrary constants; J_n, N_n, K_n, and I_n are the nth order Bessel functions of the first and second kinds and the nth order modified Bessel functions of the first and second kinds, respectively; and n is an integer that corresponds to the azimuthal mode number. For a core with a uniform refractive index, the functions N_n and I_n can be neglected [7].

The solution in the core corresponds to β_t being real, and the solution in the cladding corresponds to β_t being imaginary. The complete solution of the wave equation is now a linear combination of the functions described in Eqs. (2.15) and (2.16). In an optical fiber, these combinations make up the TE, TM, HE, and EH modes. Fortunately, in a weakly guiding optical fiber, many of these modes are degenerate, making it possible to use the much simpler LP mode designation [6, 7]. When the refractive index difference between the core and the cladding is small, the longitudinal component (z-component) becomes negligible and the transverse components of the fields become dominant, hence the name *linearly polarized modes* (LP modes). The radial dependence of the fundamental mode (LP_{01}) is shown in Figure 2.5; there is complete symmetry in φ for this mode.

The arguments in the Bessel functions are commonly written as (Ur/a) and (Wr/a), where

$$\begin{aligned} U &= a\sqrt{n_1^2 k^2 - \beta^2} \\ W &= a\sqrt{\beta^2 - n_2^2 k^2} \end{aligned} \tag{2.17}$$

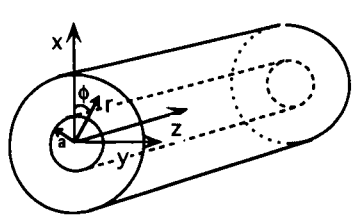


FIGURE 2.5 Radial dependence of the LP_{01} mode.

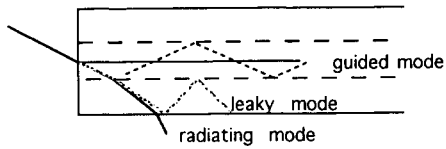


FIGURE 2.6 Guiding and radiating modes in an optical fiber.

and a is the radius of the core of the optical fiber. Therefore, Eq. (2.16) can be rewritten as

$$R_z(r) = \begin{cases} A_n J_n\left(U\dfrac{r}{a}\right) & r \leq a \\ \tilde{A}_n K_n\left(W\dfrac{r}{a}\right) & r \geq a \end{cases} \tag{2.18}$$

With this notation, it can be seen that both U and W have to be real for the mode to be guided. This corresponds to the following criteria for the phase constant β,

$$n_2 k < \beta < n_1 k \tag{2.19}$$

$W = 0$ corresponds to the cutoff for the mode and when W is imaginary, the mode is said to be radiating or leaky, as shown in Figure 2.6.

Using U and W, we can define another important dimensionless quantity to characterize an optical fiber, namely, the V number. V is defined as

$$V = \sqrt{U^2 + W^2} = ka\sqrt{n_1^2 - n_2^2} = \frac{2\pi}{\lambda} a(NA) = \frac{2\pi}{\lambda} a n_1 \sqrt{2\Delta} \tag{2.20}$$

The V number is referred to as the *normalized frequency*. For large V numbers, the number of modes that can propagate in the fiber is approximately $V^2/2$.

The Bessel functions describing the radial dependence of the modes in an optical fiber are cumbersome to work with. Because it is necessary to have a functional dependence of the radial mode field distribution for many different calculations, for example, to calculate the coupling efficiency into the optical fiber, it is fortunate that the LP_{01} mode can be approximated with a Gaussian function (Fig. 2.7),

$$R(r) = A_0 e^{-(r/\omega_g)^2} \tag{2.21}$$

where $2\omega_g$ is the mode field diameter of the fundamental mode [8].

To determine which modes will propagate in the optical fiber, we have to determine the constants A, B, and so on, in Eq. (2.16). This is done through the boundary conditions that the tangential components (z-components) of the electric and magnetic fields have to obey at the core-cladding interface

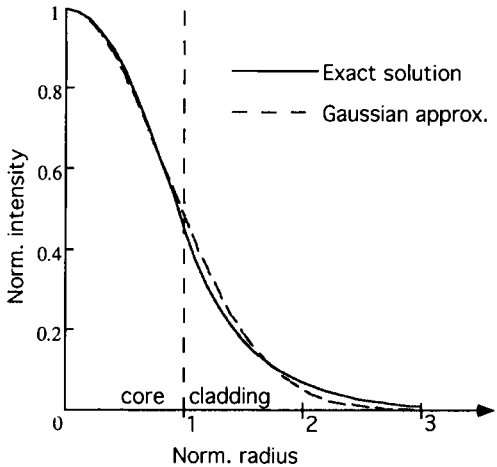


FIGURE 2.7 Comparison between exact and Gaussian approximation of the radial mode field distribution for the lowest order mode.

($r = a$). In principle, we would have to solve four eigenvalue equations, one for each of the possible modes in an optical fiber (TE, TM, EH, and HE), but because we limit ourselves to optical fibers that are important for data and telecommunication, that is, weakly guiding fibers, these eigenvalue equations converge to one equation, which then determines the phase constants β_{lm} for the LP modes. The eigenvalue equation is [7]

$$U\frac{J_{n+1}(U)}{J_n(U)} = \pm W\frac{K_{n+1}(W)}{K_n(W)} \tag{2.22}$$

The number of solutions that exist to this equation for a given optical fiber is determined by the V number. Hence, by plotting the various Bessel functions of the first kind ($J_n(V)$), we obtain from the zero crossings the cutoff frequencies for the various LP modes, as shown in Figure 2.8. Remember that at cutoff, $W = 0$, and therefore $U = V$.

A very useful dimensionless parameter is the normalized propagation constant b, it is defined from

$$b = 1 - \frac{U^2}{V^2} = \frac{(\beta^2/k^2) - n_2^2}{n_1^2 - n_2^2} \tag{2.23}$$

Because of the restrictions for a guided mode, Eq. (2.19), b must lie between 0 and 1.

The brief theory that has just been outlined is only valid for step-index fibers (uniform core fibers; Fig. 2.9a). The step-index fiber can either be SM (allowing only one LP mode to propagate) or MM (allowing several LP

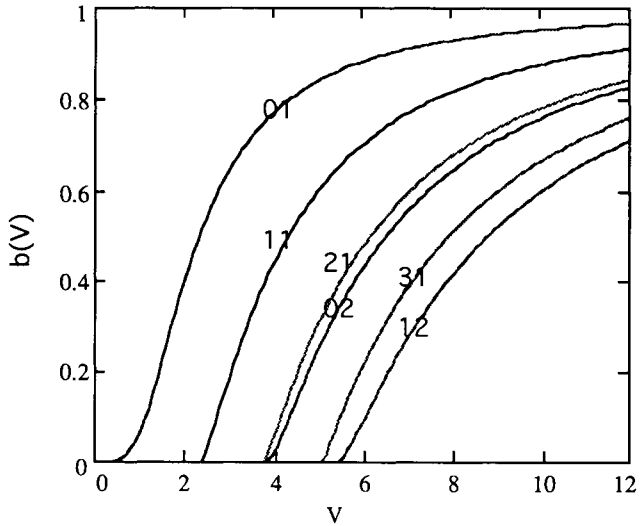


FIGURE 2.8 Cutoff frequencies for the lowest order LP modes.

modes to propagate). Because of modal dispersion (see section 2.4.4), MM fibers are, however, most commonly gradient-index fibers (Fig. 2.9b).

The index variation for a gradient-index fiber can be written as [7]

$$n(r) = \begin{cases} n_1\sqrt{1 - 2\Delta(r/a)^{\alpha}} & \text{core} \\ n_1\sqrt{1 - 2\Delta} = n_2 & \text{cladding} \end{cases} \tag{2.24}$$

where α is the profile parameter and its value is ≈ 2, giving the gradient index fiber a parabolic refractive index profile in the core. Except when $\alpha = 2$ and the refractive index in the core is allowed to, there is an analytical solution of the wave equation for a non-uniform refractive index profile in the core [6]. So, for practical purposes, it is necessary to solve the wave equation

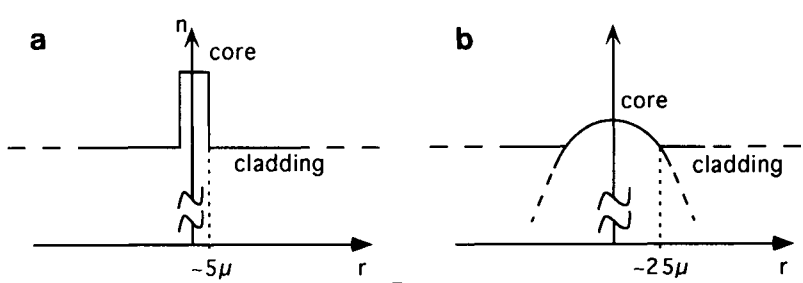


FIGURE 2.9 (a) Step-index fiber. (b) Gradient-index fiber.

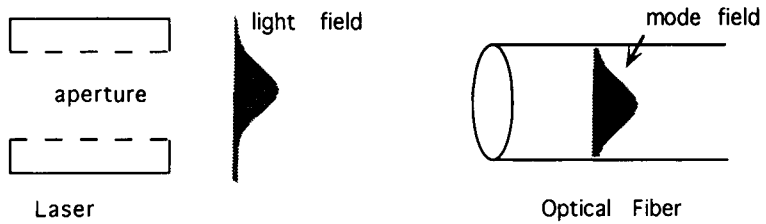


FIGURE 2.10 Coupling of light into an optical fiber.

numerically. An excellent treatment of how to solve the wave equation for gradient-index fibers has been given in Reference 6. For a given V number, a step-index fiber allows twice as many modes as a gradient-index fiber.

2.3 Coupling of Light into an Optical Fiber

In any fiber-optic system the coupling of light into an optical fiber is of fundamental importance. The source of the light can be either another fiber or a light source such as a semiconductor laser or light emitting diode (LED). Regardless of what the source of the light might be, the recipe for efficient coupling is to match the light field from the source with that of the mode field in the fiber, as shown in Figure 2.10.

In accordance with this, we define the coupling efficiency η as [8]

$$\eta = \frac{P_{lm}}{P_i} \tag{2.25}$$

where P_{lm} is the power in the mode LP_{lm} in the fiber and P_i is the power in the incoming light from the source. Rather than expressing the coupling efficiency in terms of percentage, as suggested by Eq. (2.25), it is more common to describe coupling losses in units of decibels a [8]:

$$a = -10 \cdot \log \eta \tag{2.26}$$

To understand what parameters are important for efficient coupling of light into an optical fiber, it is necessary to have a more quantitative definition than Eq. (2.24). It can be shown [8] that for LP modes, we can write the coupling efficiency as

$$\eta = \left| \frac{n_2}{Z_0} \int_{A_\infty} \mathbf{E}_i \bullet \mathbf{E}_{lm}^{*} \, dA \right|^2 \tag{2.27}$$

where n_2 is the refractive index of the cladding, Z_0 is the free-space wave impedance, and $\mathbf{E}_i$ and $\mathbf{E}_{lm}$ are the electric field amplitudes for the incoming light and for the light propagating in mode LP_{lm} in the fiber, respectively. The fields in Eq. (2.7) are just the transverse fields, because in the LP mode approximation the longitudinal component is neglected.

For SM fibers, $\mathbf{E}_{lm}$ can be approximated with a Gaussian function (as discussed in section 2.2),

$$
\begin{aligned}
E_\xi &= \sqrt{\frac{2Z_0P}{\pi w^2}}\exp\left(-\frac{\tau\rho^2}{2}\right)\exp(-j\psi) \\
\tau &= \frac{1}{w^2} + j\frac{k}{R} \\
\psi &= k\zeta - \tan^{-1}\left(\frac{kw^2}{R}\right) \\
w &= s_1\sqrt{1 + \left(\frac{\zeta}{ks_1^2}\right)^2} \\
R &= \zeta\left(1 + \left(\frac{ks_1^2}{\zeta}\right)^2\right)
\end{aligned}
\tag{2.28}
$$

where P is the total power carried by the transmitting beam, ρ is the fiber radial component, and k is the free-space wave number. In a similar fashion, we can write the field from the source as

$$
E_\xi = \sqrt{\frac{2Z_0P}{\pi w^2}}\exp\left(-\frac{\tau\rho^2}{2}\right)\exp(-j\psi) \tag{2.29}
$$

The geometry used is shown in Figure 2.11.

It is assumed that the light is propagating in the ζ direction and that it is polarized in the transverse ξ direction. An angular misalignment θ in the plane is also incorporated.

We can immediately distinguish four different cases (Fig. 2.12) in which a mismatch will reduce the coupling efficiency η [9, 10].

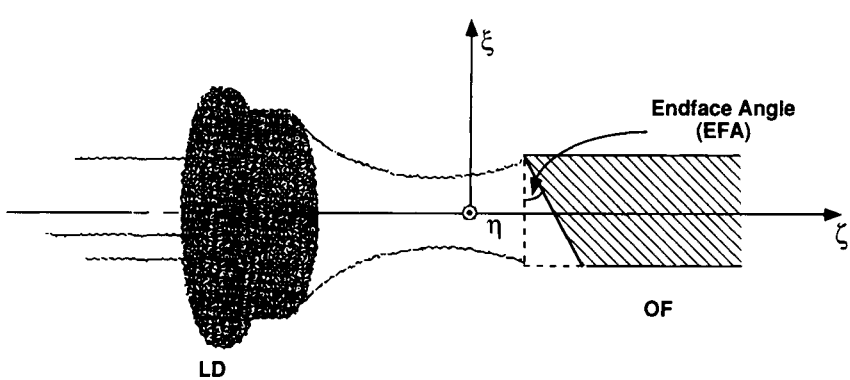


FIGURE 2.11 Coupling geometry.

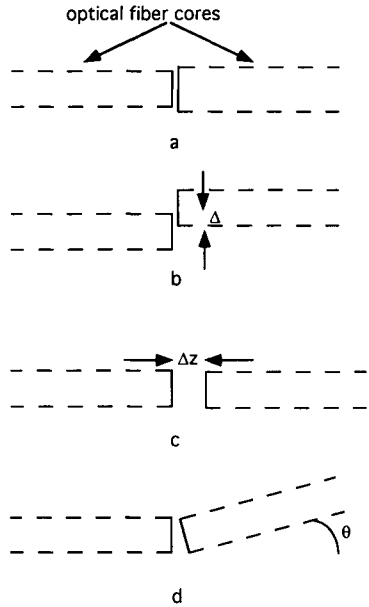


FIGURE 2.12 Different coupling cases.

Case 1: Spot-size mismatch $s_1 \neq s_2$

$$\text{Coupling efficiency: } \eta = \left(\frac{2s_1 s_2}{s_1^2 + s_2^2}\right)^2$$

Case 2: Matched spot sizes but a transverse offset Δ

$$\text{Coupling efficiency: } \eta = \exp\left(-\left(\frac{\Delta}{s_2}\right)^2\right)$$

Case 3: Matched spot sizes and fiber axes but a longitudinal offset Δz

$$\text{Coupling efficiency: } \eta = \frac{1}{1 + (\Delta z/2z_0)^2}$$

where z_0 is the Rayleigh range.

Case 4: Matched spot sizes and no offsets but an angular misalignment θ

$$\text{Coupling efficiency: } \eta = \exp\left(-\left(\frac{\theta}{\theta_0}\right)^2\right)$$

In the more general case, when two or more of these mismatches occur simultaneously, the formula for the coupling efficiency becomes much more involved. A full treatment of this has been given by Nemoto and Makimoto [11].

The preceding treatment is for SM step-index fibers illuminated with coherent light (e.g., a semiconductor laser). Coupling of coherent light into a MM graded-index fiber is much more complicated to analyze and the reader is referred to the work of Miller and Mettler [12].

For incoherent light (e.g., a Lambertian source), we can relatively easily treat the coupling of light into both a step-index fiber and a graded-index fiber. We start by describing the light source through its radiance B. The units for B is watts per area and steradians. Mathematically, B can be defined as [8]

$$B = \frac{\Delta P}{\Delta A \cdot \Delta\Omega \cdot \cos\theta} \tag{2.30}$$

where ΔP is the power emitted from an area ΔA on the source into a solid angle $\Delta\Omega$. θ is the angle between the direction of observation and the normal to the surface of the emitter, as shown in Figure 2.13.

In the case of a true Lambertian source, B does not depend on the position of the emitting element and furthermore B is not (to a good approximation) a function of the direction of the emitted light. So, for a Lambertian source, the total emitted power can be calculated using Eq. (2.30):

$$P_{ta} = \int_{\Omega}\int_{A} B \cdot \cos\theta \, dA \, d\Omega = \pi BA \tag{2.31}$$

From the point of view of coupling into an optical fiber, the total power calculated in Eq. (2.31) is of minor importance because the optical fiber can only accept light in a small cone determined by the difference in refractive index between the core and the cladding, Eq. (2.9). So, as a more realistic example of coupling incoherent light into an optical fiber, let us examine the case of coupling an LED into a step-index fiber. We assume the emitter area (diode) to be large relative to the receiving area (fiber) (see Fig. 2.14).

We assume that the emitter has circular symmetry, and therefore we can write the power emitted from one ring as [13]

$$dP = B \cdot 2\pi\rho \, d\rho \cdot \underbrace{\cos\theta'}_{\text{projection}} \cdot \frac{dA' \cdot \overbrace{\cos\theta'}^{\text{projection}}}{(R')^2} \tag{2.32}$$

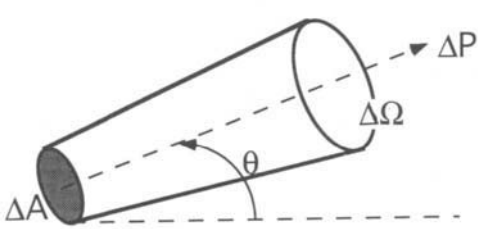


FIGURE 2.13 Lambertian source.

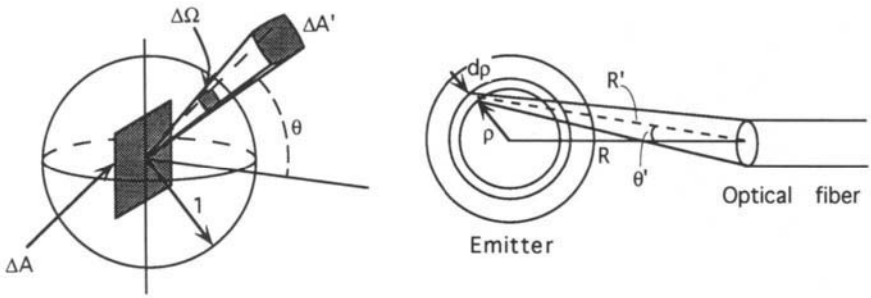


FIGURE 2.14 Coupling of incoherent light.

where ρ, $d\rho$, R', and θ' are defined in Figure 2.14. The total power accepted by the optical fiber is then obtained by integrating Eq. (2.31) from $\theta' = 0$ to $\theta' = \theta$, where θ is the acceptance angle of the fiber, i.e.

$$P = \pi \cdot B \cdot dA_{\text{det}} \cdot \sin^2\theta = \pi \cdot B \cdot dA_{\text{det}} \cdot (NA)^2 \tag{2.33}$$

A similar analysis can be done for a graded-index fiber [14].

If a lens is used in between the emitter and fiber, some modifications to the previous formulas have to be done. The geometry is given in Figure 2.15.

If we assume the refractive index to be the same on either side of the lens, $n = n'$, we must have $B = B'$. In other words, a lens cannot increase the amount of radiance at the receiver. What the lens can do for us, however, is to match the output angle of an emitter to the acceptance angle of the receiver (Fig. 2.15). In the case of coupling a LED into a step-index fiber, as in the preceding example, the power coupled into the fiber is multiplied with a factor M. This is the lens magnification factor and it is the ratio between the diameter of the emitter and the diameter of the receiver (d_{rec}/d_{em}).

All of the preceding formulas need to be corrected for reflection losses. If there is a mismatch in the refractive index between the surrounding medium and the core of the optical fiber, the coupled power is reduced with a factor R, where

$$R = \left(\frac{n_{\text{core}} - n_{\text{surround}}}{n_{\text{core}} + n_{\text{surround}}}\right)^2 \tag{2.34}$$

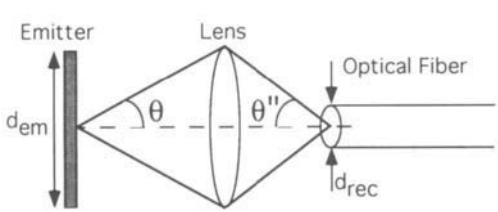


FIGURE 2.15 Lens for improving coupling efficiency.

This reflection factor occurs at every interface between the emitter and the receiving fiber.

2.4 Optical Fiber Characteristics

2.4.1 Optical Fiber Materials

Contributing to the breakthrough for low-loss fibers was the development of a successful manufacturing process. A precise, repeatable process was required that could manufacture optical fibers with extremely low contaminant levels and in high volume. Low-loss fiber was achieved when contaminants were reduced to levels measured in parts per billion (ppb). Even the best manufacturing processes, however, would not have led to low-loss fibers if optimal materials had not been selected.

A suitable fiber material must meet several criteria.

1. Highly transparent
2. Low cost
3. Highly manufacturable
4. Reliable

Optical fibers consist mostly of fused silica glass (SiO_2). One reason that silica glass is used is that silica is highly transparent, more transparent than other materials in their most pure form. Table 2.1 shows scattering loss data for several representative glass materials.

From the table, one can conclude that fused silica has the lowest scattering loss of common glasses. Scattering loss is explained later in the text, but in general terms, it is the largest component of loss inherent in glass materials. Another reason to choose silica is that fibers can be produced at a low cost. There are two reasons for this: first, silica is made from raw materials used in large volume by the semiconductor industry, and second, processes have been developed that produce optical fibers in large volume inexpensively.

As mentioned earlier, in order for the light to be guided, fibers are made with a varying refractive index through their cross-section. Step and gradient variations are the most common. The refractive index is changed by adding

TABLE 2.1 Scattering loss for several representative glass materials [15]

	633 nm	800 nm	1060 nm
Fused silica	4.8 dB/km	1.9	0.6
Soda lime	8.5	3.3	1.1
Borosilicate crown	7.7	3.0	1.0
Lead silicate	47.5	18.6	6.0

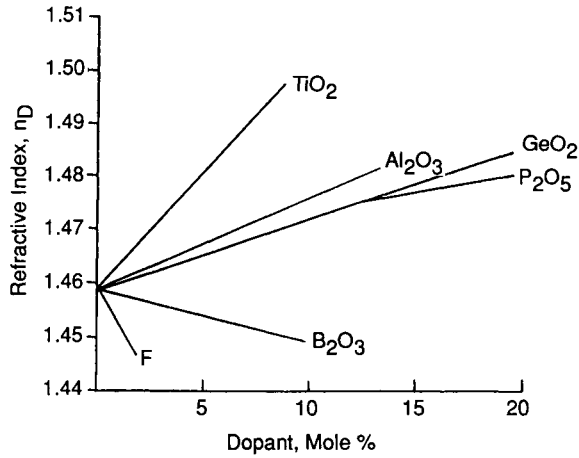


FIGURE 2.16 Influence of different dopant mole percentages on refractive index of silica glass. Reprinted from [16], p. 21 by courtesy of Marcel Dekker, Inc.

materials called dopants to silica. Dopants are materials that increase or decrease the refractive index of silica. GeO_2 and P_2O_5 are dopants that increase the refractive index of silica. B_2O_3 and Fluorine reduce the refractive index. The influence of different dopant mole percentages on the refractive index of silica is shown in Figure 2.16.

2.4.2 Optical Fiber Manufacturing

A precise process is required to make highly pure, doped silica fibers. Fibers must be very pure to limit the attenuation of light. The manufacturing of an optical fiber can, for simplicity, be broken down into two steps. First, a large blank of glass is prepared with a desired core/clad refractive index profile. Second, the fiber is made by drawing a filament from the heated blank. A viable blank production process must meet several criteria.

1. Low contamination
2. Highly manufacturable
3. Low inclusion of defects/highly homogeneous

A process that was used to produce blanks for the first low-loss fibers, a process which is still being used, is the chemical vapor deposition (CVD) method. CVD was developed by the semiconductor industry for production of extremely pure materials. In the CVD process, chemicals in the vapor phase are mixed and reacted while flowing. This results in highly pure reactants. Purity is achievable because the reacting constituents are very pure vapors that started as liquids. In a liquid mixture each component has a different vapor pressure. When a mixture is heated to and held at a constant

temperature, only one component of the mixture vaporizes and all other components in the mixture—whether liquid, dissolved solid, or suspended solid—are left behind in the remaining liquid. If the process starts with a pure liquid, the result is an even more pure vapor. CVD reduces the concentration of contaminants to the requisite ppb level for low-loss optical fibers.

Several processes based on CVD have been developed.

- Outside vapor deposition (OVD)
- Modified chemical vapor deposition (MCVD)
- Plasma chemical vapor deposition (PCVD)
- Plasma chemical vapor deposition—modified (PMCVD)
- Vapor axial deposition (VAD)

The OVD method was primarily developed by Corning Glass Works and was the first process used to manufacture fibers with less than 20 dB/km attenuation. The process starts with liquid chemicals such as $SiCl_4$ or $GeCl_4$. The materials are available in purity levels measured in parts per million (ppm). These same raw materials are also used in the semiconductor industry, which makes for very low cost. The first step in the process is to vaporize the chemicals, then mix the vapors with O_2 and H_2. The flame product is fine, highly pure silica particles in the form of soot. The flame and soot stream is blown onto a starter rod. The rod is rotated and cycled transversely to build up a uniform structure layer by layer. The glass is deposited in a partially sintered state and the resulting structure is called a preform. Glass preform composition, refractive index, and core/clad dimensions are varied by changing the mixture of constituents reacting in the flame. After a preform is completed, it is removed from its target rod and sintered at approximately 1500°C in an intert atmosphere to produce a highly homogeneous glass blank. During this process, the center hole collapses.

The MCVD process uses the same concept of purifying chemicals and reacting them in a flame to produce soot. Instead of depositing soot onto a starter rod, gases flow and react inside a rotating silica tube and deposit on its inner surface. Heat applied outside the tube vitrifies deposited soot. The final step is heating the tube to collapse it into a solid rod. The other processes in the list are variations of CVD and each has its advantages and disadvantages. VAD is a continuous production process, whereas the other four are batch processes.

Manufacturing glass blanks with controlled composition is the first step in fiber production. The second step is making a fiber from the blank. Fibers are made by drawing a filament from a heated blank. The tip of a blank is locally heated to its softening point. Silica glasses have high softening temperatures: 2000°C or higher. A filament is created when the tip of the molten glass falls under the force of gravity. The resulting filament is fed through a drawing and winding apparatus that consists of pulleys and tensioners. The apparatus sustains filament continuity as the blank is slowly lowered in the zone oven

to soften more material. Fiber diameter is continuously monitored as it leaves the furnace. A closed-loop system adjusts draw speed, furnace temperature, and blank feed speed based on inputs from fiber diameter monitors. While a fiber is in the drawing and winding apparatus it is purposely given a tensile load to force the fiber to break if strength does not meet a specified level. This process of removing weak points in a fiber is called proof testing. If a fiber breaks during proof testing, drawing is recommenced. As part of the drawing process, the fiber is wound on reals for storage. A drawing process is illustrated in Figure 2.17. Optical properties of fiber samples are tested before storage or shipment.

A fiber that is freshly drawn is susceptible to surface damage that lowers fatigue and tensile strength. For protection, fibers are coated with a thin layer of polymer on the order of 125 to 500 μm thick immediately after they are drawn yet before they reach the winding apparatus. The coating is cured after it is applied by a continuous process of heat or ultraviolet (UV) light. A soft coating offers good protection from external mechanical factors such as crush

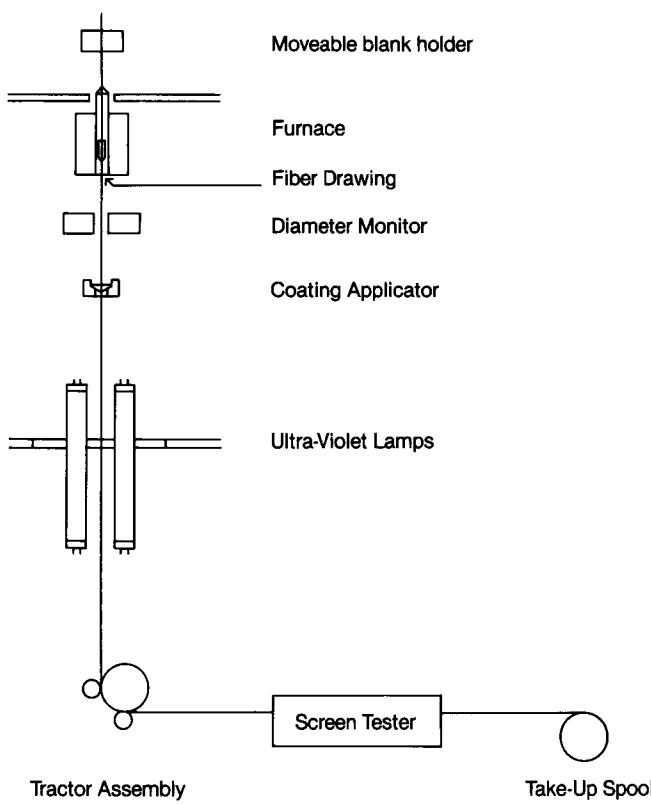


FIGURE 2.17 Fiber drawing process (Courtesy of Corning Incorporated).

loads and microbending. A coating material must be impervious to hydrogen and water, which are both detrimental to fiber performance, and must be quickly curable for practical use in continuous fiber-drawing processes. Silicone rubbers and UV-curable acrylates are two soft coatings that are used to coat fibers. During drawing, it is important to limit exposure of fibers to both airborne contaminants and sources of surface damage before the coating operation. Damage to glass blank surfaces must also be minimized. The surface of blanks can be fire polished to relieve and repair microscopic surface flaws.

2.4.3 Attenuation

When selecting an optical fiber for an application, trade-offs must be made between competing attributes. The following sections examine attenuation, dispersion, and mechanical characteristics of optical fibers with regard to trade-offs made for fiber selection. Since the time fiber-optic communication became commercially practical, in part from the development of "low-loss" fibers with losses around 20 dB/km, the goals of fiber research and development have included lower loss, higher bandwidth, and more robustness.

Signals transmitted by optical fibers lose strength like signals in any other transmission medium. The general shape of loss plotted against wavelength λ for optical fibers is illustrated in Figure 2.18. Power loss, or attenuation, is generally proportional to fiber length. The units of attenuation are decibels per kilometer.

Attenuation in fibers can be divided into two categories: instrinsic and extrinsic. Intrinsic losses are those losses that are inherent to both pure and doped silica. Intrinsic losses form the upper limit on fiber transparency and will exist even if all other forms of attentuation are eliminated. Extrinsic losses are induced by both design of fibers and manufacturing processes. Advances in fiber design and manufacture have been in part aimed at reducing the magnitude of loss from extrinsic mechanisms.

Intrinsic losses include UV absorption, infrared (IR) absorption, and scattering. Intrinsic loss occurs because optical quanta either cause electrons in a transmission medium to shift to higher energy levels or cause molecules to shift to different vibration states. First, consider UV absorption. Silica absorbs energy from light transmitted in the UV region. Absorption of a fiber doped by n% by weight GeO_2 can be generally expressed by the following formula [16]:

$$a_{uv}(\lambda) = a_0' \cdot n \cdot 10^9 \exp\left(\frac{4.63}{\lambda}\right) \tag{2.35}$$

where $a_{uv}(\lambda)$ is UV absorption at wavelength λ, n is percentage by weight of GeO_2 dopant, and a_0' is an experimentally obtained constant.

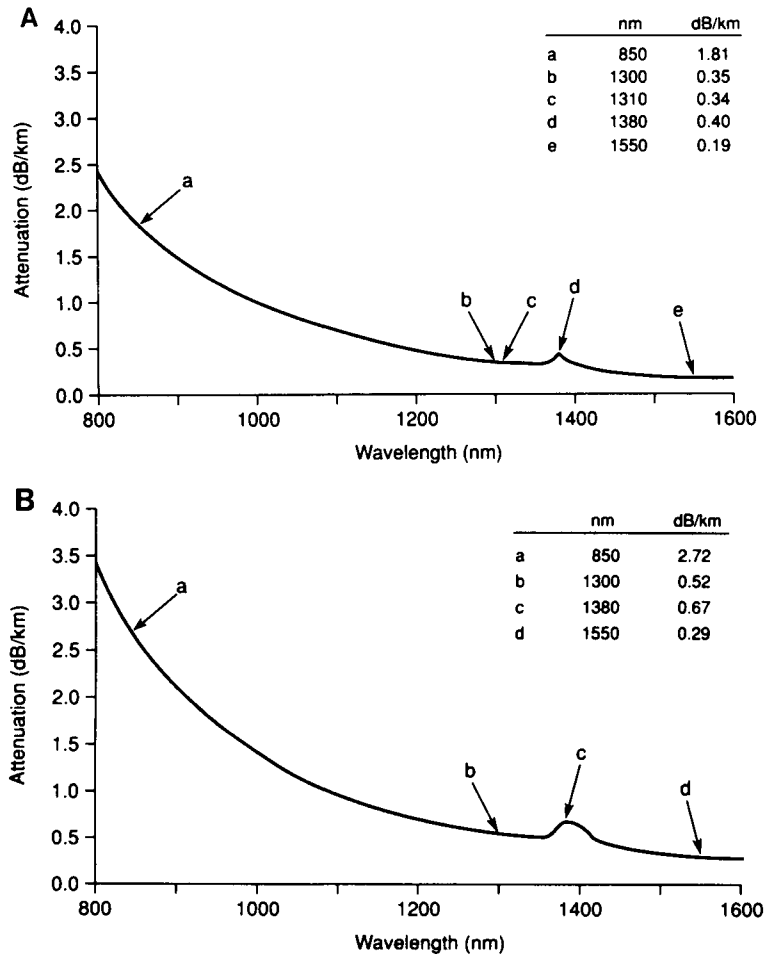


FIGURE 2.18 Transmission loss in optical fiber: (a) SM fiber; (b) graded-index 62.5/125 micrometer MM fiber. (Courtesy of Corning Incorporated.)

Ultraviolet absorption is caused by electronic transitions in the glass and increases exponentially with wavelength (the Urbach edge). Typical attenuation values are 0.02 dB/km at 1240 nm and 1 dB/km at 620 nm [17]. Ultraviolet absorption contributes relatively little to total loss of a silica-doped fiber above 600 nm wavelength. Below 600 nm, UV absorption is significant; therefore, UV absorption imposes a lower bound on transmitted wavelength of approximately 600 nm in silica fibers.

Infrared absorption, between 9 and 13 μm, occurs because photons interact with the strong Si–O vibrations in the glass. Peak absorption is

10^{10} dB/km, which occurs at approximately 9 μm. Harmonic overtones occur at shorter wavelengths in the useful range of data communication but are relatively insignificant. At 1770 nm, absorption is 1 dB/km. At 1550 nm, attenuation is typically 0.02 dB/km [18]. Infrared absorption imposes an upper bound of approximately 1800 nm on wavelength transmitted by silica fibers. UV and IR absorption lower and upper bounds leave a large window for the wavelengths between 780 and 1550 nm that are used in data communication applications.

Scattering, the third primary intrinsic loss mechanism, is composed of four types: Rayleigh, Brillouin, Mie, and Raman [16]. The last three contribute little to scattering loss compared with Rayleigh scattering. Rayleigh scattering is synonymous with molecular scattering. Molecules in optical transmission mediums form electric dipoles that oscillate when subjected to time-varying electromagnetic fields. The electrons emit electromagnetic waves almost instantly after they begin to vibrate. The reradiated light is emitted in almost all directions. In an optical fiber, light is emitted in directions that will not propagate in the waveguide; therefore, light is lost by light radiating. Strictly speaking, Rayleigh scattering is a loss mechanism, not an absorption mechanism. Rayleigh scattering can be expressed in terms of wavelength by the following equation [19]:

$$\alpha = \frac{A}{\lambda^4} \tag{2.36}$$

Coefficient A is a function of the refractive index, photoelastic coefficient, Boltzmann's constant, absolute temperature, and thermal compressibility. For silica fibers, A is related to dopant type and the core/cladding refractive index difference. For pure silica at 850, 1300, and 1550 nm, Rayleigh scattering loss is approximately 1.3, 0.3, and 0.1 dB/km, respectively. Total intrinsic loss, graphed as the sum of its constituents, is illustrated in Figure 2.19. Intrinsic loss is the theoretical minimum loss in optical fiber transmission and is on the order of 0.15 dB/km at 1550 nm.

Extrinsic losses are consequences of imperfections induced by both fiber manufacturing processes and design limitations. Primary mechanisms for extrinsic losses in silica optical fibers are contaminants, waveguide geometric imperfections, and bending.

Consider contamination first. Several compounds found in silica waveguides have a detrimental effect on transparency. Three of the primary contaminants that must be controlled are transition metal ions, hydroxyl ions, and hydrogen. Transition metal ions of copper, iron, nickel, vanadium, chromium, and manganese are present in fused silica and are difficult to remove completely. These ions have electronic absorption characteristics in the visible and near-infrared wavelengths. Metal ions are formed during manufacture of silica fibers and leave electron holes in the molecular structure. Table 2.2 shows attenuation for representative transition metals. The

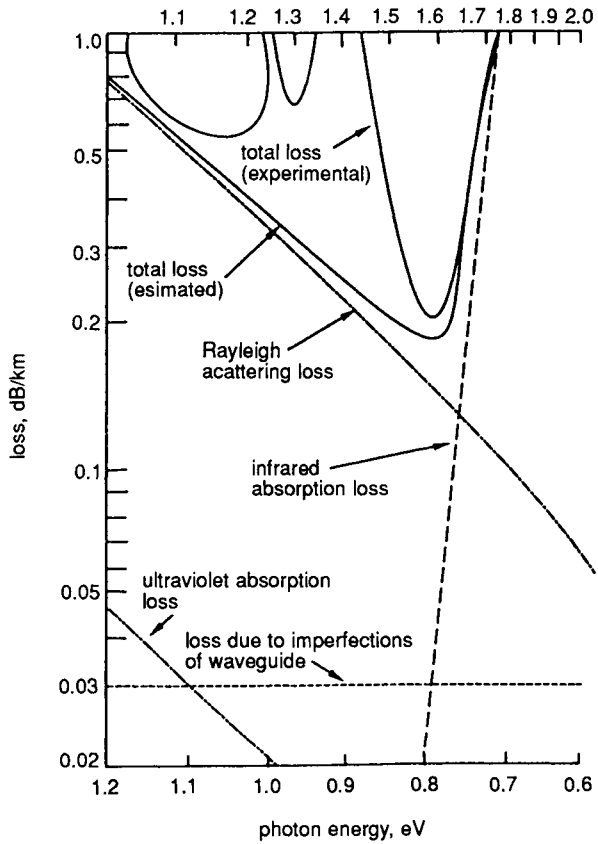


FIGURE 2.19 Total intrinsic loss and its constituents.

TABLE 2.2 Attenuation in silica glass fibers resulting from contamination by transition metal ions

Transition metal	Absorption peak (nm)	Concentration for 1-dB/km loss at absorption peak (ppb)	Concentration for 1-dB/m loss at 800 nm (ppb)
Cu^{2+}	800	0.45	0.45
Fe^{2+}	1100	0.40	0.75
Ni^{2+}	650	0.20	1.3
V^{3+}	475	0.90	1.8
Cr^{3+}	675	0.40	4.2
Mn^{3+}	500	0.90	90

data show that silica waveguides must be highly pure to achieve low attenuation. The concentration of metal ions must be reduced to as low as 4 ppb for a 20-dB/km loss and 0.2 ppb for a 1-dB/km loss.

Losses induced by the inclusion of hydroxyl ions (OH^-) from water cause the largest losses that result from contamination. Like metal ions, OH^- ions are difficult to remove. The loss mechanism for hydroxyl ion contamination is from photons interacting with the O–H vibration. Table 2.3 shows attenuation of 1 ppm OH^- in silica at the fundamental and overtone frequencies. The fundamental frequency is 2700 nm in the IR region. In the band of wavelengths used in data communication, significant harmonic overtones occur at wavelengths 720, 880, 950, 1240, and 1380 nm. Data communication systems should not be designed to operate at these wavelengths. The development of low-loss fibers has focused on minimizing OH^- ion concentration. Current production methods keep OH^- ion concentration below several parts per billion.

Hydrogen is the last of the three primary contaminants. It can diffuse into silica glass and induce attenuation. Ions formed during both high-temperature manufacturing processes and room-temperature diffusion vibrate at characteristic frequencies and attenuate signals by absorbing energy. A large loss occurs at 2.42 μm; smaller yet appreciable losses occur at several shorter wavelengths between 1 and 2.42 μm. To limit hydrogen diffusion, fiber and cable materials must be selected that do not release hydrogen and that are impermeable to it. UV-curable resins exhibit low H_2 generation and permeability, and are selected for coating fibers because of these properties.

Attenuation is induced by factors other than contamination during fiber and cable manufacture. Notable factors include geometric imperfections and bends in fibers. An example of a geometric imperfection is core diameter

TABLE 2.3 Attenuation due to hydroxyl ions at resonant frequencies and 1 ppm concentration

Wavelength (μm)	Loss per 1 ppm OH^- (dB/km)
2.72	6000.0
2.22	160.0
1.90	6.2
1.38	39.0
1.24	1.72
1.13	0.07
0.945	0.62
0.88	0.055
0.82	0.002
0.72	0.039
0.68	0.002

deviations from nominal. Points in a fiber with a diameter smaller than nominal will not propagate as many modes as a fiber with nominal size: higher order modes will radiate instead of propagate. Deviations in core diameter along a fiber's length will also cause shifts from guided modes to radiating modes. Present day manufacturing techniques typically control fiber core diameter and its deviation sufficiently to minimize losses from core geometry imperfections.

Bends in fibers that cause signal attenuation are categorized as macrobends and microbends. Microbends are microscopic deviations in a fiber from a nominal central axis of a fiber if it were straight or uniformly curved. Macrobends are visible curves in fibers. Bends may form radii of curvature that, when small enough, permit modes to either couple to higher order modes or radiate. In terms of light ray analysis, a bend in a fiber may increase the angle at which a ray intersects the core/cladding interface to one greater than the critical angle. When the angle of incidence is greater than the critical angle, total internal reflection does not occur and the ray will not propagate along the waveguide. Microbend losses in general vary inversely with Δ; therefore, fiber with large Δ reduce attenuation from microbends. For MM fiber only, reducing fiber core size and/or NA will reduce microbend losses, too.

2.4.4 Dispersion

As signal pulses traverse optical fibers they not only lose strength but also become distorted and "broadened." Broadening occurs because different parts of a pulse travel with different velocities. This concept is illustrated in Figure 2.20. Broadening alters both pulse amplitude and width. Broadening of signal pulses is generally called pulse dispersion and is expressed in terms of frequency or time. Dispersion limits the transmission bandwidth, or

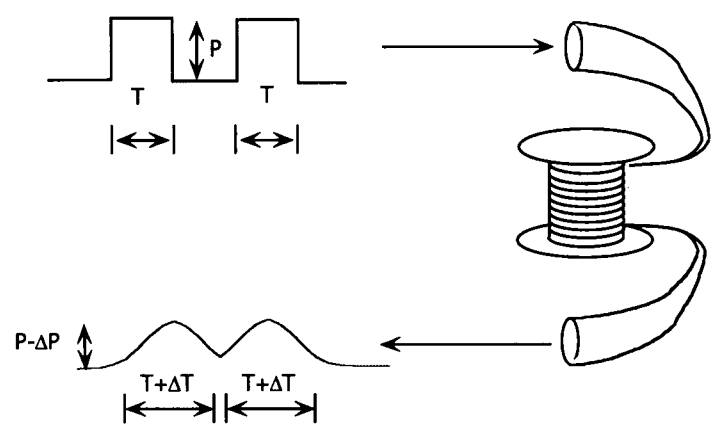


FIGURE 2.20 Pulse broadening: signal pulses are distorted as they are transmitted.

information-carrying capacity, of a fiber. Dispersion limits a fiber's capacity because an input pulse broadens as it is transmitted which causes pulses to overlap or become indistinct. Broadened pulses can become unrecognizable by detection circuits at fiber output and thus data communication would not occur. To transmit data successfully, the time between broadened pulses must be increased (pulse frequency must decrease), which reduces the data transfer rate. Transmission bandwidth is defined as the frequency at which output power is attenuated to one half of its input power (−3 dB). Dispersion and bandwidth are inversely proportional. Bandwidth can be approximated from dispersion by the following formula [20]:

$$\text{BW (MHz} \cdot \text{km)} = \frac{350}{\text{Dispersion (ns/km)}} \tag{2.37}$$

Dispersion can be categorized into intermodal and intramodal. Intermodal dispersion occurs because each mode in a MM fiber travel a different total distance. The effect is shown conceptually using rays in Figure 2.21. Each ray in a group of guided rays travels along a fiber at different propagation angles θ with respect to the fiber axis. Rays with larger θ have longer paths to travel than rays with smaller θ for a given fiber length. Recall that velocity in a transmission medium equals c/n. Therefore, the velocity of light is constant in a fiber with a core of constant refractive index. Because velocity is constant, propagating rays with a larger angle of incidence with the core/clad interface lag rays with smaller incident angles. This phenomenon is called group delay. A single input pulse that propagates in multiple modes will broaden as the different modes arrive at the output at different times. The time difference is typically measured in nanoseconds and is normalized for fiber length in kilometers. The time delay is expressed in units of ns/km. Intermodal dispersion is highest in step-index MM fibers. Intermodal dispersion does not occur in SM fibers: only one mode propagates in SM fibers and group delay does not occur.

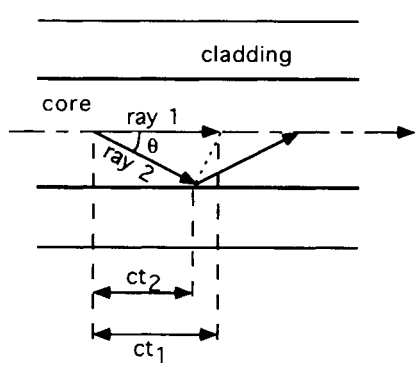


FIGURE 2.21 Group delay: ray 1 travels farther along a fiber's longitudinal axis than ray 2 in a given increment of time, but both propagate at the same velocity.

The time delay between two rays in a step-index waveguide is calculated by using the following analysis [21] if the following assumptions are made.

1. Rays propagating in a waveguide pass through the axis (the rays are meridionals);
2. Rays are totally reflected at the core–clad interface; and
3. The core is homogeneous.

First define $\tau = 1/t$ where t is the speed of light in a medium. For the two rays in Figure 2.21,

$$\tau_1 = \frac{1}{t_1} = \frac{n_c}{c} \qquad \text{and} \qquad \tau_2 = \frac{1}{t_2} = \frac{n_c}{c \cdot \cos\theta}$$

where n_c is the core index of refraction and θ is the angle of the incident ray with the interface,

$$\Delta\tau = \tau_2 - \tau_1 = \frac{n_c}{c}\left(\frac{1}{\cos\theta} - 1\right) \tag{2.38}$$

Next, consider Snell's law for a ray traveling in the core and incident on the core–clad interface, $n_c \cdot \cos\theta_c = n_{cl} \cdot \cos\theta_{cl}$, where θ_{cl} is the angle of the refracted ray and n_{cl} is the cladding index of refraction.

For the special case when the refracted ray is parallel to the interface, that is, $\cos\theta_{cl} \cong 1$:

$$\cos\theta_c = \frac{n_{cl}}{n_c} \tag{2.39}$$

Substitute Eq. (2.39) for $\cos\theta_c$ in Eq. (2.38):

$$\Delta\tau = \frac{n_c}{c}\left(\frac{n_c}{n_{cl}} - 1\right) \tag{2.40}$$

For an optical fiber with $n_c = 1.47$ and $n_{cl} = 1.46$, Eq. (2.40) predicts a delay of 34 ns/km. One can see from Eq. (2.40) that group delay varies with core and clad refractive indices.

Grading the change of refractive index between core and cladding offers a means to reduce intermodal dispersion by reducing group delay. The concept of rays propagating in a graded-index (GI) fiber is illustrated in Figure 2.22. Recall that the speed of light in a transmitting medium is c/n. If light were to travel in both the core and the cladding of the fiber (which it does to a small extent), it would propagate faster in the cladding because its refractive index is lower than the core refractive index for a given fiber. In a GI MM fiber, light traveling at the largest propagation angle θ travels faster when it travels near the core/clad boundary than near the fiber axis. Rays also bend as they travel through a GI MM fiber, and bending is more pronounced for larger θ. The result is that rays that have to travel the farthest also travel the fastest. Refractive index variation can be constructed so that all modes of a signal

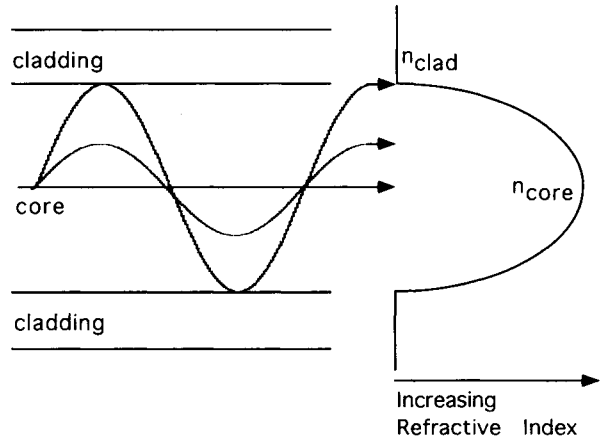


FIGURE 2.22 Ray propagation in graded-index fibers: the components of velocity parallel to a fiber's longitudinal axis for all rays are approximately equal.

pulse arrive simultaneously at any plane section through the fiber. In other words, the speed of all modes measured along a fiber axis can be made equal, which eliminates group delay. The distribution of refractive index in a fiber is its index profile. In practice, an index profile that is approximately parabolic produces the least amount of intermodal dispersion.

Intramodal dispersion is also called chromatic dispersion because it depends on the range of wavelengths transmitted by a fiber. As Figure 2.23 illustrates, chromatic dispersion is the sum of two components: waveguide dispersion and material dispersion. The figure also depicts the fact that, unlike intermodal dispersion, chromatic dispersion can be positive or negative. When material dispersion, which is always negative, is added to waveguide dispersion, the total can be positive or negative. The two components of chromatic dispersion can also cancel each other, yielding zero dispersion.

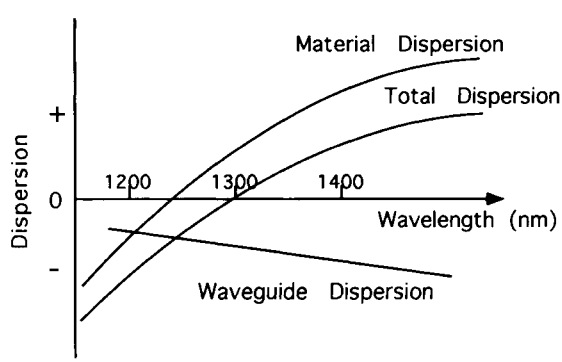


FIGURE 2.23 Chromatic dispersion and its constituents in SM step-index fibers.

The wavelength at which chromatic dispersion equals zero is called the zero-dispersion wavelength. The zero-dispersion wavelength for SM fibers is about 1310 nm, a wavelength that also has low attenuation and for which commercially available semiconductor lasers exist. When one considers that bandwidth is inversely proportional to dispersion, one might think that bandwidth could be infinite at the zero-dispersion wavelength. Bandwidth remains finite, however, because other, smaller effects dominate when chromatic dispersion becomes negligible. In addition, no light source in fiber optic communication operates at a single wavelength. Because zero dispersion occurs at one wavelength only, some chromatic dispersion will always occur.

Material dispersion results from the fact that the refractive index n of silica fibers varies with wavelength. Because the light sources used in fiber optic communication systems emit pulses with a range of frequencies rather than a single frequency, different portions of the pulse will travel with different group velocities. Group velocity variation causes group delay and pulse dispersion. The selection of light sources influences dispersion in a system because different sources have different spectral characteristics. Signal pulses generated by more spectrally narrow sources such as lasers experience a smaller range of refractive index variations. Less group velocity variation reduces group delay and pulse broadening. LED sources characteristically emit wide spectral ranges and experience more material dispersion.

Waveguide dispersion results from spectral width, fiber construction, and variation of n with λ [16]. Exact solutions for its governing equations and an explanation in physical terms are difficult.

2.4.5 Mechanical Characteristics of Optical Fibers

To some, fiber-optic cable and copper cable may look similar, but there are important differences between a glass fiber and a copper wire. Silica glass has almost 20 times the tensile strength of copper. At 14 GPa, it approaches the tensile strength of steel (20 GPa). Glass is brittle, it can only withstand up to 5% elongation before breaking, compared with copper, which can withstand 25 to 30%. Copper is ductile and can withstand bends of radius as small as its own radius and still properly conduct signal pulses. It will also deform plastically and maintain continuity when pulled axially. Glass fibers will tolerate a surprisingly small bend radius before breaking. Fibers will return to their original, straight state when released because they are inelastic. Tensile strength of silica glass drawn into fibers, however, is less than the ideal silica fibers and the actual strength is much lower: on the order of 700 to 3500 MPa. [22].

From work by Griffith on the strength of materials, the strength of a body is determined to a large extent by surface defects. For optical fibers, cracks and flaws arise primarily from contact with objects (e.g., grinding, winding, and

other mechanisms) and from defects that exist in silica blanks before fibers are drawn. Stresses generated by twisting, bending, and pulling fibers are concentrated at cracks and flaws. Cracks will grow when concentrated stress exceeds breaking strength. When a crack propagates to a point that the bulk material can no longer support the load, a fiber breaks. The largest crack on a fiber's surface concentrates stress the most; consequently, it is the largest crack that limits a fiber's strength. To illustrate the effect of surface flaws on fiber strength, first consider the theoretical strength of flawless silica σ_0 [16],

$$\sigma_0 = \sqrt{\frac{E\gamma}{a}} \tag{2.41}$$

where E is the Young's modulus, γ is the surface energy of the material, and a is the atomic spacing or bond length. For silica glass, $E = 7 \times 10^4$ N/mm^2, $\gamma = 7 \times 10^{-4}$ N/mm^2, and $a = 2 \times 10^{-7}$ mm. Equation (2.41) yields a theoretical strength of 1.6×10^4 N/mm^2. The theoretical breaking stress applied to a fiber with a crack h deep is σ_t:

$$\sigma_t = \sqrt{\frac{2E\gamma}{\pi h}} \tag{2.42}$$

Equations (2.41) and (2.42) can be solved for the quantity $E\gamma$ and equated to yield an equation for breaking stress in terms of theoretical strength.

$$\sigma_t = \sigma_0 \sqrt{\frac{2a}{\pi h}} \tag{2.43}$$

For a silica fiber with a flaw 1×10^{-3} mm deep and an atomic bond length of 2×10^{-7} mm, $\sigma_0 = 90\sigma_t$ and an actual strength that is nearly 1/100 of theoretical strength is predicted [16].

Silica fibers exhibit tensile strength lower than theoretical maximum when exposed to water or humid conditions. Silica fibers weaken over time in environments containing water in both stressed and stress-free conditions. Failure in a stressed state occurs by a mechanism called static fatigue or stress corrosion; the weakening of fibers when no load applied is called stress-free aging [23]. Both stress corrosion and stress-free aging are a result of attack

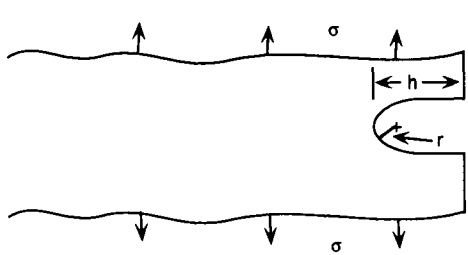


FIGURE 2.24 Microcrack on a fiber surface.

by water on silica glass. Static fatigue is noted by decreasing tensile strength versus the time a fiber is concurrently exposed to water and axially loaded. Stress-free aging is characterized by decreasing tensile strength of fibers exposed to water but not loaded concurrently. Static fatigue and stress-free aging both vary with the number and size of surface flaws. Fibers are coated after drawing to limit water ingress and surface damage.

2.5 Optical Fiber Cable Design

An optical fiber may be properly designed to meet optical requirements for an application, but there are other design considerations. Silica fiber by itself is not robust enough to operate reliably in most real operating environments—fibers are sensitive to chemical contamination and are brittle. Fibers are protected from stress by forming them into cables. Cable structures protect fibers from adverse mechanical and ambient environmental factors that result from installation, operation, reconfiguration, and service of links. Cabling fibers also eases handling of small fibers. Proper selection of a fiber optic cable requires an understanding of the application environment, cable types, and cable processes. The following sections discuss these three pertinent considerations for fiber optic cable.

2.5.1 Cable Types

Two types of fiber optic cable constructions are common: loose tube and tightly buffered. Cross-sections of each are illustrated in Figure 2.25. Tight-buffered cables have one or more layers of polymer encasing coated optical fibers. Loose-tube cables contain coated optical fibers in a tube having a diameter larger than the sum of the contained fibers. Loose-tube cables sometimes are filled with jelly for additional protection.

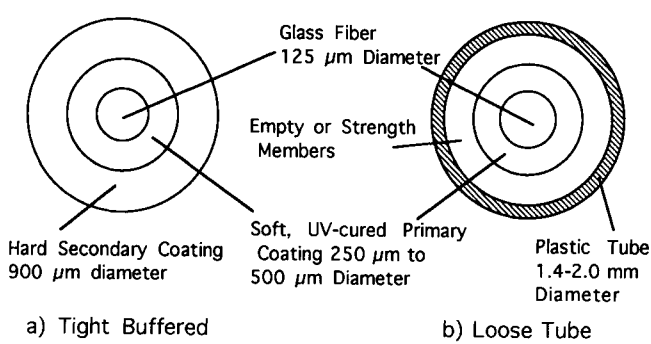


FIGURE 2.25 Common single-fiber cable types.

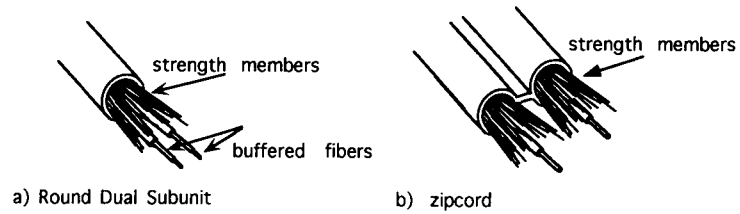


FIGURE 2.26 Duplex cable types.

Two types of cable construction are commonly used in data communication point-to-point links. They are zipcord and round dual subunit (RDS), and are illustrated in Figure 2.26. Zipcord is similar in appearance to household electrical cords. It is a tight-buffered cable with two fibers coated with a polymer jacket. A web between two fibers holds them in a fixed center-to-center distance and enables easy separation of cable halves. Between optical fibers and cable jackets are strength members, for example, aramid yarn, that are wrapped around fibers just before two fibers are jacketed. Cable jackets are generally extruded plastics that are applied as fibers pass through a die. After extrusion, the cable is cooled in long water troughs. The cooling rate is controlled by water temperature and flow rate to produce desirable jacket properties. Fibers in tight-buffered cables that are used in data communication are free to move longitudinally inside the jacket. RDS cables have an extruded jacket, too. Rather than a jacket extruded tightly around coated fibers and strength members, RDS cables are loose-tube cables. Two fibers are drawn through an extrusion die. Before the fiber reaches the die, strength members are wrapped around the fibers as a pair. RDS cable construction accommodates slack in fibers, which lessens residual strain.

2.5.2 Mechanical Environment

Cables must withstand mechanical strains imposed on them by the operating environment they are used in. A properly designed cable keeps strains from reaching the fibers. Jacket, strength member, and fiber work as a system to keep stress in fibers below critical levels. Before a cable is mass produced it is tested for reliability during development. Cable development is based on models of the mechanical operating environment. Estimates and measurements of mechanical environments provide test parameters and methodologies for developing robust cables. Tests check the ability of a cable's jacket, strength members, and fiber coating to protect glass fibers from the mechanical environment. One list of tests for cable is presented in Appendix A.

Mechanical environments for fiber optic cable applications have been studied and standards organizations have developed standardized tests.

Standardized tests benefit the fiber-optic industry because they promote common test methodologies which yield comparable, repeatable data between manufacturers. Standardized test procedures have been developed by the Electronics Industries Association (EIA) under the category of Fiber Optic Test Procedures (FOTP), for example.

Typical sources of mechanical strain found in operating environments are the following:

- Crush loads
- Tensile loads
- Impact loads
- Bending

Crush loads are forces applied to a length of cable and sustained over time for more than an instant. Crush loads are generated by people stepping on cables, machinery sitting on cables, and objects rolling over cables, for example. Crush loads are modeled by compressive strength. One approach to testing compressive strength is by placing a length of cable between two plates of specified size. A specified force is applied normal to the two plates in order to crush the cable under test. Cable attenuation is measured before and during the test. Failure is defined as either an increase in attenuation by a specified value or fiber breakage.

Tensile loads are axial forces applied to cables and can be categorized as either transient or residual. Residual loads are forces sustained through time, whereas transient loads are short in duration. For example, residual forces may be generated while a long section of cable is suspended after an installation is complete; transient forces may be generated while cables are pulled through cable conduits during installation or when a cable is tripped on after installation. All these conditions impart tensile loads on cables.

Tensile loads are modeled with and tested by simple tensile tests of cable samples. Tensile tests check how effective a cable is at buffering its fiber from axial loads. Tensile tests also indirectly confirm that fiber axial strength is not adversely affected by a cable manufacturer's processes. A general tensile test that provides attribute data is performed by wrapping a cable around two mandrels and leaving a free span of specified length between them. An axial force is applied to the cable on test at a specified load rate until a specified load is reached. The cable is loaded for a specified amount of time while light is sent through the cable. Failure is defined as fiber breakage or an increase in attenuation of greater than a specified amount. A test method that generates variable data is pulling cables to failure and measuring the force at breakage.

Impact loads are forces concentrated on a small section of cable that are large in magnitude and short in duration. Impact loads are generated by dropped objects, for example. Resilience to impact loads is modeled by an

impact test. Impact tests are performed by dropping an object with specified mass and "nose" shape from a specified height several times on a cable. The object hits a cable normal to the cable axis. The product of mass and drop height yields impact energy. Cable impact tests, like Charpy v-notch tests, impart a high-magnitude, short-duration force to a small area. Impact tests gauge a cable's effectiveness at buffering the contained fiber from the impact energy.

Routing cables in data communication applications require them to be bent. Bends are a source of stress in fibers, and tighter bends induce higher stress. A trend in equipment packaging toward smaller packages translates into requirements for smaller bends in the fiber. In addition, free space in new equipment becomes scarce as packages shrink. Routing cables in data processing equipment takes up large, relatively unproductive space which leads system designers to force cable to take less space by bending them closer to safe minimums rather than permitting large, safe bends that rob space. The ability of a cable to withstand small bend radii is gauged by testing optical performance while a cable is bent to a specified bend radius. One means to test for many possible conditions is to run a battery of tests that increase bend radius inversely with tensile load and/or duration of load application. For example, smaller bend radii are tested at lower tensile loads and larger bend radii are tested at higher tensile loads. Performance at a minimum long-term, no-load bend radius is also desirable because it serves as the tightest reliable bend in a cable. Table 2.4 lists test parameter matrices that can be used to test rigorously the reliability of bent cables. A less rigorous but applicable test matrix is listed in Table 2.5.

In addition to protecting fibers from adverse mechanical and environmental stresses, cable structures ease handling of small optical fibers used in data communication. Cable design helps make fibers less cumbersome to handle, especially groups of fibers. The diameter of coated fibers are small; diameters are typically less than 1 mm. Cables group fibers into manageable units with

TABLE 2.4 An example of parameters for cable bending tests

Duration	Load	Bend radius	Failure mode
1 to 5 s	400 N	4 mm	Fiber breakage
5 to 60 s	222 N	4 mm	Fiber breakage
5 to 60 s	450 N	48 mm	Fiber breakage
5 to 60 s	1000 N	75 mm	Fiber breakage
Long	0	12 mm	Attenuation change exceeds spec.
Long	85 N	25 mm	Attenuation change exceeds spec.
Long	231 N	31 mm	Attenuation change exceeds spec.
Long	330 N	37 mm	Attenuation change exceeds spec.
Long	401 N	43 mm	Attenuation change exceeds spec.
Long	447 N	48 mm	Attenuation change exceeds spec.

TABLE 2.5 An example of a short list of parameters for cable bending tests

Load	Bend radius	Failure mode
None	30 mm	Attenuation change exceeds spec.
80 N	50 mm	Attenuation change exceeds spec.
800 N	75 mm	Attenuation change exceeds spec.

a single, continuous jacket. Handling is also eased simply by a larger cable structure. For example, zipcord cables containing two fibers have typical dimensions of 3 mm by 6 mm; two-fiber RDS cables are typically 4.5 mm in diameter. Cables are used to group fibers into sensible groups. Data communication point-to-point links require two fibers, one for sending signals and one for receiving them. Two-fiber cables are a convenient cabling unit for data communication and are called duplex cables. Cables designed for long distance, high-fiber count links, called trunk cables, frequently have 24 or more fibers in one cable.

The material and manufacturing processes selected for a cable's outer jacket affect handling properties, too. Two properties that are affected are cable flatness and cable roughness. Cable flatness is how flat a cable will lay during and after installation. It is a combination of the stiffness and the "memory" of the jacket material. Cables that do not lie flat are a concern because they hinder the installation process. For example, they are difficult to install in routing trays or get tangled under raised floors. Cables do not lie flat in part because jacket materials "remember" the shape they were stored in. Memory begins when finished cable assemblies are either wound on small--diameter reels or coiled for packaging in a bag where they are stored for potentially long periods of time. When cables with high memory are uncoiled or unwound they retain the shape of the coil or reel. When laid out for installation they take the shape of a long coil spring and do not lie flat. Cable flatness is influenced by extrusion and cooling process parameters, jacket material, and coil or spool diameter used for storage.

Flatness can be monitored by testing how flat cable samples lie on a floor when uncoiled or unreeled. One approach has been to uncoil 10 m of cable from a reel under test on a hard floor. After a specified waiting period under specified room conditions, the height at two or three points at which the cable lifts farthest from the floor are measured. A failure is any one measured height that exceeds a minimum value. A waiting period of a few minutes between unreeling and measuring is advantageous if the intended application will permit it because some cable jackets will relax quickly even if they initially exhibit memory. Defining a set of environmental preconditioning steps and defining reel and coil size help to generate repeatable test data.

The second handling consideration affected by manufacturing processes and material selection is cable roughness. Roughness is a term given to how much friction force is generated when the cable jackets' outer surfaces rub against each other. More force is needed to pull cables with "rough" jackets along each other in crowded spaces than cables with smooth jackets. Cable jackets with low friction coefficients are desirable when cables are installed and pulled along each other. Cable roughness, like cable flatness, is adjusted by changing materials or process parameters.

Cables must perform reliably when subjected to excursions in ambient temperature, to humidity, and to corrosive chemical agents. Ambient environments are generally categorized into storage/shipping or operating. One list of tests for cables is presented in Appendix A. The EIA has developed standardized tests for ambient environment.

First, the cable must withstand excursions encountered during shipping and storage. A simple scenario is one in which a cable is shipped from its manufacturer to a data processing equipment manufacturer, then shipped with the equipment to an end customer. Temperature may begin at room levels, climb to higher temperatures in a truck sitting in a parking lot in a temperate climate, quickly decline to cold temperatures in an airplane cargo bay as the plane climbs in altitude, and increase again when the plane lands. Next, cables may be transported in a hot truck to a hot warehouse where they are stored in inventory. Several temperature cycles may occur before a cable reaches its final destination. Temperature and humidity may swing quickly during shipping and may shock cables.

To predict cable performance in rapidly changing ambient conditions exhibited during shipping, cables are frequently tested using thermal ship shock (TSS). Generally, a TSS test is performed by exposing cables to several cycles of high and low shipping temperature extremes in uncontrolled humidity with short transition periods between them. Loss measurements are periodically made at room conditions after an interval of cycles is completed. Like cables in storage, cables on test are non-operating: they are not connected to test equipment and are not transmitting light.

Second, cables may be exposed to cycles of heating and cooling and/or temperature extremes over their operating lifetime after installation. One may think that data communication cables are exposed only to steady, controlled, and benign ambient environments, but uncontrolled ambient conditions are encountered too. Air-conditioned offices and data centers are examples of controlled environments that come to mind. However, cables are routed in both outside walls of buildings and outside conduits where they encounter uncontrolled heating and cooling cycles; for example, with sunrise and sunset. Cables are routed inside data processing equipment that are turned on and off and exhibit temperature cycles. Cables are installed in less-controlled ambient environments that exhibit temperature extremes without cycling, too, such as manufacturing plants.

Performance of operating cables in temperature- and humidity-cycled environments can be tested using thermal cycle tests. In general, a thermal cycle test procedure exposes cables to several cycles of temperature extremes indicative of the operating environment, with slow temperature transitions and uncontrolled humidity. Losses in cables undergoing a thermal cycle test are periodically measured while in an operating condition. Cables are connected to test equipment while transmitting light and are measured during dwells in temperature.

Cable performance in sustained extreme ambient conditions during storage and shipping may be predicted using temperature and humidity aging (T&H aging). T&H aging tests expose nonoperating cables to cycles of high temperature and humidity.

Data communication cables must lastly perform reliably in corrosive environments. One can envision corrosive environments in industrial applications in which airborne compounds combine with humidity. Surprisingly, office, building, and interbuilding ambient environments are also corrosive. To ensure performance of cable in the field, reliability should be established by a corrosive environment test.

Fiber-optic networks include intrabuilding cables that are routed in numerous horizontal and vertical trays and conduits. Intrabuilding networks include cables that run between rooms and floors. A concern with intrabuilding fiber-optic cable, a concern extended from copper cables, is its ability to propagate a flame [24]. The specific concern is that cables can spread a fire from one room or floor to another through cable conduits that provide open channels between what otherwise might be independent spaces. A concern also exists regarding cables that generate smoke. Applications exist for cables that do not facilitate a flame or generate excessive smoke.

Since 1975, the National Electric Code (NEC) has established a set of standards that address smoke generation and flammability issues. Underwriters Lab (UL) has also established a set of tests for characterizing fiber-optic cable flammability and smoke potential. Two UL classifications are prevalent in data communication: Nonconductive Optical Fiber Riser cable (OFNR) and Nonconductive Optical Fiber Plenum cable (OFNP). Cables must meet requirements of flame spread tests UL-1666 to be classified as "riser rated" and UL-910 to be classified as "plenum rated." Riser-rated cables meet requirements for minimizing flame spread in vertical conduits; plenum-rated cables meet requirements for minimimizing flame spread and smoke density in spaces that transport environmental air. Riser- and plenum-rated cables are widely available and are prevalent in data communication.

Market opportunity for cables possessing more stringent combusion characteristics have been growing. Interest is growing in cables that not only limit flame propagation but also produce less smoke and less corrosive smoke. Thick smoke impedes visibility during an evacuation and is dangerous if

inhaled. Thick smoke may damage sensitive electronic equipment not damaged by heat or combusion, as smoke may contain corrosive gases and/or may deposit soot. In fact, the majority of damage to equipment is frequently from corrosive gases rather than combustion itself. Developing concerns for nonthermal fire damage can be categorized into toxicity (impact on people) and corrosivity (impact on equipment).

Polyethylene and PVC are common jacketing materials but their flammability characteristics differ. They are low cost, easy to process, and meet other requirements for a cable jacket. Polyethylene, however, is fairly flammable and is not suitable when fire hazard is a concern. PVC has inherently low flammability; flammability is low because of its halogen content and halogen materials are generally recognized as a combustion retardant. Another group of materials with good flame retarding characteristics is fluorinated polymers. Halogen materials; however, present a corrosion concern: materials like PVC generate corrosive gases when burned. Cables with flame-retardant jackets that are halogen-free and that produce low-corrosivity by-products when burned are viewed as an alternative. They are generally thermoplastic resins with fire-retardant additives.

Development of cables that are flame retardant and that produce low smoke requires a balance of characteristics. Thorough fire hazard analysis requires integration of ignitability, flame propagation propensity, smoke propensity, toxicity, and corrosivity. Other characteristics required in the balance are acceptable manufacturing properties and environmental performance. For example, a flame-retardant jacket material must not be so loaded with fillers that it cannot easily be processed or cannot maintain integrity in thermal cycle conditions.

General statements about the flammability attributes of halogen and nonhalogen jacket materials must be carefully made. A misconception is that all halogen materials are less flammable than nonhalogen materials. In a study by Dickinson [24], data showed time to ignition for most of the halogen materials selected for the study were less than nonhalogen materials. In contrast, it also showed that some halogen materials ignite faster. No clear, general winner existed for peak heat release rate, total heat release, or total smoke release.

Another misconception is that nonhalogen jacket materials produce less corrosive combustion by-products. The study showed that the majority of halogen/halogenated materials tested released gasses that were more corrosive than nonhalogen combustion by-products. However, some nonhalogen materials gave off significant amounts of corrosive gases.

Performance of each jacket material must be established by standardized tests for a fire performance attribute of interest. It is not accurate to assume, for example, that eliminating halogens from material formulations makes smoke from them both less corrosive and less toxic. No single test has yet been developed to address completely all flame and smoke issues.

Standardized tests are in development or in the process of establishment for fire attributes, but test development and establishment is a lengthy process. Results should be seen in the future.

2.6 Optical Fiber Connectors

Fiber-optic systems are capable of transmitting data over long distances. Specifications for several industry-standard lengths allow links of 2 km or more. It is possible to run a single cable over long distances between transceivers, but this kind of link is rarely practical in data communication. Cables comprising links that have the ability to be connected and disconnected are better suited to system needs. The ability for joining and unjoining cables is achieved by using connectors. The term *connector* is used to describe a component that permits joining of one or more fiber pairs in a manner that anticipates unjoining and rejoining by its design intent [25]. Connectors join in a separable fashion either two fiber ends or a fiber end and an active device. Connectors provide fiber-optic links with several benefits.

The primary benefit connectors lend to fiber-optic systems is that they allow partitioning of links. Partitioned links produce the following desirable characteristics of fiber-optic networks:

- Ability to reconfigure links
- Cost and performance optimization
- Ability to relocate equipment
- Replaceable link subunits

First, the ability to reconfigure fiber-optic links is an important attribute for communication networks because requirements for networks change over time. Reconfiguration involves rerouting cables in the network to different ports. Connectors ease reconfiguration because they provide field-separable joints for connecting and disconnecting cables. Second, connectors are useful for breaking links into sections that can be optimized for performance and cost. For example, links can be partitioned into cables that will be either installed permanently or reconfigured. Permanent sections might be premise wiring built into a building's infrastructure and terminated at a patch panel. Sections intended for reconfiguring might be jumper cables running from data processing equipment to a patch panel. Jumper cables generally are sections of cable under 100 m in length, terminated at both ends, and not permanently installed as part of premise wiring. Different manufacturing and installation processes are available for each category of cable. A partitioned cable network that consists of a mix of permanent and jumper cables can be optimized for performance and cost. Third, partitioning permits repositioning of equipment, which eases cable plant planning. Premise cabling is like the electrical wiring in a house; its installation is permanent, inaccessible, and

must be planned. Equipment is linked to premise cabling by accessible, shorter jumper cables that accommodate changes to equipment location. A mix of cable types lends a combination of flexibility and trunking capability to a network. Lastly, partitioning links creates smaller replacement units. Dividing links into smaller parts permits replacement of defective sections rather than entire links. Smaller replacement units are desirable because they are less costly to replace and cost less to stock.

There are some drawbacks to using connectors, however. First, connectors are the least reliable, most often damaged component in a fiber-optic link. One reason for this is that joining connectors is a manual operation that exposes them to handling. Conditions imposed by handling are difficult to predict and include mechanical force extremes that damage connectors. Useful connectors must be robust enough to endure handling. Second, connectors typically exhibit the highest loss of any single component in a link. For example, compare connector loss to attenuation by fiber. Single-mode fibers are available that attenuate signals less than 0.40 dB/km, but connector loss is typically 0.5 to 1.0 dB maximum each; attenuation in a cable 1 km in length is less than the loss in one connector. Connectors can exhibit losses lower than 0.5 dB but loss typically increases under adverse conditions. The drawbacks to connectors translate into some of the core challenges for improving connector design and manufacture.

2.6.1 Connector Types

A connector consists of two types of components: a terminal and a receptacle. Fiber-optic connections can be categorized into two types: fiber-to-fiber and fiber-to-transceiver. Fibers are assembled into terminals to provide both alignment features and a convenient unit to handle and insert. Terminals for both connector types are the same. The first type of connector, fiber-to-fiber (terminal-to-terminal), is required in networks for interconnecting two sections of connectorized cable. Joining is performed via a receptacle called an interconnection coupler (not to be confused with an optical coupler). A typical connector is illustrated in Figure 2.27.

A coupler consists of a cylindrical sleeve housed in a body that has features for guiding and retaining two terminals. Terminals are inserted into each end of the coupler sleeve. The interconnection sleeve aligns fiber cores within precise tolerances to enable optical coupling. Some connector designs actually have two terminal endfaces touch to reduce losses that arise from air gaps between coupled fibers. After terminals are fully inserted, a mechanical fastener such as a snap latch mechanically couples terminals and receptacles together. The second type of connector joins fibers to transceivers. Transceivers are designed with a receptacle to accept cable terminations that are similar to one-half of a coupler. They, like couplers, have both retention and guide features to accept a terminal, and have a sleeve to position precisely a

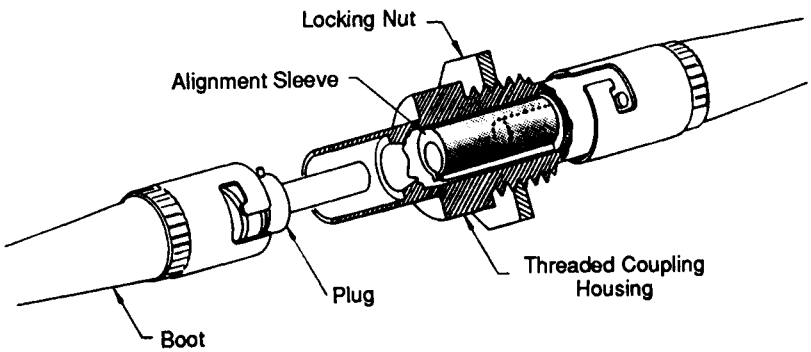


FIGURE 2.27 Typical connector. (Courtesy of AT&T.)

terminal's fiber with respect to an emitting or detecting device. Transceiver connections are examined more closely in Chapter 5, "Connector/Module Interface."

Connectors that join two fibers to complete one optical path are called simplex connectors. Simplex connectors are used quite extensively in telecommunication applications and data communication premise cabling. A connector that joins two fiber pairs to complete two optical paths simultaneously is a duplex connector. Duplex connectors are used in data communication applications because they meet some unique requirements. Duplex and simplex connectors are discussed in more detail later in the chapter. First, however, requirements for connectors and fundamental design elements are described.

2.6.2 Connector Design

Connectors must enable and maintain a separable joint between optical fibers with very small cores in adverse environments. Connectors couple two fibers with core diameters as small as 10 μm. Connectors for practical use must be easy to use, inexpensive, exhibit repeatable performance over numerous mating/demating cycles, and exhibit sustained performance over time. A goal for connector design is as close to a transparent, low-loss optical joint as possible. Improving connector design and manufacture presents a weighty challenge. Fortunately, manufacturers are taking on and meeting the challenges. Design and development of connectors is viewed by some people as less esoteric than work on other fiber-optic devices, but their importance in networks underscores the need for sound designs. The following sections describe fundamental connector elements that are commonly used to meet the requirements. Discussion focuses on connectors that complete one optical path. The elements are also utilized in duplex connectors and in connectors that join a higher number of fiber pairs.

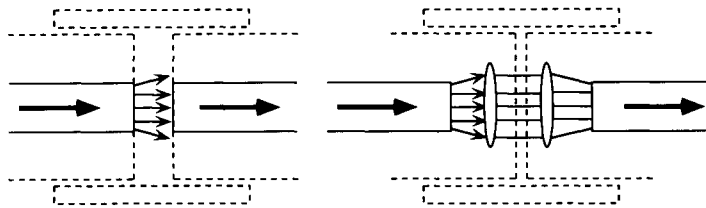


FIGURE 2.28 Connector joint types.

Two general types of joints prevail in connector designs: butt joints and expanded-beam joints. The two types are illustrated in Figure 2.28. Expanded-beam (E-Beam) joints consist of a lens positioned between radiating and receiving fiber ends. Light that radiates from the radiating fiber expands in roughly a cone shape. The lens, or series of lenses, collimates the light and focuses it onto the core of a receiving fiber. A butt joint consists of two fiber ends facing each other close together without any optical elements to focus the light from the radiating fiber. In some applications, fibers in butt joints are made to actually touch. In other applications, fibers not only touch, they are pushed together to generate a compressive force over the contact surface. These types of joints are called physical contact (PC) joints.

E-Beam and butt joints have advantages and disadvantages [25]. E-Beam joints simplify the difficult task of aligning mode fields in a direction lateral to the joint axis. The alignment task is eased because coupling occurs through a lens system that inherently tolerates laterally offset fiber cores. E-beam joints are also tolerant of foreign matter on interface surfaces. A disadvantage of E-beam joints is that they are relatively intolerant to angular misalignments. Angular misalignment is a condition in which the joint halves are tilted so that their longitudinal axes intersect at an angle when projected through the lens system. Angular misalignment is more easily controlled with butt joints, but butt joints are less tolerant of lateral offset and foreign particles than E-Beam joints.

Development of a ceramic ferrule-sleeve system and advances in ceramic component manufacture have yielded a practical means to control lateral offset in butt-joint connectors. Butt-joint connectors are a popular choice for data communication connectors.

2.6.3 Mechanical Coupling

Several methods are used for mechanically joining connectors together. Mechanical coupling is critically important for maintaining stable optical coupling over time. Mechanical coupling for most connector types is based on common fastening techniques.

One fastening method used in fiber-optic connectors is based on a screw. Connectors with screw-type fasteners are typically round in shape and have a ring that is held captive to a terminal body but that is free to rotate around it. The ring is threaded for fastening to a mating, threaded connector half. FC and Biconic connectors are two common simplex connectors that use threaded fasteners. Another method used is a quarter-turn fastener. As with screw fasteners, a ring is held captive to a terminal body and left free to rotate around it. A pin in a receptacle rides in a slot on the ring and is locked in place after the ring is turned. The ST (R)[1] connector is a popular connector that uses a quarter-turn fastener. Screw and quarter-turn connectors are illustrated in Figure 2.29.

Another common fastener is a spring latch. Spring-loaded latch hooks are typically located on one connector half and interlock with reliefs designed to capture and retain latch-hooks on a mating connector half. Connectors that use spring latches can be categorized into two types: push–pull and manual.

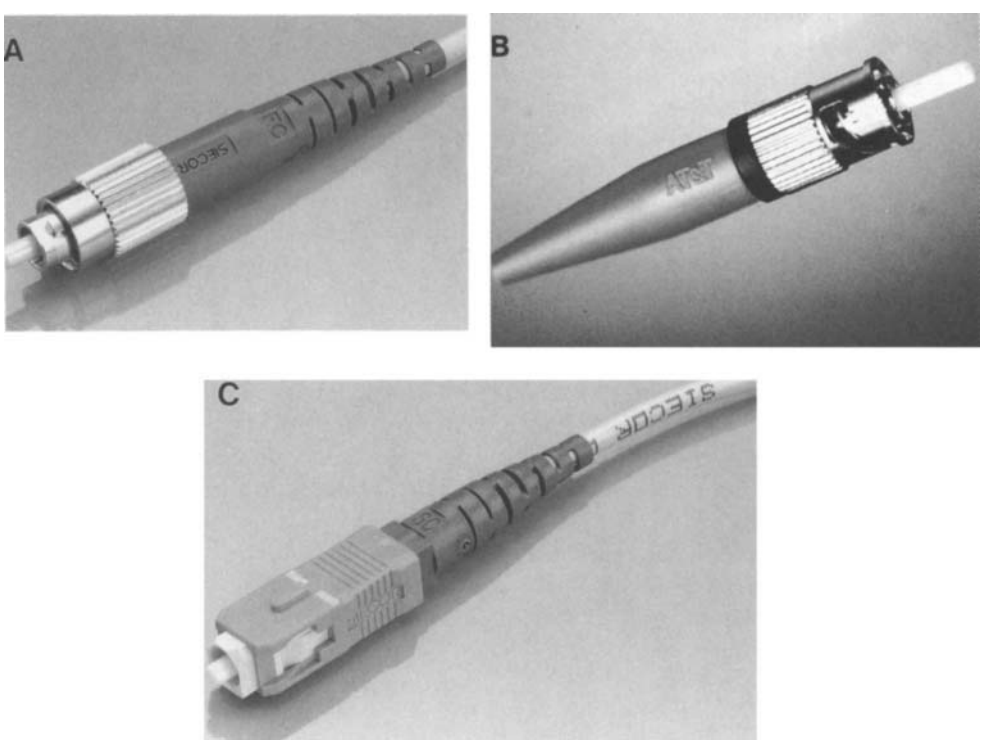


FIGURE 2.29 Simplex connectors: (a) FC, (b) ST (R), (c) SC. (a, c—Photos courtesy of Siecor Corporation, Hickory, NC; b—photo courtesy of AT&T.)

[1] ST is a registered trademark of AT&T.

Common connectors that use manual spring latches for retention are ESCON (R)[2] and FDDI connectors, as shown in Figure 2.30.

Manual-latch connectors have latch springs with small hooks are integrated into their shells. When a terminal is initially inserted into a receptacle, the latches are forced in toward the terminal's body. When the terminal is fully inserted, the latch hooks snap out into openings in the receptacle. To remove a terminal, the latches must be manually depressed to unseat the latch hooks. Push–pull connectors are also anchored by a pair of latches. As a terminal is pushed into a receptacle, latch springs are spread apart by inclined ramps in the mating connector half. After it is fully inserted, hooks on the latch springs engage reliefs in the mating half. A connector is decoupled by pulling on it. Inserting and removing a terminal forces the latch hooks out of and over the reliefs. A popular connector that uses a push–pull fastener is the SC connector. SC connectors incorporate latch springs into receptacles, unlike ESCON and FDDI connectors that incorporate latch springs on terminals.

Each type of fastener has advantages and disadvantages. Threaded fasteners in general are the most secure. They reduce optical coupling variability because they mechanically couple well. A disadvantage is that they require enough space around them to permit access and actuation by a person's fingers. Threaded fasteners also take the most amount of time to disconnect and reconnect. For these reasons threaded connectors such as the FC are popular for use in laboratory applications in which connection time and connector spacing are relatively unimportant but low variability is. Quarter-turn connectors also require space for a person's fingers to actuate a retention ring, but they can be connected more quickly than threaded connectors.

Round connectors in general must have features that eliminate rotation of terminals during fastener actuation. Rotation has a deleterious effect because sliding contact between ferrules and sleeves can cause wear. Wear will change dimensional relationships between connector components and will generate particulates that can interfere with either insertion or the optical interface. Rotation of ferrules can also damage polished endfaces in PC connectors. An effective means to limit rotation is a pin and guide combination. One connector half has a pin that rides in a guide on the mating connector half during insertion. The guide permits axial motion but not rotation.

Latched connectors generally are box-shaped and eliminate rotation. Like quarter-turn fasteners, latch fasteners can also be quickly disconnected. Latch systems require finger access space to two opposing sides of a connector only Consequently, they can be positioned close to each other along the free two sides or "brick-walled." Latch systems snap into retaining recesses and provide feedback in the form of an audible click after terminals are fully inserted. Latch systems have a disadvantage: they often require more complex designs and molding techniques than threaded or quarter-turn connectors.

[2] ESCON is a registered trademark of IBM Corporation.

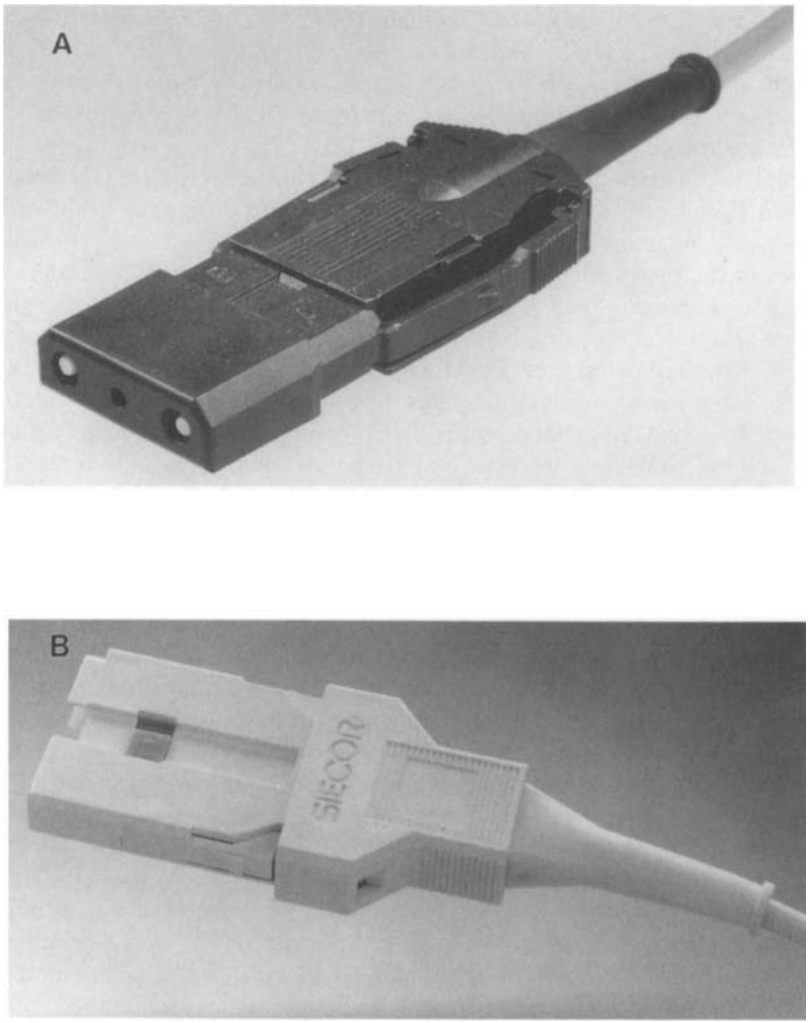


FIGURE 2.30 Duplex connectors: (a) ESCON (R), (b) FDDI, (c) FCS duplex (Fiber Channel Standard Compliant), (d) SC compliant duplex. (a, c, d—Photos courtesy of IBM Corporation; b—photo courtesy of Siecor Corporation, Hickory, NC.)

2.6.4 Precision Ferrule System

Mechanical fasteners couple terminal and receptacle together but contribute little to aligning fiber cores for optical coupling. An alignment system used for butt joints must be capable of aligning fiber cores to within a micrometer (multimode) or submicrometer (single-mode). A widely used system that meets these requirements is a precision ferrule assembly.

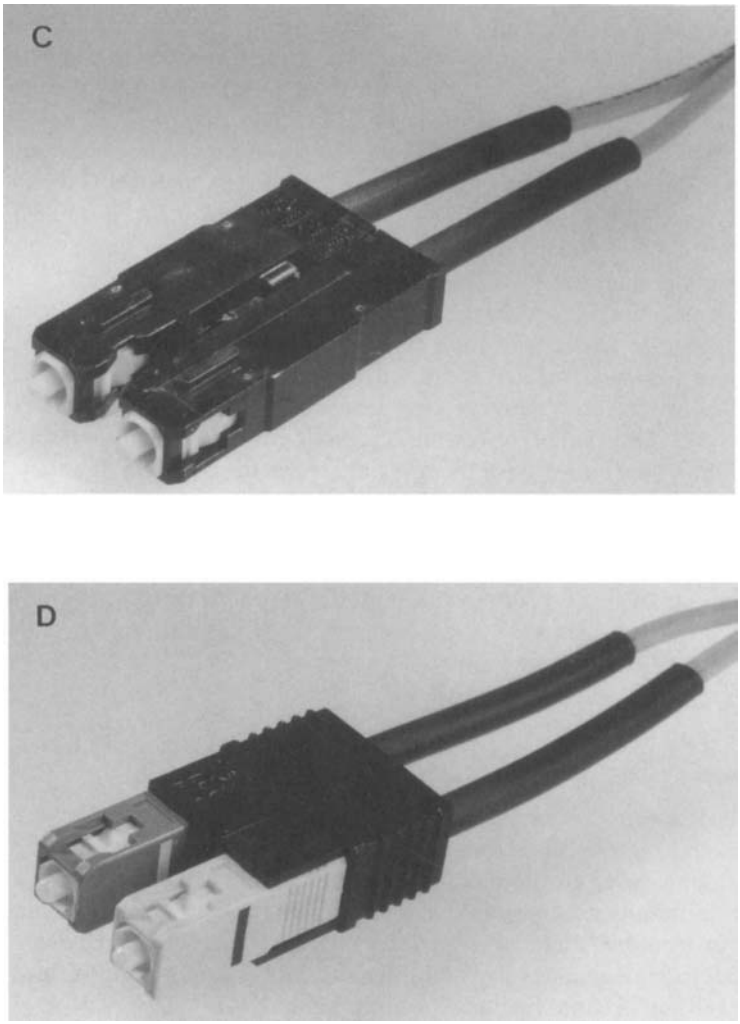


FIGURE 2.30 *Continued.*

A precision ferrule system packages fiber ends into a small axial hole in a precision-manufactured cylindrical terminus called a ferrule. A fiber is epoxied in a ferrule hole to fix it into position. The end of a fiber is left exposed at one end of a ferrule, and the end is polished to leave a smooth fiber/ferrule endface surface. The resulting fiber/ferrule assembly is lastly packaged in housing components to complete a terminal.

In precision ferrule systems, a joint between fibers is achieved by inserting two ferrules into opposite ends of a "sleeve." A sleeve is a cylinder with a hole or bore that leaves a thin wall. A sleeve with these characteristics is a

solid-bore sleeve. A sleeve aligns two fibers together via the ferrules. Sleeves are housed in a body to form a receptacle. Terminal and receptacle housings have fastener features built into them for mechanically coupling the components. To achieve the alignment requisite for good optical coupling, ferrule and sleeve dimensions must be tightly controlled. Almost mirror-like surfaces result from the polishing processes. Dimensions for a typical MM ferrule are specified as $2.499 + 0.001/-0.002$ mm; specifications for a typical SM ferrule are 2.499 ± 0.0005 mm.

Precision ferrules can be categorized into two types: capillary and through hole [26]. Capillary types are formed by pressing a precisely manufactured ceramic cylinder with a pin hole through the middle into a metal sleeve. The sleeve's outside surface is turned down to the desired dimension. A through-hole type is a cylinder with a precisely finished outer surface and through-hole. One end of the ferrule is pressed into a metal component called a backbone. Backbones have features that are designed to interact with features in a terminal housing for the purposes of providing guidance to and retention of a ferrule. Ferrules and backbones can be made economically in batches. The free end is polished for low optical loss.

Ceramic ferrules are made using the following general steps [27]:

1. Mix compounds
2. Extrude
3. Sinter
4. Hole polishing
5. Outer surface finishing

Material selection is important for ferrule, and thus connector, performance. Several materials are chosen that include ceramics, metals, and polymers. Ceramics used are alumina, zirconia, silicon nitride, and silicon carbide. A metal commonly used is stainless steel. Alumina and zirconia are frequently chosen over other materials for their ability to meet requirements for high-performance connectors. Consider the properties of alumina and partially stabilized zirconia that are listed in Table 2.6.

Note that zirconia has higher bending strength. It is also tougher; consequently, it resists chipping and abrasion better. Small grain size enables very

TABLE 2.6 Material properties of ceramics used for fiber-optic connectors [27]

Properties	Units	Alumina	Zirconia
Bending strength	MPa	340	1180
Young's modulus	GPa	370	190
Vicker's hardness	GPa	18	12
Grain size	μm	5	0.5

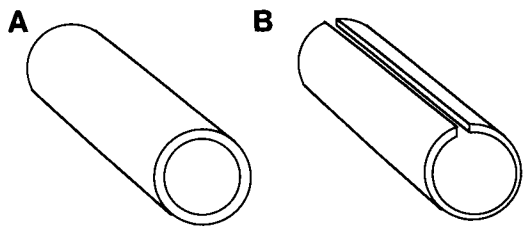


FIGURE 2.31 Interconnection sleeve types: (a) solid bore, (b) split bore.

smooth polished surfaces. Low Young's modulus, although it yields poor grinding characteristics, permits more stable physical contact connections.

An interconnection sleeve must be able to position the ferrules precisely. Good wear properties and tight dimension control and stability are required. A typical sleeve is manufactured to micrometer tolerances to yield a close slip fit. A method employed to ease the tight tolerances necessary for good optical coupling is to add an open slit along the length of the sleeves. Such sleeves are called "split" sleeves. A split sleeve and a solid-bore sleeve are illustrated in Figure 2.31.

Split sleeves act like very stiff springs whose inner diameter will expand. One property that results is that they will accept a wider range of ferrule sizes, including oversized ferrules that would interfere with a similarly toleranced solid sleeve. Another useful property is that they will center a ferrule that is larger than its inside diameter. For that reason, split sleeves are designed for a slightly smaller diameter than its corresponding ferrule so that the ferrule will always self-center. The two attributes of compliance and self-centering permit looser sleeve and/or ferrule tolerances, which requires less manufacturing precision and yields lower costs. A disadvantage of solid-bore ferrule systems is that they require very tight tolerances and close fits to center a ferrule.

Two sleeve materials that are frequently chosen to meet connector requirements are ceramics and phosphor-bronze ("phos-bronze") alloys. Phos-bronze sleeves are generally less expensive than ceramics. A disadvantage sometimes cited concerns their wear properties. Wear induced by repeated terminal insertions and withdrawals can generate particles of debris. Particles can interfere with the ferrule/sleeve interface and can increase terminal insertion/withdrawal force. Particles can also interfere with optical coupling between terminals and/or can damage ferrule endfaces in physical contact connectors. In addition, wear changes sleeve inner-surface dimensions and can change connector physical characteristics. Ceramic sleeves and ferrules offer better wear characteristics but are generally more expensive. The surfaces of ceramic ferrules and sleeves are hard, and smoothly polished. An advantage phos-bronze has is it is a tougher material than ceramic.

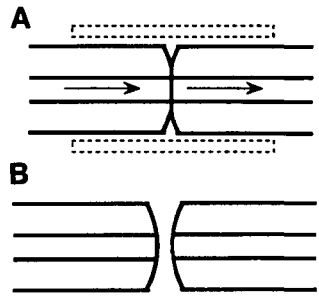


FIGURE 2.32 Physical contact butt joint.

The choice between ferrule material, sleeve material, and sleeve design must be weighed against requirements for the application. A characteristic of data communication environments is that the number of insertions and withdrawals for a connector are typically high: specifications for 100 insertion/withdrawal cycles or more are not uncommon. In data communication applications that require high reliability and durability, ceramic ferrules and ceramic split sleeves are frequently chosen.

Endfaces of ferrules for PC butt joints are polished in a manner that leaves a spherical, convex face. A PC joint is illustrated in Figure 2.32. In addition, fibers in PC joints are left to protrude slightly (on the order of micrometers) from the ferrules. Fiber protrusion ensures that the fibers will come in contact. When two PC terminals are joined they are pushed together by a biasing force built into the terminals to concentrate stress at the surface of contact, causing them to deform and flatten over a very small area. PC joints eliminate gaps between fiber surfaces by this process, and eliminating this gap lowers connector loss. Loss mechanisms and characteristics are discussed in the following sections.

2.6.5 Connector Optical Characteristics

A combination of components must be selected and matched to achieve a link that will operate as a communication channel. One parameter that must be carefully addressed is that enough light is detected at the receiving end of a link. The parameter is influenced by two considerations: sensitivity of the receiving device and optical power emitted by a connected fiber. Receiver sensitivity defines the minimum optical power that must be coupled into it from a connected fiber. Minimum optical power depends on emitter power at the other end of a fiber and all connection and transmission losses between emitter and receiver. The primary sources of loss in a link are at emitter-to-fiber and fiber-to-receiver optical connections, terminal-to-terminal connections via couplers, and fiber attenuation. For reliable channel operation, a link

budget must be met. A link budget is the maximum amount of loss tolerable to still achieve proper receiver operation. All sources of loss in a link must be quantified and totalled. The total must not exceed the link budget. A detailed discussion of emitter and receiver connection loss is presented in Chapter 5. Optical coupling and fiber attenuation in fibers have been presented in sections 2.3 and 2.4.3, respectively. The remaining topic is the mechanics of aligning fibers to achieve good optical coupling.

Loss occurs when a factor causes a reduction in optical power. Sources of loss can be envisioned at points where a waveguide deviates from a perfectly transparent path for the transmitted light. Connectors are one example of a discontinuity in optical paths that causes loss of optical power. There are several causes of loss at connector interfaces. Most factors are a result of either manufacturing variability or allowances in a design made to accommodate variability. The factors can be divided into two categories: intrinsic and extrinsic. Examples are given in Table 2.7.

Intrinsic factors are a result of either variability in fiber attributes or dissimilar attributes between interconnected fibers. An example of the former is when cores for two fibers are slightly out of round. If the major axes of the elliptical cores are joined 90 deg out of phase, the mode fields will not be completely coupled and some loss will occur. An example of the latter is when a step-index fiber is connected with a graded-index fiber.

Extrinsic factors are a result of variability in connector manufacture, manufacturing process limitations, and design attributes. Contributions to loss factors from manufacturing variance comes from the fact that all manufactured parts in their final form deviate from their basic ideal dimensions or states. The first three conditions from Table 2.7 that contribute loss are illustrated in Figure 2.12.

Offset is a term used to describe the condition when fiber axes are parallel but not collinear. Offset is a function of the basic dimensions of the ferrule and sleeve, tolerance ranges, and wear. It is also a function of ferrule hole position variation with respect to ferrule center, hole size variation, and fiber diameter variation. Gap is the amount of space between terminal endfaces. Gap is a function of designed gap and tolerances for both ferrule length and endface polished surface. Tilt is the term used to describe when ferrule axes

TABLE 2.7 Connector loss factors

Extrinsic	Intrinsic
Lateral offset	Mode field diameter mismatch
Axial gap	Index profile mismatch
Tilt	Core eccentricity
Endface quality	Fiber and ferrule hole clearance
Reflections	Ferrule hole eccentricity

are not parallel. It is a function of many of the same parameters as lateral offset. Submicrometer tolerances are required for SM connectors and micrometer tolerances for MM connectors to achieve connection losses of less than 0.5 dB.

To approach a transparent joint terminal, endfaces must be highly polished and free of discontinuities. Ferrule endface polish quality is a function of the polishing process selected and manufacturing variability. Limitations in the process selected may leave scratches, chips, or roughness which increase loss over defect-free surfaces. Lastly, loss can be induced by changes in ambient conditions; changes in humidity or temperature may cause dimensional changes, for example.

The nominal value, variability, and repeatability over time of connector loss are important parameters used to select connector designs, testing programs, and manufacturing processes. Two loss parameters are typically considered. First, optical power is lost at connector discontinuities because of coupling between fibers that is less than perfect. This type of loss is called connection loss. Loss is expressed by the following relation:

$$\text{Loss} = -10 \cdot \log \frac{P_{\text{out}}}{P_{\text{in}}} \tag{2.44}$$

Transmitted power is also lost by reflections at fiber-to-fiber interfaces. This second type of loss is called return loss. Some light at joints is reflected back toward emitting fibers rather than coupled to a receiving fiber. Emitting fibers then transmit the reflected light back to the source. Return loss occurs because light that travels through a boundary between media having different refractive indices is reflected and refracted. From Fresnel equations that describe the state of light traveling at a media boundary, the percentage of radiation reflected at a boundary is the reflectivity R [28]. Reflectivity for an angle of incidence normal to the boundary is given in Eq. (2.34). Light transmitted between two fibers with a gap between them encounters two boundaries, which causes more loss than a single boundary.

There are methods to reduce return loss. One method is to use PC joints because they eliminate gaps between fiber surfaces and reduce the effect of two boundaries. Return loss decreases rapidly with increase in contact force. A point is reached, however, when loss flattens out with increasing force [26]. Another method is to put a gel between ferrule endfaces. A gel with a refractive index equal to the fibers in a joint will, in theory, eliminate the effect of boundaries between fibers. A third method involves polishing a ferrule endface at an angle to the ferrule axes. A boundary at an angle to the direction of light transmission greatly reduces the Fresnel loss.

In addition to causing power loss, reflections can degrade performance of semiconductor laser emitters. Reflections at connector joints can be transmitted back to the emitter and interfere with laser operation. The magnitude of degradation depends on the amount of light reflected back to the source

and the length of fiber. The connector in a link nearest to an emitter is of the most concern: it has the shortest length of fiber between it and the emitter to attenuate reflected light.

2.6.6 Connector Application Requirements and Testing

Connectors and attached cables are exposed to stresses that can degrade performance or cause failure. Connectors in operating, nonoperating, free, and connected states are exposed to stress. Stresses come from conditions in the mechanical and ambient environment and are imposed during storage and shipping, installation, operation, and service. Connectors must be rugged enough to endure repeated exposure in both environment types. Telecommunication and data communication applications expose cables to many of the same types of conditions. For example, connectors in both types of applications are exposed to forces transmitted by attached cable. Ambient conditions in fiber-optic applications are not always constant and benign. Telecommunication and data communication connectors are both exposed to extremes in temperature, humidity, and airborne contaminants.

Although telecommunication and data communication applications expose connectors to the same types of conditions, environments for data communication applications differ in some ways. Data communication conditions are similar in type to telecommunication but differ in magnitude for some types of exposure. Data communication networks are generally reconfigured more frequently than telecommunication. More frequent handling leads to more opportunity for mishandling and damage. Data communication environments include connectivity to equipment that are an integral part of office environments. Cables leading to the equipment may not be isolated from foot traffic, chair rollers, equipment relocation, and other activity. These day-to-day activities increase opportunities for damage over relatively isolated telecommunication and premise wiring installations. Damage during use is a large source of connector/cable failure. Data communication jumper cables are generally viewed as operating in more benign ambient conditions. For example, some data processing equipment is located in raised-floor data centers where ambient conditions are well controlled. Although exceptions exist, data communication connectors in general are not exposed to the extremes in temperature that telecommunication cables are exposed to.

Usage in an application is one way to determine performance of a connector; however, for obvious reasons, this is not usually a good idea. Tests would necessarily take a long time and results would not be timely. Instead, connectors are tested during performance and reliability verification phases of development with a battery of tests designed to approximate exposures to ambient and mechanical extremes that a connector is likely to see over its lifetime. Tests are shortened by selecting acceleration factors for test parameters. One list of tests is presented in Appendix A. Standardized tests have

been developed by the EIA; they enable generation of comparable data for the same test parameters throughout an industry.

Users of fiber-optic connectors, like users of copper connectors, place emphasis on ease of use. In data communication environments, ease of use is important because connectors experience frequent mating/demating as a result of network reconfigurations. Fiber-optic terminals must be easy to handle, insert, fasten, and disconnect. Characteristics of connectors that relate to ease of use are insertion and withdrawal force. The force (or torque) an average person can comfortably exert on a connector must be sufficient to operate it, so terminals are designed to insert, fasten, and withdraw at a maximum force that does not exceed these levels. Typical maximum insertion and withdrawal force values specified for a duplex connector are 80 to 98 newtons. Verification testing typically involves repeatedly inserting and withdrawing a terminal and measuring the force required. Other ease-of-use factors one must consider when selecting a connector are subjective; for example, fastener ease and terminal shape.

A terminal should be comfortably inserted into and withdrawn from a receptacle, but it should not be easily removed after they are mechanically couplied. Forces transmitted by cables to connectors must be accommodated to avoid affecting optical coupling and to prevent a cable from being damaged or pulled out of its terminal by the loads imposed on it. Loads can be categorized as short or long term. Axial and off-axis loads can occur as short, jerking loads while a connector is in use; for example, someone trips on a cable. Terminals in a free state can also be exposed to short-term loads. For example, terminals are sometimes seen as convenient handles with which to pull cables along floors, trays, and through wiring ducts. Long-term loads are those that are experienced over time. Loads transmitted to a terminal by its cable can be simulated with axial and off-axis pull tests. Pull tests are performed by applying a specified load to the cable exiting a terminal mated to a receptacle for a specified amount of time. The load is applied to a cable coincident with a connector centerline for axial tests and for off-axis tests, a cable is pulled at an angle to the axial centerline (i.e., 45 deg). There are two common methods for off-axis tests. One is to pull on a cable in a continuous fashion around the connector axis to form a cone-shaped line of action. A second method is to pull on a cable at 45-deg increments around a cone-shaped line of action. Typically, an acceptable connector must not exhibit more than a specified connection loss during any application of the load.

Cables, over their lifetime, become bent at the point where they enter the terminals. Bending is induced in cables when terminals are demated and remated, reconfigured to different ports, and in spaces crowded with other cables or by confines of a wiring closet. In these conditions, terminals can be repeatedly flexed relative to their cable over time. Repeated flexing is detrimental to fibers because it induces fatigue. Fatigue failure occurs with fewer cycles as the band radius decreases. Response to repeated flexing is

tested by bending a cable back and forth past the connector centerline for numerous cycles. Twist-flex tests are also a means to predict connector performance in applications where flex occurs. Flex tests are typically performed for a specified number of cycles with a load applied to the cable.

Another source of mechanical stress and damage is dropped terminals. A robust terminal must survive being dropped from a reasonable height. Drop tests are conducted by releasing a terminal from a specified height onto a surface for a specified number of times and measuring connection loss to determine any change and performing a visual inspection for physical damage.

Environmental tests and the rationale for them are the same as for cable tests. Tests are performed on mated, free, operating, and nonoperating connectors to simulate a variety of temperature and humidity conditions. Like cables, connectors are tested for corrosion resistance. To avoid corrosion, relatively inert materials such as polymers and metallic alloys such as brass, stainless steel, and phosphor-bronze are used. Unlike for cables, dust is detrimental to fiber-optic connector performance. Dust can cause damage to surfaces in contact and can interfere with good optical coupling. Connector designs must limit dust flow at the connection interface. Dust is even more of a concern in applications in which terminals will be mated and demated numerous times over their lifetime because interfaces are exposed more often. Dust is also a concern in applications that use an index-matching gel. Index-matching gels, as discussed previously, are used to reduce reflection losses. Gels are a viable option for separable and splice connections. However, connectors with terminals that will be demated and remated several times over their lifetime may not be good candidates for gels because gels are messy, difficult to apply consistently in a joint, and attract dust. Because links are reconfigured in data communication gels have not been a preferred method for reducing the return loss.

2.6.7 Termination Packaging

The key part of a precision ferrule connector is the ferrule assembly. (As a reminder, ferrule assemblies are the assembled unit composed of a ferrule tip and backbone, inserted and epoxied fiber, and polished endface.) However, ferrule assemblies used alone as terminals are neither robust enough to survive the rigors of application environments nor easy to handle. They are also insufficient for maintaining an optical joint through time. Ferrule assemblies require packaging to be practical. The combination of a ferrule assembly and its packaging is what commonly referred to as a terminal. Terminal packaging is designed to do the following:

- Ease handling
- Improve reliability of joints
- Provide protection
- Decrease optical loss

The following sections present typical design features of simplex and duplex terminals that are used to meet these requirements.

Simplex connector is the name given to connectors that couple two fibers into one optical path. There are several simplex connectors commercially available. Several popular simplex terminals are illustrated in Figure 2.29. The terminals illustrated are designed to connect two fibers through a coupler receptacle. Several elements are common between different simplex terminals. A typical terminal is packaged in a housing that holds furrule assemblies captive and permits ferrule tips to protude and cable to exit. Typical housing shapes are round or square. Housings serve several purposes: they package everything into a single manageable unit, they protect ferrule assemblies and, they aid guidance and alignment of terminals and receptacles during insertion via features that are built into them. Housings also have fastner features for mechanical coupling functions.

A popular choice for low-loss simplex connectors is physical contact joints. Terminals for PC connectors use a spring to generate contact forces. PC ferrules are free to move axially and springs provide biasing force over the range of motion. Housing and ferrule backbones typically have a system of ridges, guides, and cavities molded into them to captivate ferrule assemblies, captivate springs, guide axial motion, and resist rotation of a ferrule. The system is also designed to limit extension and compression of the ferrule. Ferrules in PC terminals are biased towards maximum extension when a terminal is in a free state to ensure contact with a mating terminal. When in a connected state, ferrules of two terminals are displaced from maximum extension to generate contact force.

Connectors must accommodate variance in geometric attributes of receptacle and terminal subcomponents. Minimizing variance by tightening tolerance ranges would be one method for reducing variance, albeit an expensive one. Instead, ferrules are designed to move a small amount laterally about their centerline or to "float." Floating allows a ferrule to move with respect to both its own package and a receptacle. It also allows a ferrule to seek a proper position for entering a sleeve. Floating enables looser tolerance requirements, which can translate into lower cost.

Ferrule assemblies are delicate and must not be exposed to excessive strains. Cables transmit forces from the application environment to ferrules that must be accommodated. To relieve strain generated by cable forces, features are added to terminal packaging. Recall that fibers in cables are protected from strain by strength members such as aramid yarn which is wound around fibers inside cable jackets. The strength members absorb strain caused by cable stretch in an effort to leave the enclosed fibers strain free. The same principle is used to "strain relieve" fibers inside terminals: strength members are typically captured in a two-piece metal sleeve by crimping the sleeves against each other, and the crimp is held captive in the terminal by features built into its housing. The result is a transmission path for strain from strength member

to crimp to terminal housing to receptacle that does not include fibers or ferrules. It is important to strain relieve ferrules to avoid damage to ferrule assemblies and to absorb forces that would cause ferrules to move axially and cause a change in optical coupling.

Connectors must resist damage from off-axis pull forces exerted on their cables. These forces can induce a tight radius bend where a cable leaves a teminal housing. Damage may result from a bend radius that is too small or from concentrated stress at the point of contact between housing and cable. Terminals typically incorporate a component called "boot" to reduce the effect of off-axis forces on cable bending. Boots are plastic or rubber sleeves that cover cables for a short distance beyond the point where cables exit a terminal. One end of the boot is fixed to the terminal body, the other end is free. Boots act as cantilever beams and stiffen cables over a boot's length. Boots are effective in part because they, like cantilever beams, are stiffest near their base where a cable exits a terminal. It is at this point that fiber radius is likely smallest and the most aid is needed. Boots become gradually more flexible toward their free ends. The result of a properly designed boot is a gradual cable bend radius from the point it exits a terminal to the point it exits a boot. Boots also reduce damage to cables at the point where they exit a terminal by acting as a cushion between hard housing edges and a cable. Additionally, boots offer another opportunity to strain relieve a fiber from axial loads applied to cables: a cable can be bonded inside a boot to couple them together. Because boots are held captive to terminal housings, strain transmitted by a cable jacket to a terminal can be diverted to the housing through the bond in the boot.

Because it is sometimes desirable to distinguish connectors as MM or SM, or to ensure only MM or SM terminals plug into certain receptacles, keying features are added to connectors. Distinguishing MM and SM connector types by keying reduces plugging errors when MM and SM coexist in the same application environment. SC connectors, for example, use two combinations of key/keyway width for keying. Keyways are present on receptacles and keys are built into terminal housings. Wide key/keyways are used for MM connectors; narrow keying is used for SM. Another motivation for keying is safety. SM semiconductor lasers emit a high-powered, spectrally narrow beam. Not all optical power is fully coupled into a SM fiber because they are very small in diameter. If a MM fiber, which is 5 to 10 times larger in diameter, was plugged into a SM transceiver, it could capture considerably more optical power and present a safety concern at the opposite end of the fiber.

Connector keys are also used to reduce connection loss in SM connectors. Because SM cores are very small, connection loss is greatly affected by offset misalignment of fiber cores. Variability in core position inside a ferrule hole from terminal to terminal is a natural result of manufacturing. If two fibers are coupled with fiber cores diametrically opposite from each other with

respect to the joint centerline, coupling efficiency is less than maximum. When two fibers with cores offset in similar directions and distances from the joint centerline are mated, favorable conditions for high coupling efficiency occur. Rather than leaving core-to-core relative position to chance, some terminal designs deliberately orient core offset in the same direction for every terminal. One approach is to orient the direction of maximum core offset toward a key built into a terminal housing when a ferrule is assembled with a housing. Connection loss from core-to-core misalignment is systematically reduced for all mating/demating cycles if keys on mating terminals are repeatedly oriented in the same direction by keyways in receptacles. MM connectors generally do not require core offset positioning because they are less sensitive to lateral offset by virtue of their relatively large core diameters.

Packaging for data communication link differs from telecommunication links. Telecommunication sends data in both directions on a fiber link. Data communication sends data in only one direction in a fiber but needs two fibers for send and receive operations. Packaging of data communication link components is driven to accommodate this dual attribute.

Duplex connectors are terminals and receptacles designed for mating two optical paths in one step. In contrast, simplex connectors require two operations to connect two optical paths. Although both simplex and duplex connectors are used for data communication applications, duplex connectors uniquely satisfy a subset of data communication requirements that simplex connectors do not.

Recall that duplex cable simplifies handling of two fibers that are needed for send and receive functions in data communication. Consequently, selecting a connector designed for duplex cable has advantages. In addition, space is often constrained in data communication applications. Duplex connectors have an advantage in that they are typically smaller than two simplex connectors. Duplex connectors generally are used for data communication applications, whereas simplex connectors are not uniquely identified with data communication. Duplex connectors provide additional function, but the additional capabilities come at a cost: duplexing produces more complexity in a terminal.

Duplex connector technology uses many design elements from simplex technology. The primary building block is a precision ferrule system. Physical contact joints typically are used as well. Fibers are coupled by sleeves housed in receptacles that have features for mechanical fastening. The material of choice for sleeves and ferrules has been ceramic. Differences between simplex and duplex connectors appear in packaging. Packaging is designed to adapt a two-fiber, send/receive communication system.

Duplex terminals can be designed in a variety of ways. Four examples are illustrated in Figure 2.30: ESCON, FDDI, and FCS Duplex (Fiber Channel Standard Compliant), and SC Complaint Duplex. They all mate with a duplex

receptacle that has two sleeves in one package. All of them position ferrules in a side-by-side position at a nominal center-to-center spacing. ESCON and FDDI house ferrules with 0.700 in. between centerlines in a rigid single-body housing made of injection molded plastic. A dissembled ESCON terminal is illustrated in Figure 2.33.

Figure 2.33 shows the familar precision ferrule assemblies with springs for generating preloading. FDDI terminals are similar in concept to ESCON terminals. SC Compliant Duplex builds on two fully assembled simplex terminals. A clip couples two simplexes together in side-by-side positions. The clip is flexible and enables relative motion between simplexes. FCS Duplex utilizes terminal components for a pair of SC simplexes. It is designed with a low package height. FCS and FDDI connectors were adopted as standards by American National Standards Institute (ANSI).

In data communication applications, the need for two ferrules and two fibers per communication channel potentially complicates fiber and connector management. Duplex connectors offer improvements over simplex connectors alone. First, duplex terminals fix ferrules in a nominal position relative to each other and they do so in a single, "unitary" housing that is large enough not to demand great manual dexterity. This characteristic permits completion of two optical paths in one insertion operation. Second, duplex connectors simplify the important task of properly interconnecting transceivers. For proper operation, an emmiter in one transceiver must be coupled via an optical path to a receiver in a second transceiver, and the emitter in transceiver 2 must be coupled via a second optical path with the receiver of transceiver 1. This requirement is shown schematically in Figure 2.34.

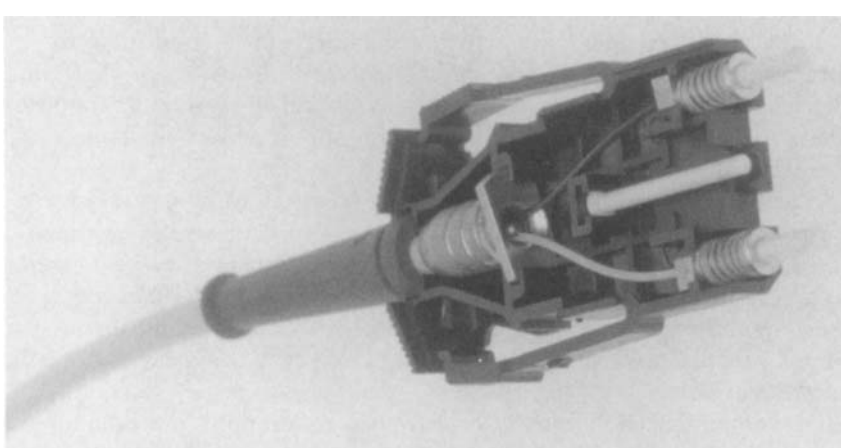


FIGURE 2.33 Duplex connector with housing cover removed. Strain relief rotated out of position for clarity. (Photo courtesy of IBM Corporation.)

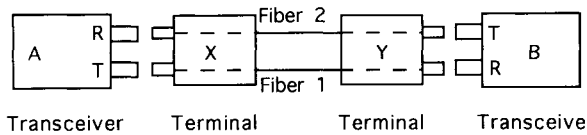


FIGURE 2.34 Fiber crossover in duplex jumper cables.

Referring to Figure 2.34, a jumper cable that interconnects transceivers A and B must join T in transceiver A to R in transceiver B via one fiber and must join T in transceiver B to R in transceiver A via a second fiber. Note that for proper linkage, fibers 1 and 2 must occupy different positions when viewed from the ends of the two terminals. Fiber 1 in terminal X occupies the right position, fiber 2 occupies the left position. In terminal Y, however, fiber 1 must occupy the left position and fiber 2 the right. Two fibers in a jumper cable must, in effect, cross. They do not actually cross but because they cross in effect the requirement is called "crossover". Crossover is needed between any two identical receptacles.

Duplex connectors are designed to perform fiber crossover automatically. To start, transceivers are designed so that emitter and receiver always occupy the same respective port (i.e., emmiter occupies the left port, receiver the right port). Second, terminals and receptacles are polarized. Polarized connectors have features on terminal and receptacle that allow insertion in only one orientation. Figure 2.30 (ESCON terminal) illustrates an example of polarizing features: the terminal nose is chamfered along its two top edges. An ESCON receptacle has identical chamfers. Third, fibers in a jumper cable always interconnect ferrules in the same position in a terminal. These three features, working as a system, ensure repeatable, error-free crossover.

Keys, chamfers, and other feature variations, can be used for error-free interconnection of MM and SM connectors in MM or SM links. Personalizing these types of variations is called *keying* a connector. Designing two variations of key/keyway or chamfer attributes easily accommodates MM and SM keying.

A pair of simplex connectors can also be polarized to achieve crossover. A variety of methods are possible. Some methods used are visual differentation and keying. For example, each simplex connector in a pair can be marked or colored differently from the other. Two different colors of housing can be used, or different color tape can be attached to each terminal in a pair. Receptacles are marked or colored to indicate keying. Another example is a key/keyway pair.

Data communication transceivers have two optical ports, one each for send and receive functions. Duplex connectors engage two ferrule/sleeve pairs in a single insertion operation, whereas a pair of simplex connectors engage each port one at a time. Engaging two ports places demands on duplex connectors

that are not placed on a pair of simplex connetors. The task of engagement is complicated by additonal sources of manufacturing variation. For example, the position of both send and receive ports within a transceiver vary. This variance places demands on a duplex terminal with a unitary body to position two ferrules for insertion. On the other hand, variance in terminal ferrule position places alignment demands on transceiver and receptacle packaging. A choice between a more tightly controlled transceiver package or terminal package variation arises. Tighter control of port position and receptacle dimensions eases requirements for tight control of ferrule position. This approach tends to reduce terminal cost, yet increases transceiver cost. The converse is also an option: tighter control of ferrule and terminal dimension variation eases requirements for receptable variation control, but terminal cost increases. Manufacturing variance also increases the difficulty of achieving a reliable mechanical coupling between terminal and receptacle in duplex connectors.

Connectors exhibit rough and fine guidance from initial insertion to final seating, respectively, to create reliable optical and mechanical joints. Fine guidance is accomplished by the ferrule/sleeve system or other joint system chosen: tight tolerances for sleeves, ferrules, and fibers ensure good optical coupling. Rough guidance is the process of bringing ferrule and sleeve into alignment so that fine guidance can take over. Rough guidance also positions fastener elements so that fastening can be performed.

Rough guidance prevents two conditions that must be avoided. First, each ferrule must not miss its sleeve opening and catch the edge at its entrance. Second, a ferrule must not seize in a sleeve during insertion. These two conditions are called stubbing and binding, respectively. Binding occurs from angular and lateral skew of ferrules with respect to sleeve axis. One means used to achieve rough guidance is tapering receptacle cavities and terminal bodies. A receptacle cavity tapers from a larger opening at its entrance to a smaller opening at the ports. A terminal body tapers from a smaller cross-section as its ferrule end to a larger cross-section at its cable end. The scheme provides gradually tighter alignment as a terminal enters a receptacle and positions ferrule and sleeve for insertion.

As discussed previously, the distance between port centerlines varies in data communication transceivers. One means used to accommodate the variance is to increase ferrule float. Float allows ferrules to shift during insertion so that they align with sleeve openings and reduce skew. For duplex terminals with unitary bodies, float is simply increased by design over their simplex counterparts. In duplexed simplex terminals, simplex ferrule float can be augmented by incorporating a spring element between terminal sub unit halves. A thin web of material integral to a terminal housing and formed during the injection molding process for the housing comprises the spring element. For example, illustrated in Figure 2.30 is an FCS duplex connector with an integral spring between simplex units that enhances float.

To improve the chances of engagement during initial insertion, terminals are designed to bias ferrules in a free state to their nominal, centered position and to a position free from skew. The centering action reduces chances for stubbing or binding between ferrule and sleeve. Centering action can be accomplished by the spring that is used for generating physical contact force. Another scheme is to incorporate chamfer pairs, one chamfer on the leading edge of a ferrule endface and one on the opening edge of a sleeve. These chamfers meet during initial engagement and force ferrule and sleeve to align through a shifting action from float.

Duplex terminals, like simplex terminals, must accommodate axial and off-axis forces transmitted to them by cable. Typically, a crimp and boot system are used. Recall that in simplex terminals a crimp captures strength members from a simplex cable for strain relief and a boot protects cables from off-axis loads. This approach works well for simplex cable, but what is done for duplex cable?

Figure 2.33 illustrates one approach for a unitary terminal and RDS duplex cable. In the figure, the crimp is shown rotated for clarity. Strength members are captured in a single crimp and the two fibers pass through the centre of the crimp rings. The crimp is captivated in the terminal body so that loads transmitted from the cable are transferred to the terminal body. A single boot guides the unitary duplex cable out of the housing.

Duplexed simplex terminals typically use strain-relieving features present in the simplex terminals. This approach generally requires that duplex cable be separated into two simplex units. RDS cable is separated by stripping the outer jacket to expose either two simplex fibers or two simplex cable subunits. Zipcord cable is separated by pulling apart the two halves at the web between them. In both cases, the point at which the duplex cable is divided into two subunits, called the bifurcation joint, must be protected. For example, at bifurcation joints, it is desirable to prevent propagation of tears in webs between zipcord halves and to prevent tears from developing in RDS outer jackets. One inexpensive approach is to apply adhesive and shrink tubing over the joint. Another, more robust approach is bifurcation kits that are commercially available.

2.6.8 Compatibility and Standards

Designing, developing, and tooling up a new, high-reliability fiber-optic connector is a lengthy and expensive task. Only a few connectors have been successfully launched and are commercially successful, in part because of the tremendous startup costs. Because demand for fiber-optic connectors is large, more than a few companies want to supply them. To meet demand, several vendors manufacture the same type of connector. A difficulty users face is having assurance that a connector from one vendor will operate reliably in their mission-critical applications when mated with the same type of

connector from a different vendor. To achieve this, users should be concerned with connector compatibility [29].

For connectors from a multiplicity of vendors to interoperate, they must do so without damaging each other. One may at first think that connectors must only be the same dimensionally in critical matting areas. The property of dimensions similarity is generally called intermateability. To some users, compatibility means intermateable connectors. Damage may be avoided by dimensional similarity, but a level of performance may not be guaranteed. An intermateable connector with mating halves manufactured by different vendors may not offer the same performance in ambient or mechanical environment halves from the same vendor. To some users, then, compatibility means intermateable by the previous definition plus a minimum level of repeatable performance. To still another group of users, compatiable may mean identical. The term compatibility is open to interpretation.

In an attempt to assure users of compatibility, standards bodies have adopted criteria for connectors. Vendors and users affected by proposed standards typically make up the standards committees. Fiber-optic connector standards typically define intermateability criteria and a minmimum level of performance. Connectors from two different vendors that meet an industry standard may not meet all users' definitions of compatibility, however.

There is another, indirect way a standard is created. A connector with desirable characteristics may become widely used. A connector's popularity and installed base may create a defacto standard. Defacto standards can be more powerful than formal standards because of the sheer number of users and number of installed connectors. The developer of a connector that has become a defacto standard may license design and technical know-how to other vendors so that they can manufacture the connector. Licensing gives users some assurance of compatibility across the manufacturing base, but it does not guarantee full compatibility. Although a licensed vendor may purchase know-how and manufacturing rights, there is no guarantee that the vendor will practice the know-how entirely or in the same way as the prime developer.

The best assurance that two connectors are compatible is an established specification for them based on requirements in the intended application. Connectors from prospective vendors should be tested to the specification. If more than one vendor is selected, cross-plugging tests between vendors' connectors should be performed to develop confidence that connectors are in fact compatible.

2.6.9 Connector Assembly

Several types of connectors are commercially available. The types differ in design but the processes for manufacturing them are similar in concept. In general, first the fiber is prepared for termination in a ferrule, which is

followed by completing a ferrule assembly, then packaging of the ferrule assembly into housings. The following text describes the processes in more detail.

A simplex termination is prepared by first removing the outer jacket, trimming the strength members, and removing buffer layers from a simplex fiber. All trimming is performed to specific dimensions for repeatable assembly. Duplex cable is first separated into simplex segments. It is important to trim two segments from one duplex cable close in length to eliminate extra bends that would occur in a longer segment. Fiber buffer layers are removed by sharp blades, much like electrical insulation is removed from copper wires. More precision is required for stripping optical fibers than for copper wires because tensile and fatigue strength of glass fibers are adversely affected by nicks or scratches on their surface. Fibers are completely stripped of buffer layers to leave a bare cladding layer (125 μm is typical).

The next step in terminal assembly is assembling a stripped fiber and a ferrule together. Before terminantion proceeds, packaging components that will not fit over a ferrule once attached to a fiber must be slid onto the fiber. To avoid damage to exposed fibers, packaging components are best placed on a cable before fibers are stripped. A ferrule is prepared by injecting a small amount of epoxy into the ferrule hole. For epoxy, shelf life and mixed age are important parameters to monitor. Monitoring mixed age is important because epoxy becomes more viscous as it sits out, which can cause fibers to break when inserted into epoxy holes. In addition, bond performance may degrade as epoxy ages. Before inserting into a ferrule hole filled with epoxy, a fiber is wiped clean with alcohol to improve adhesion. A sufficient length of fiber is stripped to leave a small portion protruding from a ferrule tip when fully inserted. Several fiber/ferrule assemblies are prepared and then "baked" in an oven to cure the epoxy. Some designs use fast-curing epoxies that cure at room temperatures to reduce assembly time and cost. After the epoxy has cured, the ferrule is ready for endface preparation.

Endfaces are prepared by first cleaving a fiber protruding from a ferrule close to its point of exit. After the extraneous piece of fiber is detached, a smooth, flat surface with sharp edges remains on the fiber stub that protrudes slightly from a ferrule. Ferrule endfaces with their protruding fiber stubs are next polished to achieve a high-performance optical interface. The process begins with light sanding with fine-grit sandpaper. Polishing proceeds with progressively finer grit lapping films and light ferrule pressure against the films. A typical grit progression is three to five steps from 30 μm dry to 0.3 μm wet. Care and control is required during at least the first polishing step to avoid generating cracks in protruding fibers. The end result of polishing is ferrule endfaces with mirror-like finishes.

Any polishing process employed must not leave scratches, pits, or chips across a fiber core. Examples of acceptable and unacceptable endface conditions for high-performance connectors are illustrated in Figure 2.35. Surface

damage on fiber cores is potentially detrimental for two reasons. First, surface damage scatters light at the optical interface and increases connection loss. Second, surface damage can become sites from which fractures propagate.

Two types of endface polish are common for precision ferrule terminals: flat polish and PC polish. Flat and PC polishing begin similarly but diverge to achieve endfaces that are shaped either flat or convex, respectively. Polishing for repeatable low-loss joints done in factory environments is typically

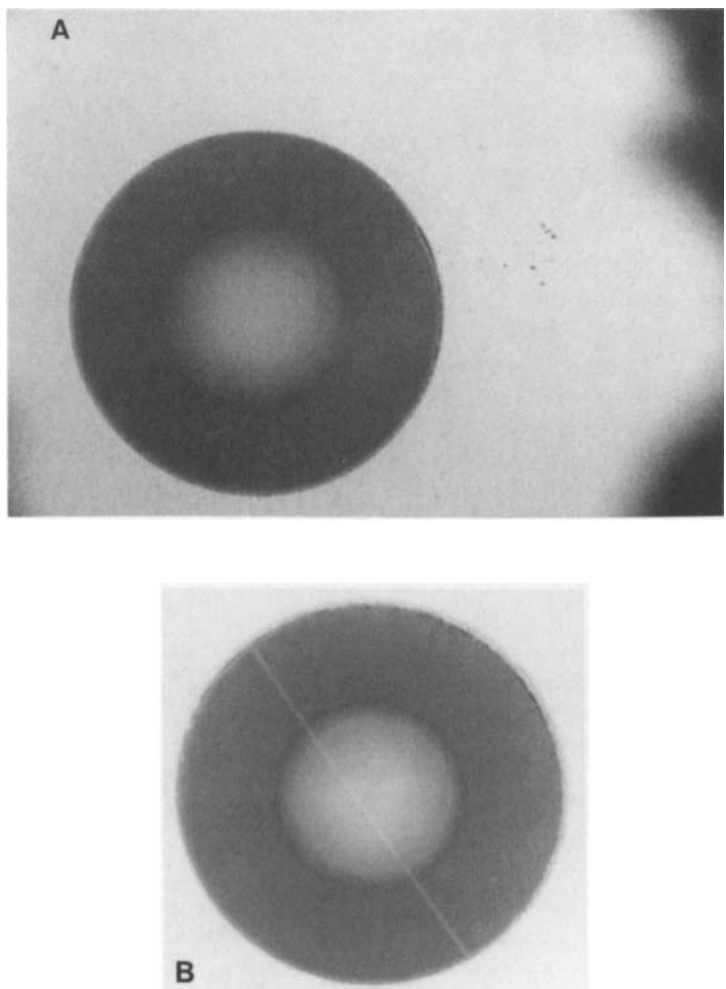


FIGURE 2.35 Examples of ferrule endface polish quality: (a) acceptable, and (b) typically unacceptable.

performed on specialized equipment that closely controls polishing parameters. Polishing field-terminated terminals is necessarily simplified to reduce cycle time, cost, and equipment required. A typical factory polishing process uses a ferrule fixture and an automated, movable platen. Ferrules are loaded into a fixture that resembles a plate with round openings to accept the ferrules and the endfaces are exposed uniformly out of one face of the plate. Polishing films are attached to the platen. The fixture is loaded into the machine with ferrule endfaces against the film/platen. The platen moves in a pattern that causes ferrules to polish in a figure-eight path that slowly rotates to cover all the film. A figure-eight path prevents any bias in endface shape that might result from polishing in only one direction. The platen moves instead of the ferrule fixture to prevent stress in the fibers attached to the ferrules.

A compliant mat such as a rubber pad is placed between polishing films and platen for the first polishing step. A mat reduces shock to protruding fibers that are prone to damage. As finer polish steps are taken, flat and PC polish processes diverge: the compliant mat is used throughout PC polishing; the mat is removed after the first polishing step for flat polish. When the mat is retained under films, ferrules causes the film to deform. The deformation is axisymmetric and has a curved cross-section. This deformation leaves a spherically shaped convex profile on ferrule endfaces [30]. It is desirable to permit a fiber to protrude slightly (on the order of micrometers) from its ferrule for PC joints to reduce return loss, as discussed previously. Protrusion is achieved during polishing processes when a glass fiber is harder than its ferrule. A convex spherically-shaped endface that is axisymmetric with respect to the ferrule is desirable for low-loss PC joints. Deviation of dome axis from ferrule axis is often called eccentricity.

After ferrule/fiber assemblies are prepared, they are assembled in terminal packaging. Cable strength members are crimped to strain relieve the fiber, springs are positioned to provide ferrule travel and preloading, and ferrules are placed into housings. The sequence of steps depends on which of the several terminal types is being assembled. Boot strain reliefs are also completed at this stage. In some applications, boots are filled with epoxy to form a bond between cable jacket and boot. Bonding and cable increases a boot's ability to absorb axial loads, which further insulates fibers from strains. For terminals on zipcord or RDS duplex cable, a breakout joint is installed to seal the cable at the bifurcation joint. Caps are installed over terminal faces to protect optical interface surfaces. Dust caps must protect the ferrule endface but not contact it, which could cause damage.

Following assembly, jumper cables are tested for acceptance. Acceptance testing is performed during production to ensure that key specifications are met before shipment. Acceptance tests are selected by weighing the ability to produce meaningful results against contribution to manufacturing cycle time. For example, connection and return loss measurements may be performed on every terminal because they yield useful information about fiber continuity

and ferrule endface condition, and these measurements are also easily performed. Thermal ship shock, however, is typically not performed on every cable because it is very time consuming. Time consuming tests like TSS are sometimes performed on a sample basis instead. If a terminal fails a test, it is either repaired or replaced.

Traceability of subcomponent manufacturing data such as lot or date code can be valuable should a cable assembly fail. To achieve traceability, cables are marked with information such as part number, date manufactured, and lot number, which in turn uniquely identifies subcomponent traceability data. Common means of identification include attaching printed labels or printing on cables directly. Cables must also be marked with mandated data such as the flammability rating.

2.6.10 Manufacturing Considerations

Optical fibers are brittle and must be handled and processed carefully to avoid fractures. Fractures are caused by applied stress imposed during ferrule and terminal assembly that concentrate at sites of surface damage and flaws. It is important to minimize damage to fibers during fiber stripping, endface polishing, fiber handling, and other activities. These nicks and scratches frequently are not sufficient to cause immediate fiber breakage: however, fiber strength is significantly reduced and subsequently applied loads can cause failure. Another source of stress comes from the fact that expansion rates between fibers, ferrules, epoxy, and other packaging materials are different. The difference in expansion rates causes stress to develop in these subcomponents when heated or cooled in assembly processes, shipping, and field use by differential thermal expansion (DTE) between subcomponents. Sufficient DTE from thermal excursions can generate enough stress to cause fiber fracture.

Whether a fiber that is assembled in a termination breaks because of DTE depends on several factors, such as the type of ferrule termination assembly process used and the type and number of coating layers used [31]. Failures of terminated fibers have been studied and categorized into two types. Both types occur in specific, respective precision ferrule termination types. One type occurs during cooling after epoxy cure by buckling. A second type occurs by tensile stress during repeated thermal excursions. This second type of failure most frequently happens at the point where a stripped fiber enters a ferrule hole. For termination of fibers with dual UV-cured coating layers—a soft inner layer and hard outer layer—it is important that fibers are inserted fully into a ferrule such that the unstripped buffer layers are against the ferrule hole opening. It is important to eliminate free spans of stripped fibre because they aggravate this second type of failure mechanism. Steps must be taken to reduce or eliminate bubble or void formation in epoxy as well. Two effective safeguards are (1) ensuring that ferrule holes are completely filled with epoxy,

and (2) removing dissolved gases in epoxy with a vacuum pump before the epoxy is used.

Some manufacturers and users of fiber-optic jumper cables (cables with terminals at both ends and generally under 100 m in length) require tensioning of cable strength members. Tensioning is removing slack in cable strength members before a section of cable is assembled into jumper cables. Slack is removed by pulling strength members at each end of a cable section. Typically, the tension is released after application so that no residual stress remains. Removal of slack reduces the amount of elogation that will occur in a cable under tension and protects an encased fiber at higher tensile loads. Tensioning, some practitioners feel, makes a jumper cable more robust.

Appendix A: Examples of Jumper Cable Specifications

Multimode Jumper Cable

Cable Assembly	
Parameter	**Requirement**
	Mechanical Integrity
Vibration	20 → 2000 → 20 Hz over 10 min, 10G peak acceleration, 4 complete cycles, 3 orthogonal axis
Impact Drop	EIA-455-2, light service
Impact Shock	500G, half sine wave, 5 ± 0.5 ms pulse duration, 3 blows in each of 6 directions
Connector Flexing	±90°, <45 cycles/min, 0.5 kg, 200 cycles
Twist	±90°, 4.0 kg, 10 cycles, 1 cycle/min minimum
Axial Pull	90 N minimum
Axial Push—Connector	20 N
Axial Push—Boot	20 N
Flammability	UL type OFNP (UL-910, NEC 770 6 (c))
	Physical
Insertion, Withdrawal Force	80 N maximum
Ferrule Diameter	2.499 ± 0.001 mm
	Optical
Attenuation per Mated Pair	0.5 dB maximum
Return Loss per Mated Pair	25 dB minimum
Repeatability	Δ0.3 dB maximum, 150 mating cycles
	Environmental
Nonoperating	−40° to 60°C, 5 to 95% RH
Operating	10° to 60°C, 5 to 80% RH
	Environmental Tests
Thermal Shock	−40° to 60°C, 200 cycles, fast transition
Corrosive Environment	GIT 336 hours (or salt spray, or Battelle FMG)
ATC	−10° to 70°C, 200 cycles, slow transition
Temperature and Humidity Aging	60°C, 85% RH, 480 hours
Particulates	Airflow dust (or Battelle AFI test, 300 hours)

Fiber, Bulk Cable	
Parameter	**Requirement**
	Mechanical Integrity
Compressive Strength	888.8 N ≤ 45 N/s on 100 mm of cable
Impact Resistance	4.5 Nm, 12.5 mm impact radius, 10 times
Cable Flex	2000 cycles @ ±90 deg, 2 kg
Minimum Bend Radius—No Load	30 mm
Minimum Bend Radius—80 N	50 mm
Minimum Bend Radius—800 N	75 mm
Flammability	UL type OFNP (UL-910, NEC 770)
	Optical
Attenuation (780 nm)	4.0 dB/km maximum
Bandwidth (780 nm)	400 MHz-Km minimum
Attenuation (850 nm)	3.0 dB/Km maximum
Bandwidth (850 nm)	500 MHz-Km minimum
Attenuation (1300 nm)	1.25 dB/Km maximum
Bandwidth (1300 nm)	500 MHz-Km minimum)
	Physical
Core Diameter	50 ± 3 μm
Numerical Aperture	0.2 ± 0.015
Core-to-Cladding Aperture	1 μm maximum
Cladding Diameter	125 ± 2 μm
Cladding Noncircularity	<2%
	Environmental
Nonoperating	−40° to 70°C, 5 to 100% RH
Operating	−10° to 70°C, 5 to 95% RH

Single-mode Jumper Cable

Cable Assembly	
Parameter	**Requirement**
	Mechanical Integrity
Vibration	20 → 2000 → 20 Hz over 10 min, 10G peak acceleration, 4 complete cycles, 3 orthogonal axis
Impact Shock	500G, half sine wave, 5 ± 0.5 ms pulse duration, 3 blows in each of 6 directions
Connector Flexing	±90°, <45 cycles/min, 0.5 kg, 200 cycles
Twist	±90°, 4.0 kg, 10 cycles, 1 cycle/min minimum
Axial Pull	90 N minimum
Axial Push—Connector	20 N
Axial Push—Boot	EIA-455-6
Flammability	UL type OFNP (UL-910, NEC 770-6 (c))
Rotational Pull	1 cycle, 45 deg, 80 N

Cable Assembly *continued*

Parameter	Requirement
	Physical
Insertion, Withdrawal Force	80 N maximum
Ferrule Endface Radius	20 ± 10 mm
Eccentricity	50 mm maximum
Core Concentricity	0.5 μm maximum
Ferrule Diameter	2.499 ± 0005 mm
	Optical
Attenuation per Mated Pair	1.0 dB maximum
Return Loss per Mated Pair	30 dB minimum
Repeatability	Δ 0.3 dB maximum, 150 mating cycles
	Environmental
Nonoperating	−40° to 70°C, 5 to 95% RH
Operating	0° to 60°C, 5 to 95% RH
Environmental Tests	
Thermal Shock	−40° to 70°C, 200 cycles, fast transition
Corrosive Environment	GIT 336 hours (or salt spray, or Battelle FMG)
ATC	−10° to 70°C, 200 cycles, slow transition
Temperature and Humidity Aging	85°C, 80% RH, 500 hours
Particulates	Airflow dust (or Battelle AFI test, 300 hours)

Fiber, Bulk Cable

Parameter	Requirement
	Mechanical Integrity
Compressive Strength	500 N/cm, 10 min, 45 N/s maximum
Impact Resistance	4.5 Nm, 150 mm drop height, 12.5 mm impact radius, 10 times
Cable Flex	2000 cycles @ ±90 deg, 2 kg
Minimum Bend Radius—No Load	30 mm
Minimum Bend Radius—80 N	50 mm
Minimum Bend Radius—800 N	75 mm
Flammability	UL type OFNP (UL-910, NEC 770-6(c))
	Optical
Attenuation (1300 nm)	0.8 dB/km maximum
Zero Dispersion Wavelength	1300 to 1322 nm
Dispersion 1285 to 1330 nm	3.5 ps/(nm•km) maximum
Zero Dispersion Slope	0.95 ps/(nm^2•km) maximum
Cutoff Wavelength	1260 nm maximum

Fibre, Bulk Cable *continued*	
	Physical
Proof Strength	345 MPa minimum
Mode Field Diameter	8.7 to 10 μm
Core-to-Cladding Offset	1 μm maximum
Cladding Diameter	125 ± 2 μm
Cladding Noncircularity	2% maximum
	Environmental
Nonoperating	−40° to 70°C, 5 to 95% RH
Operating	−10° to 70°C, 5 to 95% RH

References

1. D. Gloge (1971). Weakly guiding fibers. *Applied Optics* **10**, 2252–2258.
2. D. M. Bloom, L. F. Mollenauer, C. Lin, D. W. Taylor, and A. M. DelGaudio (1979). Direct demonstration of distortionless picosecond-pulse propagation in kilometer-length optical fibers. *Opt. Lett.* **4**, 297–299.
3. P. Lemaire. *AT&T Bell Labs. Murray Hill, private communication.*
4. E. A. Guillemin (1963). *Theory of Linear Physical Systems.* New York: John Wiley & Sons.
5. S. Ramo, J. R. Whinnery, and T. van Duzer (1994). *Fields and Waves in Communication Electronics*, 3rd ed. New York: John Wiley & Sons.
6. T. Okoshi (1982). *Optical Fibers.* New York: Academic Press.
7. J. Senior (1985). *Optical Fiber Communications: Principles and Practice.* Englewood Cliffs, NJ: Prentice Hall.
8. E. G. Neumann (1988). *Single-Mode Fibers: Fundamentals.* Berlin: Springer-Verlag.
9. H. Kogelnik (1969). *Coupling and Conversion Coefficients for Optical Modes in Quasi-Optics.* N.Y.P. Press, Microwave Research Institute Symposia Series, 333–347.
10. D. Marcuse (1977). Loss analysis of single-mode fiber splices. *Bell System Technical Journal* **56**, 703–718.
11. S. Nemoto and T. Makimoto (1979). Analysis of splice loss in single-mode fibers using a Gaussian field approximation. *Optics and Quantum Electronics* **11**, 447.
12. C. M. Miller and S. C. Mettler (1978). A loss model for parabolic-profile fiber splices. *Bell System Technology Journal* **57**, 3167–3180.
13. K. O. Möller (1988). *Optics.* Mill Valley, CA: University Science Books.
14. J. P. Powers (1993). *An Introduction to Fiber Optic Systems.* Homewood: Aksen Associated Incorporated Publishers.
15. R. D. Maurer (1973). Glass fibers for optical communication. *Proceedings of IEEE* **61**, 452–462.
16. H. Murata (1988). *Handbook of Optical Fibers and Cables.* New York: Marcel Dekker.
17. D. B. Keck, R. D. Maurer, and P. C. Schultz (1973). On the ultimate lower limit of attenuation in glass optical waveguides. *Applied Physics Letters* **22**, 207–309.
18. T. Izawa, N. Shibata, and A. Takeda (1977). Optical attenuation in pure and doped fused silica in the IR wavelength regions. *Applied Physics Letters* **31**, 33–35.
19. I. L. Fabelinski (1968). *Molecular Scattering of Light.* New York: Plenum.
20. J. Hecht (1987). *Understanding Fiber Optics.* Indianapolis, IN: H. W. Sams.
21. D. G. Baker (1985). *Fiber Optic Design and Applications.* Reston, VA: Reston Publishing Company.

22. C. Yeh (1990). *Handbook of Fiber Optics: Theory and Applications.* San Diego: Academic Press.
23. D. R. Roberts, C. Enrique, and M. Kennedy (1991). Fiber construction for improved mechanical reliability. *Proceedings of SPIE* **1366**, 129–135.
24. P. R. Dickinson (1992). Evolving fire retardant material issues: A cable manufacturer's perspective. *Fire Technology* **28**(4), 345–368.
25. W. C. Young and D. R. Frey (1988). *Optic Fiber Telecommunications II.* Boston: Academic Press.
26. E. Sugita, K. Iwasa, and T. Shintaku (1986). Design for push–pull coupling single fiber connectors featuring zirconia (ZrO_2) ceramic ferrules. *12th European Conference on Optical Communication.* Telefonica. Madrid, Spain, 141–144.
27. E. Sugita, R. Nagase, K. Kanayama, and T. Shintaku (1989). SC-type single-mode optical fiber connectors. *Journal Lightwave Technology* **7**(11), 1689–1696.
28. J. R. Meyer-Arendt (1989). *Introduction to Classical and Modern Optics*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
29. J. E. Denny (1993). Optical connector standards and licenses. *Fiber Optic Product News* **3**, 27.
30. T. Szotstak (1987). New ST connector design eliminates fiber pistoning problems. *Connection Technology*, December, 22–26.
31. B. G. LeFevre, W. K. King, A. G. Hardee, A. W. Carlisle, and K. B. Bradley (1993). Failure analysis of connector terminated optical fibers: Two case studies. *Journal of Lightwave Technology* **11**(4), 537–541.

Chapter 3

Light Sources, Detectors, and Basic Optics

A. Alex Behfar
IBM Corporation
Microelectronics Division
Hopewell Junction, New York

Janet L. Mackay
IBM Corporation
Microelectronics Division
Essex Junction, Vermont

3.1 Introduction

We live in the electronic age. Our computers, networks, and control systems are electronic. Input from the world, if not already in the form of an electrical signal, is first converted into one (and usually further converted to a digital signal). Next, information processing is performed; then, if need be, outgoing signals are converted back from an electric signal into one that humans can understand.

Control systems, networks, and computers operate on electrical signals. This simple fact explains why a wire is the most common and most straightforward link between two computers. However, for a long distance, high data rate, or EMI-sensitive application, it may be advantageous to convert an electrical signal into an optical one, transmit the optical signal over optical fiber, then convert it back into an electrical signal at the other end. These applications are becoming more prevalent, which is one reason why this book was written.

Optoelectronics for Data Communication

Light sources and detectors are key components of an optoelectronic data link. A light source converts an electrical signal into an optical one; a detector converts an optical signal into an electrical one. Sources and detectors are the link between the electronic world of the control systems and the optical world of the fiber.

Electronic devices (transistors) and optoelectronic devices (sources and detectors) both operate by virtue of the unique properties of semiconductors. Silicon rules the electronics world primarily because it is amenable to high-volume production. The enormous volume of silicon device production makes silicon by far the least expensive semiconductor. However, silicon has an unfortunate property that makes it less than ideal for use in optoelectronic devices, as will be explained later in this chapter. This property makes it extremely difficult to use silicon for light sources; so light-emitting diodes (LEDs) and laser diodes are made from other semiconducting materials. It turns out that silicon will work as a light detector, just not as efficiently as other semiconductors. So, depending on the application, a silicon detector or another type of detector may be chosen.

In this chapter, we start with some basic optics and discuss the optical properties of semiconductors in general and a few common semiconductor materials in particular. Next, we discuss the p–n junction diode—the basic structure is common to all sources and detectors. Then, we look at each type of optoelectronic device in turn, photodiode detector, avalanche photodiode detectors, LED, and several different lasers.

3.2 Basic Optics

In this section we review some fundamental optics required for a basic understanding of lasers, LEDs, and detectors. We begin by looking at Fresnel and Bragg reflections—the two most common methods of obtaining reflection within semiconductor devices. Next, we look at waveguiding in a multilayer structure, and finally, we look at the effective index technique that approximates a two-dimensional waveguide to one that is one-dimensional. The objective here is to introduce these concepts without delving too deep into the theory. However, for those readers who need to understand the theoretical background with derivation from first principles, references are provided in each subsection.

3.2.1 Fresnel Reflection

You are walking in front of a store window and you see your reflection. Although faint, you still see your reflection. This type of reflection occurs when light travels from a medium of one index of reflection to another, like air to glass. Consider light traveling from medium 1 with index of refraction

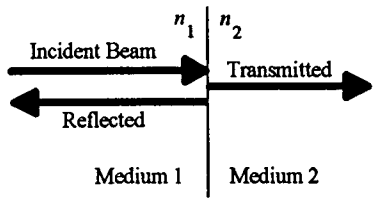


FIGURE 3.1 Incidence of light at an interface.

n_1 to medium 2 with n_2. The reflectivity and transmission are governed by the Fresnel equation [1, 2]. In the case where light is incident at right angles to the interface between the media (see Fig. 3.1), the Fresnel equation can be written as

$$R = \frac{(n_2 - n_1)^2}{(n_2 + n_1)^2} \tag{3.1}$$

So, if medium 1 is air ($n_1 \approx 1$) and medium 2 is glass ($n_2 = 1.5$), we find the reflectivity, R, to be 0.04 or 4% of the incident light. Note that the transmitted light and the reflected light are in line with the incident beam and are shown in Figure 3.1 (with a small offset for clarity). When light travels from one medium to another medium of higher index of refraction, no phase change is observed in the reflected beam, as is the case for the preceding example. However, if light travels from one medium to another one of lower index, then the reflected beam experiences a phase change of 180 deg.

3.2.2 Bragg Reflection

Consider Figure 3.2 in which a beam of light with wavelength λ_0 is incident on a medium at right angles. The medium is made of a stack of layers of high, n_h, and low, n_l, refractive index layers on a substrate with an index of n_s. The stack is terminated with a layer of index n_h, with the outside medium being of index of refraction n_0, where $n_h > n_l > n_0$. The respective thicknesses of these layers are $t_h = \lambda_0/4n_h$ and $t_l = \lambda_0/4n_l$. Because the wavelength of light in a material is equal to the wavelength of light in vacuum divided by the index of refraction of the material, these layers are referred to as quarter-wavelength or quarter-wave films. The reflectivity is given by the following equation, derived from Reference 1:

$$R = \left[\frac{n_0 - (n_h/n_l)^{2p}(n_h^2/n_s)}{n_0 + (n_h/n_l)^{2p}(n_h^2/n_s)}\right]^2 \tag{3.2}$$

where $(2p + 1)$ is the number of individual layers in the stack. The reflectivity increases as the number of quarter-wave pairs is increased. This is an example of a first-order Bragg reflector.

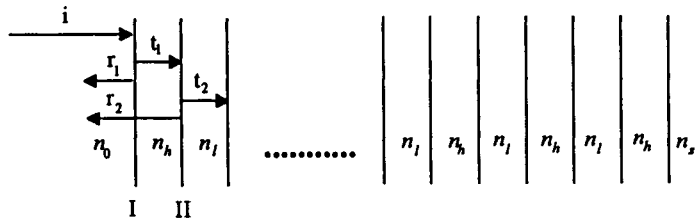


FIGURE 3.2 Incidence of light on a Bragg reflector.

Let us now look at where reflections occur and how reflected beams of light are perfectly in phase when they leave the surface of the stack, or interface I. As the incident light, i, goes from n_0 to n_h at interface I in Figure 3.2, the reflected component, r_1, experiences no phase change because $n_0 < n_h$. Meanwhile, the transmitted component, t_1, travels to interface II and gives rise to r_2 and t_2. But this time, because $n_h > n_l$, r_2 experiences a 180-deg phase shift. Now to see why r_1 and r_2 will be in phase as they leave interface I. By traveling from interface I to II, t_1 experiences a quarter-wavelength—or a 90-deg—phase shift because it has traveled in a film of a quarter-wave thickness. r_2 is generated with a 180-deg phase shift from that of t_1 at interface II. Then, r_2 travels from interface II to I and sees a quarter-wavelength phase shift. So, in total, a 360-deg phase shift occurs and that is why r_1 and r_2 are in phase as they leave interface I. Continuing the same exercise for other components, we find the same exact result.

Example

A multilayer structure with a GaAs substrate is placed in air and a beam of light of wavelength $\lambda_0 = 0.8265$ μm is incident at right angles to the structure. The multilayer structure is made of quarter-wave pairs of $Al_{0.2}Ga_{0.8}As$ and $Al_{0.6}Ga_{0.4}As$. We calculate the thickness of both quarter-wave layers and the reflectivity of the stack for 1, 2, 3, 4 and 5 quarter-wave pairs, assuming $n_{GaAs} = 3.6566$, $n_{Al_{0.2}Ga_{0.8}As} = 3.4859$, and $n_{Al_{0.6}Ga_{0.4}As} = 3.2194$. Here, the

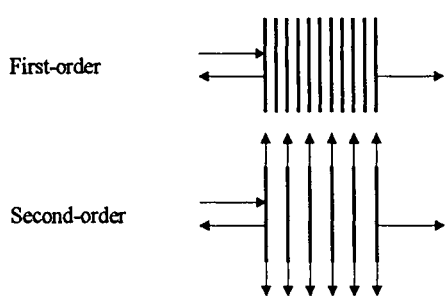


FIGURE 3.3 Illustration of first- and second-order Bragg reflection.

high-index layer is $Al_{0.2}Ga_{0.8}As$ and the low-index layer is $Al_{0.6}Ga_{0.4}As$. The thicknesses of the quarter-wave layers are

$$t_h = t_{Al_{0.2}Ga_{0.8}As} = \frac{\lambda_0}{4n_{Al_{0.2}Ga_{0.8}As}} = \frac{0.8265\ \mu m}{4 \times 3.4859} = 0.0593\ \mu m, \text{ and}$$

$$t_l = t_{Al_{0.6}Ga_{0.4}As} = \frac{\lambda_0}{4n_{Al_{0.6}Ga_{0.4}As}} = \frac{0.8265\ \mu m}{4 \times 3.2194} = 0.0642\ \mu m,$$

The reflectivity can be found using Eq. (3.2):

$$R = \left[\frac{1 - (3.4859/3.2194)^{2p}\,(3.4859^2/3.6566)}{1 + (3.4859/3.2194)^{2p}\,(3.4859^2/3.6566)}\right]^2$$

$$= \left[\frac{1 - 1.0828^{2p} \times 3.3232}{1 + 1.0828^{2p} \times 3.3232}\right]^2$$

and for $p = 1, 2, 3, 4$, and 5, we find:

p	R
1	0.35
2	0.41
3	0.47
4	0.53
5	0.58

As the thickness of the layers in the stack is changed from quarter-wave to half-wave, second-order Bragg reflection is obtained. Figure 3.3 shows how first- and second-order Bragg reflectors differ. We have already discussed first-order Bragg reflectors and saw how we obtain reflected and transmitted components in line with the incident beam. The second-order Bragg reflector has the two components of the first-order Bragg reflector, but adds two extra perpendicular components: one in the upward direction and another downward. As we will see later on in this chapter, Bragg reflectors provide a very useful way of incorporating mirrors in the semiconductor devices.

3.2.3 Multilayer Waveguides

Here, we introduce one of the most fundamental requirements of a semiconductor laser: waveguiding. A waveguide can be formed by sandwiching one layer—known as the *core*—between two layers with lower index of refraction—known as *cladding layers*. An example of such a three-layer waveguide is shown in Fig. 3.4a, where $n_2 > n_3 > n_1$.

The index of refraction is shown in Fig. 3.4b. Casey and Panish [3] give a detailed account of obtaining the eigenvalues, or allowed modes, of such a waveguide as well as the corresponding eigenfunction, or modal intensity.

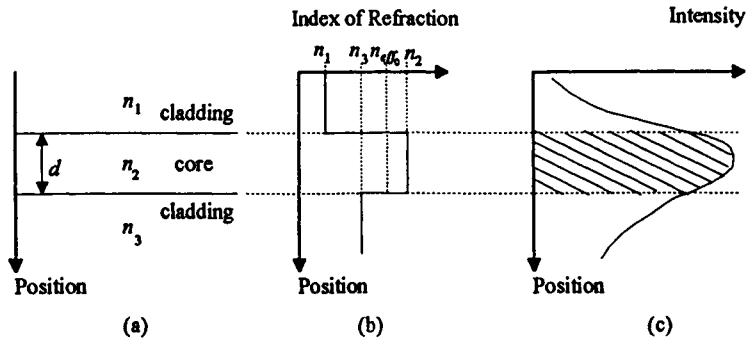


FIGURE 3.4 Three-layer waveguide.

Assuming that only the light of one wavelength will be present in the waveguide, Fig. 3.4b schematically shows the only eigenvalue of this structure, indicated by n_{eff_0}, with the associated modal intensity shown in Fig. 3.4c. Generally, the number of allowed modes increases with increasing core thickness, d. The number of allowed modes also increases with larger values for $(n_2 - n_1)$ and $(n_2 - n_3)$.

A waveguide parameter commonly used in semiconductor lasers is the confinement factor. The confinement factor is the ratio of the light intensity in the core (indicated by the shaded area in Fig. 3.4c) to the total light intensity (shaded and unshaded regions under the curve). Reference 3 also covers the confinement factor in detail. The confinement factor will be used in our discussion on lasers (subsection 3.9.1).

3.2.4 Near-Field and Far-Field

The intensity versus position curve shown in Fig. 3.4c is an example of a near-field plot. The near-field is a profile of the light intensity at the surface where the guided beam emerges from a waveguide. The near-field is usually shown as a two-dimensional image or one-dimensional plots of intensity versus a given direction (normally x and y) on the surface where light exits the waveguide.

The far-field is a Fourier transform of the near-field. It is the dispersion of a beam of light as it emerges from a waveguide. The far-field is generally shown as a one-dimensional plot of intensity as a function of angle in a vertical or horizontal direction.

3.2.5 Effective Index Technique

In our discussion of LEDs and semiconductor lasers, we will encounter waveguide cross-sections such as that shown in Figure 3.5, where

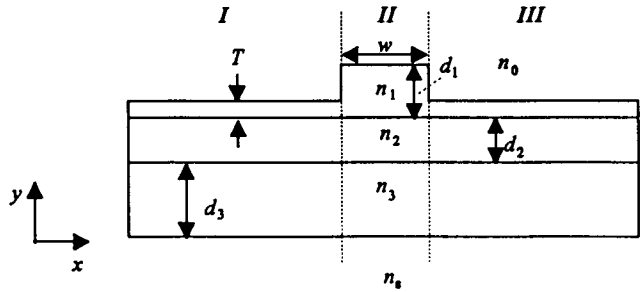


FIGURE 3.5 Ridge waveguide.

$n_s > n_2 > n_1 = n_3 > n_0$. Because the index of refraction varies in two dimensions, numerical techniques need to be employed to solve this problem accurately. However, there is an approximation called the effective index technique that gives reasonably good results [4].

The effective index method takes the two-dimensional waveguide and breaks it up into several one-dimensional waveguides. The one-dimensional waveguides are individually solved and the eigenvalues found. Then the eigenvalues are combined to form a virtual one-dimensional waveguide representation of the two-dimensional waveguide. Finally, the virtual one-dimensional waveguide is solved to get the eigenvalue and modal intensity. We can see how this works by taking the two-dimensional waveguide in Fig. 3.5 and breaking it into three segments, *I*, *II*, and *III*.

Figure 3.6 shows the index of refraction of these segments as a function of position. Note that because segments *I* and *III* have identical cross-sections, they result in the same position versus index plot, as shown in Figure 3.6a. With the thicker layer of material having refractive index n_1, segment *II* gets its own plot of position versus index (Fig. 3.6b). The effective index for each

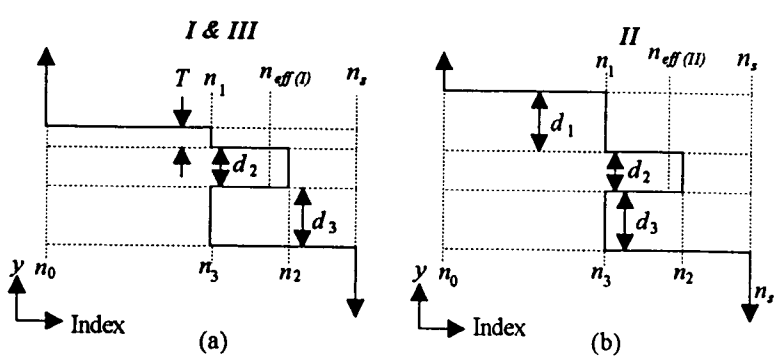


FIGURE 3.6 Position versus index plots for cross-sections in the ridge waveguide.

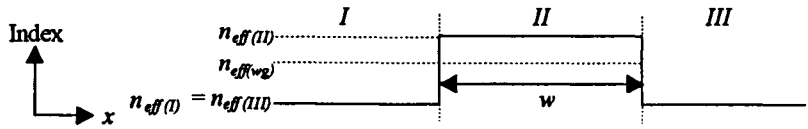


FIGURE 3.7 Effective index technique representation of the ridge waveguide.

of the guided modes in the two one-dimensional waveguides are indicated in Figure 3.6, a and b with $n_{\text{eff}(I)}$ and $n_{\text{eff}(II)}$, where $n_{\text{eff}(II)} > n_{\text{eff}(I)} = n_{\text{eff}(III)}$.[1] $n_{\text{eff}(I)}$ is smaller than $n_{\text{eff}(II)}$ because the optical mode "sees" more of the material with low index of refraction, n_0 (material with index n_2 is closer to material with index n_0 in *I* compared with *II*).

With effective indices of refraction for regions *I*, *II*, and *III* in hand, we now represent the two-dimensional waveguide of Figure 3.5 with its "effective" one-dimensional representation, as shown in Figure 3.7. The same process of finding an eigenvalue is performed for the waveguide of Figure 3.7 and results in effective index $n_{\text{eff(wg)}}$ that is an approximation to the effective index of the mode that is guided by the two-dimensional waveguide. Although the most accurate results can be found using a two-dimensional numerical calculation of the eigenvalue, for many situations, the preceding technique performs an adequate approximation with far less complexity.

3.3 Optical Processes in Semiconductors

3.3.1 Semiconductors

A semiconductor is a crystalline solid. Some of the basic properties of semiconductors may be understood by recalling atomic properties. An atom has distinct energy states; an electron is allowed to occupy one of those states. Each state has a particular energy. An electron within the atom is not allowed to be at just any energy, but only at one of the energies of allowed states. Each state has room for two electrons, after which it is full, and any additional electrons must occupy other states. In the ground state, if one imagines "adding" the right number of electrons to fill the allowed states, the lowest energy states are filled first. The valence state (the highest energy state that is filled in the ground state of the atom) may be partially filled (one electron) or filled (two electrons). By using an electrical discharge or by illuminating

[1] A description of how to find eigenvalues is well beyond the scope of our discussion. The objective here is to familiarize the reader with the effective index technique in a qualitative manner. A reader who needs a more detailed description on finding eigenmodes of multilayer structures is referred to K.-H. Schlereth and M. Tacke, *IEEE J. Quantum Electron. QE-26, 627 (1990).*

the atoms at a suitable wavelength, electrons may be excited to higher energy states. When an electron subsequently falls back to a lower energy state, the energy difference is given off as a photon (light).

In a crystalline solid, individual atoms are brought together in close proximity in an ordered array. The energy states of neighboring atoms intermix, and new electron energy states, characteristic of the whole crystal (not just a collection of independent atoms), form. This is shown schematically in Figure 3.8. Whereas isolated atoms have a set of exact allowed energies, it turns out that crystals have "bands" of allowed energies. An electron is not allowed at an energy that falls in between bands. Within a band, there is a range of allowed energies. In the ground state, the lowest bands fill first, then higher energy bands. The highest energy band that contains electrons is called the valance band.

Crystals can be classified as one of three types: insulator, metal, or semiconductor. What distinguishes a metal is that its valence band is only partly filled with electrons, even at zero temperature. The energy of the last electron at zero temperature is (loosely speaking) the Fermi energy. (Note: This is not a rigorous definition of Fermi energy, and only applies to a metal.)

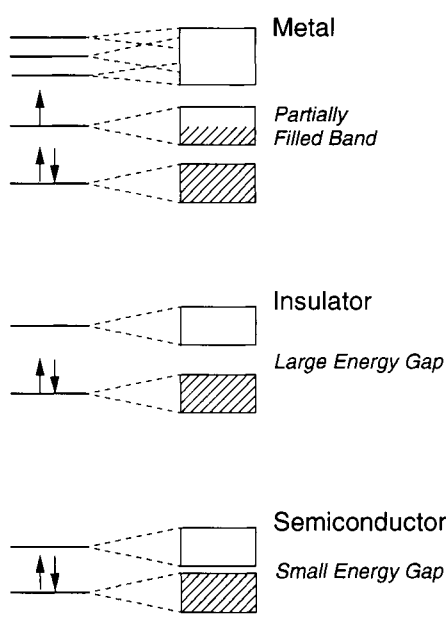


FIGURE 3.8 Schematic diagram of energy bands in crystals. Left side of figure represents energy levels in the constituent free atoms; right side represents energy bands in the crystalline solid. Each arrow represents an atomic electron with either up or down spin. The shaded areas on the right represent large numbers of electrons in the solid. A metallic crystal is shown at the upper right; an insulating crystal in the middle right, and a semiconducting crystal is at the bottom right. All materials are represented in their ground states.

This means that empty electron states are available at energies only infinitesimally higher than the Fermi energy, and that only a very slight energy above the ground state would need to be added to the crystal to cause some electrons to move from states just below the Fermi energy to those just above. In fact, at any nonzero temperature, some states just below the Fermi energy will always be empty, and some states just above the Fermi energy will always be partially occupied as a result of thermal energy.

Insulators and semiconductors, in contrast to metals, have no partially filled bands at zero temperature. All bands are either completely full or completely empty at zero temperature. The highest filled band is the valence band. The next highest band (which is the lowest energy empty band) is the conduction band. The difference in energy between the top of the valance band and the bottom of the conduction band is called the bandgap. It is usually measured in eV. The bandgap of a semiconductor (or insulator) is perhaps its most important single property for understanding its electronic and optoelectronic behavior.

What distinguishes an insulator from a semiconductor is the size of the bandgap. A semiconductor has a relatively small bandgap, so that at room temperature, a few electrons will have enough thermal energy to occupy the conduction band (leaving an equal number of empty states in the valance band). An insulator has a large bandgap, so that at room temperature, no electrons can be thermally excited to the conduction band. The cutoff is somewhat arbitrary, but a material with a bandgap below about 3 eV is usually considered a semiconductor, and a material with a higher bandgap is considered an insulator.

To understand why metals conduct electricity easily, pure semiconductors only poorly, and insulators hardly at all, consider the following analogy. Imagine a party at a house with two floors, as shown in Figure 3.9. (The first floor is the valence band, the second floor the conduction band and the party-goers are electrons.) First, imagine that the first floor is about half full of people (with the second floor empty). Potato chips are brought to the left side of the room and people flock to the left like electrons in a metal when a voltage is applied across it. Now, imagine that the first floor is completely packed with people and the second floor is inaccessible to the party-goers. When the chips are brought out, there is no room to move, so hardly anyone gets chips (Fig. 3.9, a and b). In an insulator, the valance band is completely full, so the electrons do not have "room" to move when a voltage is applied.

Lastly, imagine that the second floor is available. As long as a few people have enough energy to go to the second floor, there will be at least a little room to move around on the first floor. Imagine that there are 997 people on the first floor and 3 on the second. If chips are brought to the right side of both rooms, the three people on the second floor will move right. The 997 people on the first floor will also move right (Fig. 3.9, c and d). However,

instead of keeping track of those 997 people, we could instead track the three "holes" of empty space in the first floor. The holes will move left.

The unoccupied states in a semiconductor valance band really are called holes. Because they move opposite to electrons in an electric field, they have positive charge. If an electron and a hole come together, they can recombine and annihilate each other. This results in one fewer electron in the conduction band and one fewer hole in the valence band. Following our analogy, if a second-floor party-goer comes downstairs, there is one fewer "hole" on the first floor and one fewer "electron" on the second floor. When an electron

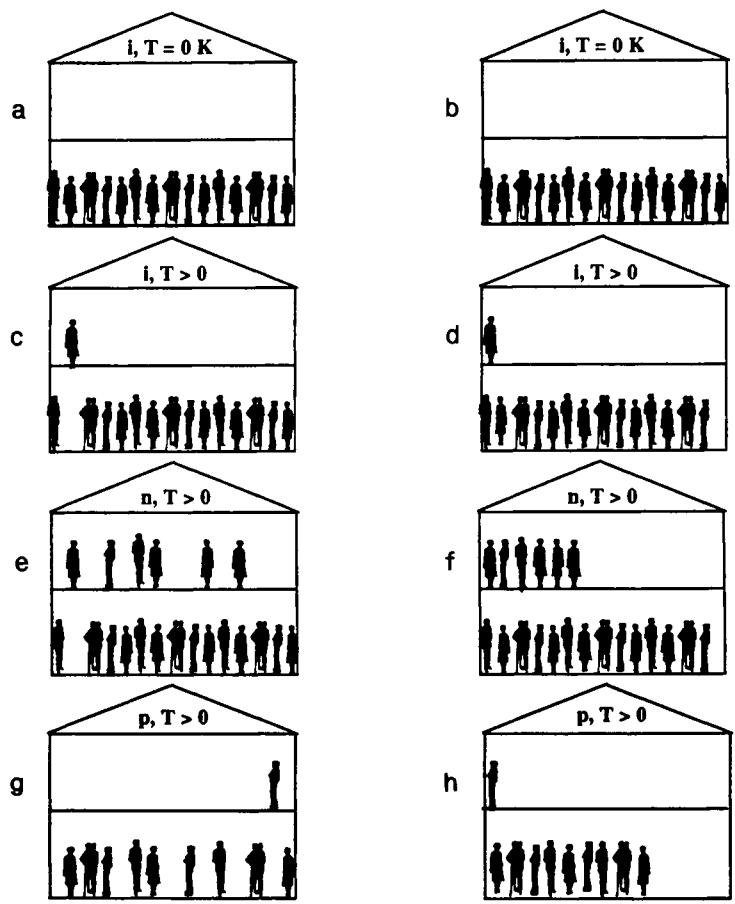


FIGURE 3.9 Party analogy to band theory. In (a), the floor is full and nobody can move in response to the potato chips brought out on the left (b). In (c), a few people have had enough energy to go to the second floor so that when the potato chips are brought out (d), the people move left and the "holes" move right. In (e), there are "extra" people on the second floor, analagous to an *n*-type semiconductor. In (f), these "extra" people are free to move. In (g) and (h), the "holes" on the first floor are free to move, analagous to a *p*-type semiconductor.

and a hole recombine, the energy given off (which is roughly equal to the bandgap energy) can be given off as a photon. Electron–hole recombination is the way LEDs and laser diodes generate light. Semiconductor detectors work in the opposite way, by converting photons into electron–hole pairs.

Through this crude analogy, one can visualize holes and the conduction behavior of metals, insulators, and pure semiconductors. From now on, we will focus our attention on semiconductors.

There is one more attribute of energy bands in semiconductors that needs to be explained. In an atom, different energy states are distinguished by their angular momentum, which is quantized. It would be meaningless to talk about the linear momentum of an electron in an atom—if the electron moved in a linear fashion relative to the nucleus, it would no longer be part of the atom. However, in a semiconductor (or metal), the crystal itself is very large compared to the atomic cores, and valence and conduction states to extend over the entire volume of the crystal. Electrons can have linear momentum; they must move from one end of the crystal to the other in response to an electric field. In solid state physics, it is conventional to use the wavevector $\mathbf{k}$ instead of linear momentum $\mathbf{p}$. They are proportional: $\mathbf{k} = 2\pi\mathbf{p}/h$, where h is Plank's constant. Figure 3.10 shows band diagrams for two different semiconductors. It turns out that in essentially all semiconductors, the highest energy state at the top of the valence band (which is the lowest energy state for holes) is a zero-momentum state or $\mathbf{k} = 0$. However, the state at the bottom of the conduction band (the lowest energy electron state) is sometimes a zero-momentum state and sometimes not. If the bandgap in a semiconductor is between the zero-momentum lowest energy hole state and a zero-momentum electron state, the bandgap is said to be direct. A semiconductor with a direct bandgap is called a direct semiconductor, but if the bandgap is between a zero-momentum hole state and a nonzero-momentum electron state, it is called an indirect bandgap and the semiconductor is an indirect semiconductor. The energy difference between the top of the valence band (which is zero momentum) and the lowest energy zero-momentum state is the direct bandgap. For indirect semiconductors, the indirect bandgap is always smaller than the direct bandgap. Silicon's (indirect) bandgap is 1.12 eV, and its direct bandgap is 2.5 eV (at 300 K).

The distinction between direct and indirect bandgaps is very important when considering photoemission in semiconductors. If an electron and a hole both have zero momentum, they can easily recombine. But if the electron has some momentum and the hole does not, conservation of momentum dictates that they cannot simply recombine. (In order to recombine, sometimes they can get the required momentum from scattering off the lattice, but this is a relatively slow process and it generates heat.) Because of the "slow" electron–hole recombination in indirect semiconductors, they are not used for light-emitting devices. This is why there are no commercial silicon LEDs or lasers. Direct semiconductors often used for photoemitting devices (as well as

detectors) include GaAs, InAs, and InP, and the alloys AlGaAs and InGaAsP. Nevertheless, indirect semiconductors are still good absorbers of light at the "right" wavelengths and silicon photodetectors are used in many applications.

3.3.2 Absorption and Emission of Photons

3.3.2.1 Absorption

When a photon enters a material, it may be absorbed or transmitted, depending on the electron energy levels of the material and the energy of the photon. The energy of a photon, E_p, is related to its frequency, v, and wavelength in vacuum, λ_0, by:

$$E_p = hv = \frac{hc}{\lambda_0} \tag{3.3}$$

where h = Plank's constant and c = speed of light. If E_p is equal to the difference in energy between a filled energy state of the material and an empty

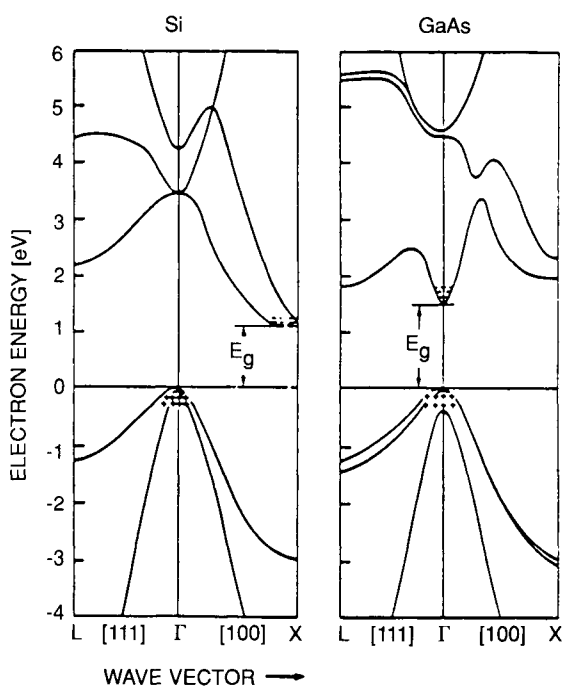


FIGURE 3.10 Simplified band diagrams of Si and GaAs. The vertical axis is electron energy. The horizontal axis is wavevector **k**, where $\mathbf{k} = [2(\pi)/h]p$. Electrons are allowed only at the (**E**, **k**) values which lie on one of the curves in the diagram. Si is a direct semiconductor and GaAs is indirect. (From S. M. Sze (1981). *Physics of Semiconductor Devices*, 2nd ed. John Wiley & Sons, New York, p. 13.)

one, the photon may be absorbed. Recall that the electrons are "forbidden" to occupy the bandgap. So the lowest energy transition that an electron can make is from the top of the valence band at E_V to the bottom of the conduction band at energy E_C. Higher energy transitions are also allowed, as there are many empty states available above E_C and many filled states below E_V. Therefore, absorption takes place when

$$E_p \geq E_g \tag{3.4a}$$

or, equivalently, when

$$\lambda_0 \leq \lambda_g \tag{3.4b}$$

where λ_g is defined as

$$\lambda_g = \frac{hc}{E_g} \tag{3.4c}$$

Table 3.1 shows a list of some commonly used optoelectronic materials, their bandgaps, λ_g, lattice constant, and refractive index. The last two materials listed, AlGaAs and InGaAsP, are alloys. Their properties depend on the relative concentrations of the elements, so only ranges are listed in the table.

3.3.2.2 Emission

A material in an excited state (not its ground state) will give off heat, light, or both as it falls to a lower energy state. Electron energy transitions may be classified as nonradiative (no photons created) or radiative. A radiative transition results in the emission of a photon; the photon energy is equal to the difference in energy between the higher and the lower state.

TABLE 3.1 Alloy semiconductors may be direct or indirect, depending on the exact alloy composition

Material	Bandgap	E_g (eV)	λ_g (μm)	Lattice Constant (Å)	Refractive Index at E_g
Si	Indirect	1.12	1.11	5.43	3.4
Ge	Indirect	0.68	1.82	5.65	4
GaP	Indirect	2.24	0.55	5.45	3.45
AlAs	Indirect	2.09	0.59	5.66	3.18
GaAs	Direct	1.42	0.87	5.65	3.66
InP	Direct	1.33	0.93	5.87	3.45
InAs	Direct	0.34	3.65	6.04	3.52
AlGaAs	Indirect	1.96–2.17	0.57–0.63	5.65–5.66	
	Direct	1.42–1.96	0.63–0.87	5.65–6.06	
InGaAs	Direct	0.36–1.42	0.87–3.44	5.65–6.06	
InGaAsP	Indirect	1.96–2.26	0.55–0.63	5.45–5.56	
	Direct	0.74–2.26	0.55–1.67	5.56–6.06	

In a semiconductor, a transition may be further classified as intraband or interband. An intraband transition occurs within a given band (conduction band or valence band). An interband transition is a transition from the conduction band to the valence band; in other words, electron–hole recombination.

In a semiconductor, a continuum of electron states exist above the conduction band edge, and a continuum of hole states exist below the valence band edge. If a photon of energy $E_p > E_g$ is absorbed, an electron is created in the conduction band at or above the band edge E_C, leaving behind a hole at or below the valence band edge E_V. Because of the continuum of states above E_C, if the electron has energy greater than E_C, it very quickly "trickles down" by a series of nonradiative intraband transitions to the conduction band edge. Similarly, if the hole is below E_V, it quickly floats up (again nonradiatively) to the valence band edge. These fast nonradiative intraband transitions to the band edge happen in both direct and indirect semiconductors.

The stage is now set for an interband transition. The interband transition may be radiative or nonradiative. If the interband transition is radiative, a photon of energy equal to the bandgap will be emitted. (Actually, the energy is not always exactly E_g, for reasons that will not be gone into here. Typically, the emitted light is within about 50 nm of λ_g.) Whether the transition will be radiative or nonradiative depends greatly on whether the semiconductor is direct or indirect.

Radiative electron–hole recombination is governed by quantum mechanical "selection rules," which basically allow transitions only when the filled and empty states have compatible wave functions. The most important selection rule is conservation of momentum, $\hbar\mathbf{k}/(2\pi)$. Because a photon has momentum E_p/c that is negligible compared with typical electron and hole momenta, this rule is usually stated as follows: The electron and hole must have opposite wavevectors, $\mathbf{k}$ and $-\mathbf{k}$, for fast radiative recombination to occur.

Refer back to Figure 3.10. Because the hole has floated up to the top of the valence band (through intraband transitions), the hole wavevector is $\mathbf{k}_h = 0$. If the semiconductor is direct (like GaAs), the electron (which has trickled down to the conduction band edge) has wavevector $\mathbf{k}_e = 0$. In this case, $\mathbf{k}_e = \mathbf{k}_h = 0$, so radiative recombination occurs. (Nonradiative recombination is also possible, but radiative recombination is usually favored when the wavevector selection rule is satisfied.)

However, if the semiconductor is indirect (like silicon), the electron has a large wavevector. The wavevector selection rule is not met. It turns out that radiative recombination is still possible through higher-order processes. But nonradiative recombination is greatly favored over the higher-order radiative processes (except at low temperatures).

Because of the wavevector selection rule, radiative recombination is more "difficult" for an indirect semiconductor compared with a direct

semiconductor, and the competing process—nonradiative recombination—tends to dominate in indirect semiconductors. This is why indirect semiconductors are such inefficient light sources. Diode lasers are made out of direct semiconductors, as are the most efficient LEDs.

As just noted, even in direct semiconductors, there is competition between radiative and nonradiative processes. For a light emitting device, nonradiative processes should be minimized. Disorder (e.g., caused by impurities or crystal defects) tends to enhance nonradiative rates. For this reason, the active (light-emitting) region of lasers (and LEDs) should be as pure as possible (no dopants) and have excellent crystalline quality.

Returning to radiative transitions, it turns out that there are two types of light emission: spontaneous and stimulated. Spontaneous emission is a random process (although the probabilities are well defined by quantum mechanics); an excited atom or material decays at some point in time. The energy of the emitted photon is fixed by the energies of the two states, but the phase and direction of the emitted photon is random. Spontaneous emission is common in our daily lives: lightbulbs, CRTs, and LEDs all emit photons spontaneously.

Stimulated emission can occur when an atom or crystal in an excited state is hit by an incoming photon of just the right energy. The incoming photon can cause the excited electron to fall to the lower state, creating another photon. This second photon has the same phase and direction as the incoming photon (as well as the same energy). Stimulated emission can happen only when the incoming photon has exactly the same energy as the difference between a filled excited state and empty lower state. Stimulated emission is the basis for the unique properties of laser light.

3.4 Impurities

A pure semiconductor has equal numbers of electrons and holes at all times, as charge neutrality of the crystal must be maintained. Electron–hole pairs exist at nonzero temperatures (temperatures above absolute zero, 0 K, which is obtainable only in theory) due to thermal excitation. Additional electron–hole pairs may also be created under nonequilibrium conditions such as light absorption. But there are always equal numbers of electrons and holes. A pure (or nearly pure) semiconductor is called *intrinsic* (sometimes abbreviated "*i*").

3.4.1 Dopants

Minute concentrations of desired impurity atoms, called dopants, are intentionally introduced into silicon to alter its electronic properties. The impurities replace (substitute for) silicon atoms in the lattice. These impurities have a profound effect on the behavior of a semiconductor, even though

their concentration is small. Typical doping concentrations for silicon range from about 1×10^{15} cm^{-3} to about 1×10^{19} cm^{-3}. Even a heavily doped sample of 1×10^{19} cm^{-3} impurity atoms represents only 1 in 5000 atoms, as there are roughly 5×10^{22} silicon atoms per cubic centimeter.

Dopants are classified as either donors or acceptors. Donor atoms typically have one more valence electron than the lattice atom being displaced. In silicon, donors typically have five valence electrons, which is one more than silicon with four valence electrons. They are called donors because when a donor replaces a silicon atom in the lattice, it gives up one of its electrons to the crystal as a whole. In other words, this "donated" electron fills a formerly empty state in the conduction band. No corresponding hole is created; rather, the donor ion has a positive charge. This positive charge is fixed on the donor ion and cannot move about the crystal as a hole can.

The other types of impurities are called acceptors. Acceptor atoms typically have one fewer valence electron than the displaced atom; acceptors in silicon typically have three valence electrons. When an acceptor replaces a silicon in the lattice, it "accepts" an electron from the valence band. A hole (free to move about the crystal) is created in the valence band. The acceptor ion has a fixed negative charge.

Silicon doped with donors has many more electrons than holes (see Fig. 3.9, e and f). At room temperature, the concentration of electrons is approximately equal to the concentration of donors. There are also thermally created electrons and holes; the number of thermally created electrons is very small compared to the number of "donated" electrons, so they are usually neglected. Essentially, all the holes are thermally generated (unless electron–hole pairs are created by light absorption or other means). Because there are so many more electrons than holes, in this type of doping, the electrons are called majority charge carriers and the holes are called minority charge carriers. Donor-doped silicon is called *n*-type silicon, because the majority carriers are electrons (*n* for "negative").

Conversely, a semiconductor doped with acceptors has many more holes than electrons (see Fig. 3.9, g and h). The majority carriers are holes and the minority carriers are electrons. An acceptor-doped semiconductor is called *p*-type (*p* for "positive"). A doped semiconductor (whether *n*-type or *p*-type) has a much higher electrical conductivity than its intrinsic version. In the party analogy, this is because either the people on the second floor (*n*-type; Fig. 3.9, e and f) or the "holes" on the first floor (*p*-type; Fig. 3.9, g and h) have more room to move in response to potato chips (electric field).

3.4.2 Fermi Level

The carrier concentrations in a semiconductor, n and p, are extremely important parameters for understanding electrical and optical behavior: n is the concentration of electrons (in the conduction band) in cm^{-3}, p is the

concentration of holes (in the valence band) in cm^{-3}. n and p are functions of position in a non-uniform semiconductor (such as a device) or under nonequilibrium conditions (such as an applied electric potential or field, applied magnetic field, light illumination, temperature gradient, etc.). In general, n and p are functions of temperature, and very strong functions of dopant concentrations n_D and N_A (also specified in cm^{-3}).

The intrinsic concentration, n_i, is a constant for a given semiconductor material at a given temperature. For silicon, $n_i = 1.45 \times 10^{10}\ cm^{-3}$. In an intrinsic semiconductor, as discussed previously, the electron and hole concentrations are equal, so

$$n = p = n_i \tag{3.5}$$

The minority carrier concentrations can be calculated from the preceding equations.

At any temperature near room temperature (for silicon, from about 150 K to about 450 K), the carrier concentrations are dominated by the impurities N_D and N_A. To an excellent approximation, the majority carrier concentrations are given by

$$n = N_D - N_A \approx N_D, \qquad \text{for an } n\text{-type semiconductor, and} \tag{3.6a}$$

$$p = N_A - N_D \approx N_A, \qquad \text{for a } p\text{-type semiconductor} \tag{3.6b}$$

within the temperature range of interest. (At very low temperatures, $n < N_D - N_A$ and $p < N_A - N_D$; at very high temperatures, $n > N_D - N_A$ and $p > N_A - N_D$.)

It turns out that for all semiconductors at any doping level, the following equation is valid:

$$np = n_i^2 \tag{3.7}$$

There is an alternate way to express carrier concentrations, rather than discussing n and p directly, which can be helpful. This alternate point of view is the Fermi level, E_F, which is expressed in energy units (electron volts, eV). E_F is typically drawn on band diagrams as a dashed line. The two viewpoints are comlementary because E_F can be derived from n and p, or n and p can be derived from the E_F. A plot of E_F as a function of impurity concentration is given in Figure 3.11.

For an intrinsic semiconductor, the Fermi level is (approximately) midgap, $E_F = E_i \approx E_C - E_V$ and $n = p = n_i$.

For doped semiconductors, the Fermi level is given by

$$E_F = E_i + kT\ln(n/n_i) = E_i - kT\ln(p/n_i) \tag{3.8a}$$

Therefore, at temperatures of interest, and when only one impurity type dominates (i.e., $N_D \gg N_A$ for n-type and $N_A \gg N_D$ for p-type),

$$E_F = E_i + kT\ln(N_D/n_i) \approx E_C - E_g/2 + kT\ln(N_D/n_i), \qquad \text{for } n\text{-type} \tag{3.8b}$$

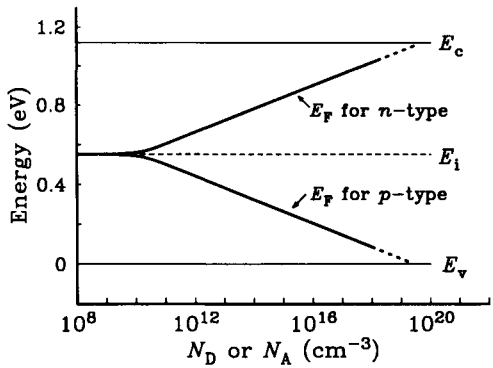


FIGURE 3.11 The position of the Fermi level E_F in Si at 300 K as a function of doping concentration.

and

$$E_F = E_i - kT\ln(N_A/n_i) \approx E_V + E_g/2 - kT\ln(N_A/n_i), \qquad \text{for } p\text{-type} \quad (3.8c)$$

For an intrinsic semiconductor, the Fermi level is E_i, which is near midgap. For an n-type semiconductor, the Fermi level is close to the conduction-band edge, E_C; for a p-type semiconductor, the Fermi level is close to the valence-band edge, E_V. The more heavily doped the semiconductor, the closer the Fermi level will be to a band edge. Except in unusual circumstances, the Fermi level is somewhere within the bandgap.

Example

(a) A silicon sample contains 5×10^{17} cm^{-3} boron. (Boron is an acceptor in silicon.) The temperature is 27°C. What are the carrier concentrations? What is the Fermi energy?

$$p \cong N_A = 5 \times 10^{17}\ \text{cm}^{-3} \qquad \text{and} \qquad n = n_i^2/p = 421\ \text{cm}^{-3}$$

27°C is 300K, so the thermal energy $kT = 8.62 \times 10^{-5}$ (eV/K) $\times$ 300 K = 0.026 eV. Using Eq. (3.8c):

$$E_F \cong E_V + \frac{E_g}{2} - kT\ln\left(\frac{N_A}{n_i}\right)$$

$$= E_V + \frac{1.12\ \text{eV}}{2} - 0.026\ \text{eV} \times \ln\left(\frac{5 \times 10^{17}}{1.45 \times 10^{10}}\right)$$

$$= E_V + 0.110\ \text{eV}$$

The Fermi level is 110 meV above the valence band edge.

(b) Another silicon sample contains $5 \times 10^9\ \text{cm}^{-3}$ phosphorus. (Phosphorus is a donor in silicon.) The temperature is 27°C. What is the electron concentration n?

This is a trick question. In this case, $N_D < n_i$, so the silicon is intrinsic. The electron concentration is $n = n_i = 1.45 \times 10^{10}\ \text{cm}^{-3}$. The Fermi level is

$$E_F = E_i \cong E_V + E_g/2 = E_V + 0.56\ \text{eV}$$

One extremely important property of the Fermi level is that under equilibrium conditions, the Fermi level is uniform throughout the crystal, even if the dopant concentration N_D and N_A are different in different parts of the crystal. Inasmuch as the position of E_F relative to the band edges E_C and E_V is fixed by the dopant concentrations (Eq. (3.8) and Fig. 3.11), what this means is that the band edges themselves move up or down in response to changes in dopant concentrations. We will see this in the following discussion of the diode.

3.5 The p–n Junction Diode

3.5.1 The p–n Junction under Equilibrium Conditions

The junction diode is an important semiconductor device composed of p-type material adjacent to n-type material. The imaginary surface inside the semiconductor that divides p-type material on one side from n-type material on the other is called the junction. For our purposes, equilibrium means that no electric or magnetic fields or potentials are applied, no light is incident on the device, and the temperature is uniform.

A junction diode may be formed as follows. (Note: practical devices are never fabricated as described here, but it is a useful "thought experiment" that aids in understanding.) Imagine that a p-type piece of semiconductor is fused to an n-type piece to the right. Immediately after this imaginary fusion, because there are many more holes on the left than on the right, some of the holes to the left of the junction diffuse right. Similarly, some of the electrons on the right diffuse left. These holes and electrons recombine and annihilate. Because there are now fewer holes than negatively charged acceptor ions just to the left of the junction, there is a net positively charged region. Similarly, just to the right of the junction, positively charged donor ions are no longer balanced by electrons. This sets up an electric field from right to left across the junction. The electric field prevents too many electrons and holes from diffusing. In fact, only a thin region is depleted of mobile carriers, typically on the order of 1 μm thick. This region is called the depletion region or the space-charge region. In equilibrium (no applied voltage), the depletion region is essentially devoid of mobile carriers.

Some of the depletion region is in the p-type side, with width x_p; some of the depletion region is in the n-type side, with width x_n. The ratio of the depletion widths is given by

$$x_p/x_n = N_d/N_a \tag{3.9}$$

Note that more of the depletion is in the lightly doped side than in the heavily doped side of the junction. The total depletion width W is given by

$$W = x_p + x_n = \left[\frac{2\varepsilon_r\varepsilon_0 V_j(N_a + N_d)}{qN_aN_d}\right]^{1/2} \tag{3.10}$$

where ε_r is the relative dielectric constant and V_j is an effective potential difference. At equilibrium, $V_j = V_{bi}$, where V_{bi} is the "built-in potential," which will be discussed next. Its value is given by

$$V_{bi} = \frac{kT}{q}\ln\left(\frac{N_DN_A}{n_i^2}\right) \tag{3.11}$$

So "one-sided" junctions (which have one side much more heavily doped than the other) have the largest total depletion width.

One way to make a very wide depletion region is with so-called "p-i-n diode." A PIN dioide is a sandwich of a p-type region, a nominally undoped (or very lightly doped) intrinsic i-region, and an n-region. This structure usually results in a depletion width extending across all of the i region, plus small depletion widths in both the p-type and n-type regions.

The equilibrium state of a p–n junction is shown in Figure 3.12. Figure 3.12a shows the charge distributions of fixed ionic impurity atoms in the space-charge region (depletion region). (The approximation is made that there is a uniform charge distribution in the depletion region on each side of the junction.) The resulting electric field as a function of position is graphed in Figure 3.12b from

$$\varepsilon = \rho x, \text{ for a uniform charge distribution.} \tag{3.12}$$

Integrating results in the electric potential, as shown in Figure 3.12c,

$$\phi(x) = -\int \varepsilon(x)\,dx \tag{3.13}$$

The change in electric potential from the extreme left to the extreme right is the built-in potential, V_{bi}. Typical built-in potentials are in the range 0.5 to 1.0 V.

Band diagrams are drawn from an electron's point of view, with

$$E = -q\phi \qquad (E \text{ here is energy}) \tag{3.14}$$

where q is the magnitude of the electronic charge. The band diagram for the semiconductor is distorted by the curve for ϕ (upside down because of the negative charge of the electron), resulting in the band diagram for a p–n

junction shown in Figure 3.12d. To minimize their potential energy, electrons would like to "fall down" in an energy diagram; holes would like to "float up." From Figure 3.12d, The energy "potential hill" that a hole from the p-side (majority carrier) would have to overcome in order to cross to the n-side is qV_{bi}; and the energy potential hill for an electron to cross from the n-side (majority carrier) to the p-side is also qV_{bi}. On the other hand, an electron on the p-side (minority carrier) would *gain* energy qV_{bi} from sliding down the hill, and a hole on the n-side (minority carrier) would also gain qV_{bi} from floating uphill.

The same resulting band diagram, Figure 3.12d, can be drawn using a different method. Because the diode is in equilibrium, we know that the Fermi level must be the same everywhere; so we can draw a horizontal (dashed) line across the page representing the Fermi level. At the extreme left side of the diagram, we know that the semiconductor is p-type, so E_V must be a horizontal line just below E_F on the left side of the diagram. The conduction band edge E_C can be drawn in as a horizontal line a distance E_g above E_V. This completes our drawing of the left portion of the figure.

Turning now to the right side of the figure; because the material is n-type, we can draw in E_C as a horizontal line just above E_F. Next, E_V can be drawn

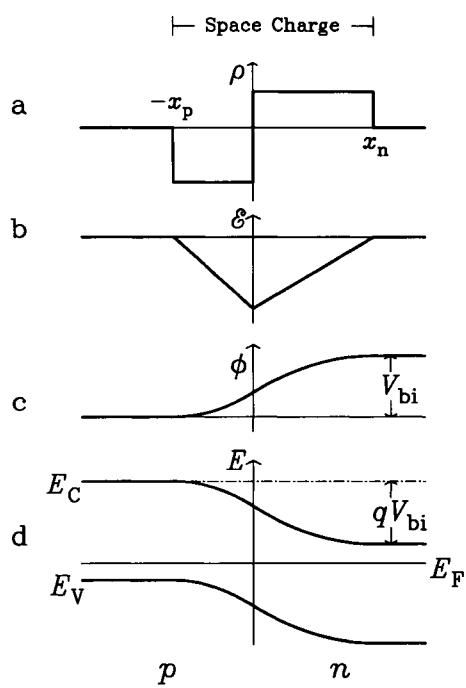


FIGURE 3.12 An abrupt p–n junction with no applied bias. (a) Net charge distribution. (b) Electric field distribution. (c) Electrostatic potential. (d) Energy-band diagram.

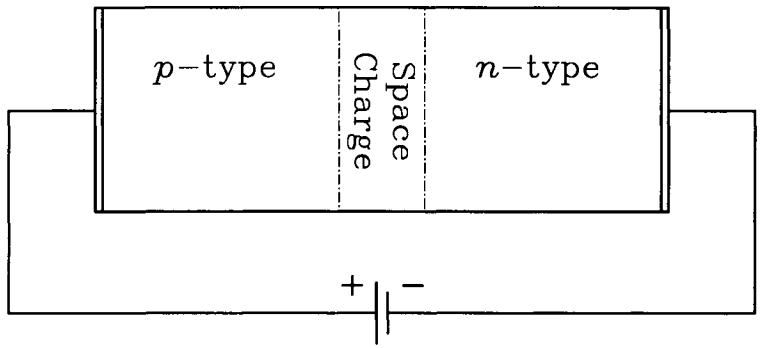


FIGURE 3.13 A diode with an applied forward bias. The space-charge region is typically on the order of a micron wide.

in a distance E_g below E_C. Finally, we can complete the diagram by smoothly joining the left and right sides of the diagram, always keeping E_C a distance E_g above E_V. This results in Figure 3.12d, as promised.

3.5.2 The p–n Junction: Forward Bias

So far, we have considered the p–n junction at equilibrium. Now, we consider what happens when each side of the junction is contacted, and an external forward voltage is applied (Fig. 3.13). In general, a small voltage drop will appear at each contact (where the wire touches the silicon), but we will not be concerned with the contact regions or voltages. In addition, we will neglect any voltage drop across the bulk silicon on each side of the junction; this is a very good approximation because doped silicon is a good conductor of electricity. Therefore, we will consider that all of the applied voltage drop occurs near the p–n junction itself; away from the junction, the *p*-type silicon has one uniform potential—and the *n*-type silicon has another.

A diode is said to be forward-biased if the applied voltage (bias) of the *p*-type side of the junction is positive relative to the *n*-type side; it is reverse-biased if a negative voltage is applied to the *p*-type side. Figure 3.12 shows the band diagram of a diode at equilibrium. Figure 3.14 shows what happens to the diode as a voltage is applied.

Strictly speaking, the Fermi level is defined only in equilibrium. Outside the space-charge region, the Fermi level can be defined on each side, as we are neglecting any voltage drop across the bulk regions. The applied voltage V causes the Fermi levels on each side to separate by energy qV. For forward bias ($V > 0$), the right side (*n*-type side) of the diagram slides up by energy qV relative to the left side. For reverse bias ($V < 0$), the right side moves down. This is shown in Figure 3.14d. For all three curves (forward bias, no applied bias, reverse bias), the Fermi level on the *p*-type side is on the *x*-axis. But on

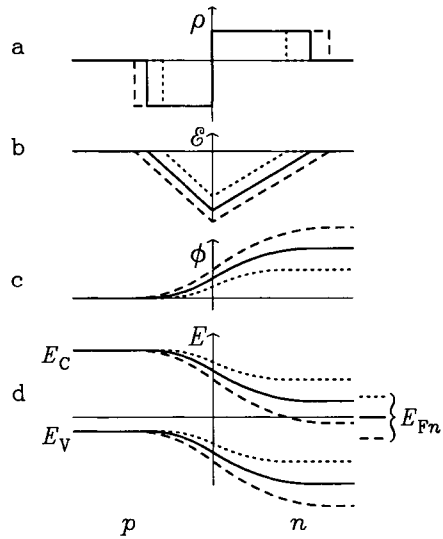


FIGURE 3.14 An abrupt p–n junction with no applied bias (solid curves), with a small applied forward bias (dotted curves), and with a small reverse bias (dashed curves). The solid curves are identical to those in Figure 3.12.

the n-type side of the space-charge region, the Fermi level moves with the applied bias, as shown to the extreme right of Figure 3.14d.

Note that under forward bias, the bands get "flatter," but under reverse bias, the bands get "steeper." The flatter bands make current flow easier in response to the applied field; the steep bands make current flow very difficult.

Under forward bias, the applied field partially counteracts the built-in field. The energy "potential hill," qV_j, changes from qV_{bi} to $q(V_{bi} - V)$. The junction potential V_j is $V_j = V_{bi} - V$, where V is the externally applied bias ($V > 0$ for forward bias). Recall that at equilibrium, the built-in field was just enough to balance the tendency of the carriers to diffuse to regions of lower concentration. Because the field at the junction becomes smaller under forward bias, some holes now flow from the p-type side across the junction to the n-type side. (A few holes may recombine with electrons in the depletion region and not make it to the n-type side.) Holes that make it from the p-type side to the n-type side are called injected minority carriers because the holes are minority carriers in n-type material and they were injected from the p-type side by forward biasing. The injected minority carriers continue to diffuse right in the n-type side until they are recombined with majority carriers (electrons).

Similarly, under forward bias, electrons from the n-type side now also face a lower electric potential barrier to diffusion, so some electrons diffuse from right to left across the depletion region (where a few may recombine with holes) to the p-type side, and continue leftward (now as injected minority carriers) until they recombine with majority holes.

Because holes (positive charge) are moving left to right and electrons (negative charge) are molving right to left, there is a net positive (left-to-right) current flow through the diode. As the forward voltage is increased, the bands get ever "flatter" as the electric potential hill of height $qV_j = q(V_{bi} - V)$ gets ever smaller, and there is more and more current flow.

Also, because the field at the junction decreases under forward bias, the majority carriers on each side can get closer to the junction. This causes the depletion layer to narrow under forward bias. This is illustrated in Figure 3.14a and can be shown by Eq. (3.10). Because the depletion layer W is narrower under forward bias, the capacitance of the junction C_j increases. C_j is given by

$$C_j = \frac{\varepsilon_0 \varepsilon_r A}{W} \tag{3.15}$$

where W is the total depletion width, A is the area of the junction, and ε_r is the relative dielectric constant of the semiconductor.

3.5.3 The p–n Junction: Reverse Bias

Under reverse bias, the applied field reinforces the built-in field and widens the depletion region. The potential barrier for majority carriers, $qV_j = q(V_{bi} - V) = q(V_{bi} + |V|)$ increases so there is no majority carrier flow. Minority carriers can flow across the junction, however, because each minority carrier that flows across gains energy $q(V_{bi} + V)$. In other words, electrons will flow from the p-type side into the junction, and holes will flow from the n-type side into the junction, creating a right-to-left (negative) current in response to the reverse bias. However, this current is ideally extremely small, because the concentration of minority carriers on each side is so small. (Typically the ratio of majority to minority carriers is at least 1×10^{10} cm^{-3}.) In real diodes, the true reverse current is larger than predicted from this simple model, but it is still very small for reverse biases of several volts or less.

Under reverse biases of more than a few volts, one of two additional mechanisms can become important, which can dramatically increase the reverse current flow from that predicted from the simple model just presented. These two mechanisms are Zener breakdown and avalanche breakdown Zener diodes have important uses in electronics, but they are not used for optoelectronic devices. Avalanche breakdown, however, is important for one type of photodiode detector, so it will be explored in more detail.

Zener breakdown is a quantum tunneling process. It is favored by a very narrow depletion region, which can be achieved by doping both sides of the junction heavily. Zener diodes are fabricated for use in reverse bias as voltage references. Once a Zener diode goes into breakdown, the voltage across the junction remains nearly constant over a wide range of reverse currents. A

more complete description of Zener breakdown is beyond the scope of this book. Zener diodes are not used as optoelectronic devices and will not be discussed further here.

Avalanche breakdown results from electron–hole pair generation in the depletion region under reverse bias. As discussed in section 3.3, energy greater than or equal to the bandgap energy E_g is required to create an electron–hole. As discussed earlier, this energy may come from a photon. However, it is also possible for this energy to come from the kinetic energy of a very energetic electron or hole. For instance, a strongly reverse-biased p–n junction will have a high electric field in the depletion region. The kinetic energy that an electron acquires while in the strong electric field of the depletion zone may be more than E_g. In such a case, there is a good probability that this electron will give up some of its excess kinetic energy to create an electron–hole pair. This electron and hole then separate because of the electric field, adding to the reverse current across the diode.

If the field is strong enough, not only will one electron create an electron–hole pair, but then the created electron and hole will in turn gain enough energy from the field to excite two more electron–hole pairs, and these may in turn create four more pairs. This process is called avalanche breakdown or multiplication, as one carrier crossing the depletion zone can create an "avalanche" of additional carriers. The multiplication factor (how many carriers out per carrier into the junction) depends strongly on the field, and thus on the applied reverse bias. Above some critical field, the average multiplication factor becomes greater than 1, and the junction is said to be undergoing avalanche breakdown. Even below the critical field, some statistical fraction of carriers will have energy (gained from the field plus their thermal energy) greater than E_g, causing a few electron–hole pairs to be generated in the junction. This generated current explains why the current under reverse bias is larger than would be expected if this generation mechanism is neglected. Note that the junction capacitance C_j decreases with reverse bias (see Eqs. (3.10) and (3.15)).

Next, we turn to the question of what happens when light is absorbed near a p–n junction and how a p–n junction may be used as a light detector.

3.6 Photodetectors

Photodetectors are devices that are designed to convert light energy (of the appropriate wavelengths) into an electrical signal. The most common type of photodetectors used in optoelectronic communications is the photodiode. Photodiodes are diodes (p–n junctions) that convert light energy into reverse current.

The relative spectral response of any semiconductor-based photodetector depends almost exclusively on the material. In other words, various types of

photodetectors may differ in many important ways, but the shape of their sensitivity versus wavelength curve will be very similar. Because λ_g for silicon is 1.1 μm, silicon detectors are usable only for shorter wavelengths. For this reason, silicon detectors cannot be used for fiber-optic systems operating at the popular telecommunication wavelengths of 1.3 μm or 1.55 μm. For these applications, InGaAs or InGaAsP alloys are often used. Silicon is an efficient absorber, however, in the 0.8-μm wavelength region (the wavelength region emitted by GaAs LEDs and lasers). Some type of silicon detector will usually be the best choice for a data communication application using 0.8-μm light; silicon detectors can have very good responsivities and low dark current in this wavelength region. Best of all, silicon devices are typically far less expensive than equivalent devices made from any other semiconductor because silicon's dominance of the enormous electronics industry. GaAs may also be used in detectors for this wavelength region.

The efficiency of a given photodiode is specified by its quantum efficiency, η, or (equivalently) by its responsivity, r. Both of these quantities are functions of wavelength, applied bias, and temperature. Quantum efficiency η is a unitless positive number that equals the number of electrons generated at the output of the detector divided by the number of photons incident on the detector. Responsivity r is the output current divided by the input light energy, so it is measured in A/W (or, equivalently, mA/mW).

Responsivity and quantum efficiency are related as follows:

$$r = \frac{\eta q \lambda_0}{hc} \tag{3.16}$$

where λ_0 is the wavelength in vacuum, h is Plank's constant, c is the speed of light, and q is the charge on the electron.

Dark current is the current across a photodetector when there is no incident light. It is ideally zero. However, when a bias is applied to a detector, a small current will result. Any crystal defects in the active region of the device will tend to increase this undesired leakage current. Even if there were no defects, there would be some current as a result of thermal noise (a source of random noise). Dark current tends to be greater for semiconductors with small bandgaps, and it tends to increase with temperature.

Another important figure of merit for any photodetector is the noise equivalent power (NEP). It is a measure of the signal-to-noise ratio and the sensitivity. NEP is defined as the optical power level into the photodetector that results in an output signal that is equal to the root-mean-square noise of the signal. NEP for a given device depends on the wavelength at which it is measured and on the bandwidth used for the noise measurement. The lower the NEP, the better the detector.

A photon incident on a semiconductor crystal may be reflected at the air–semiconductor interface, or it may enter the semiconductor. The probability of reflection is given by the Fresnel coefficient, which is equal to $\sqrt{R}$,

where R is given by Eq. (3.1) in the case of normal incidence. (Note that n_2 is the index of refraction of the semiconductor *at the wavelength of the photon.*) The intensity of the reflected light is R times the intensity of the incident light.

For a photodetector, reflection at the air–semiconductor interface is undesired, because it reduces the efficiency of the detector (and can have other undesired effects). For this reason, commercial photodiodes are typically antireflection coated.

If the photon is not reflected at the air–semiconductor interface, it may be absorbed by the semiconductor crystal. From what we have already learned about semiconductors, we know that long-wavelength light ($\lambda_0 > \lambda_g$) will not, for practical purposes, be absorbed, but will be transmitted through the crystal. But if the wavelength of the light is shorter than $\lambda_{g(\text{direct})}$, it will tend to be absorbed by the crystal, with the light energy being converted to electron–hole pairs.

The intensity of the light at a distance z from the air–semiconductor interface is given by

$$I = I_0 \exp\left(\frac{-z}{\Lambda}\right) \tag{3.17}$$

where Λ is the absorption length of the semiconductor, and I_0 is the light intensity just inside the semiconductor. Λ is a property of the semiconductor and is a function of wavelength. Λ is very long for $\lambda_0 > \lambda_g$, as there is (almost) no absorption. But for $\lambda_0 < \lambda_g$, a typical value of Λ is on the order of 1 μm. The fraction of light absorbed in a material of thickness W is given by

$$\left[1 - \exp\left(\frac{-W}{\Lambda}\right)\right] \tag{3.18}$$

For maximum efficiency, the absorbing thickness W of a photodetector should be equal to two or three times Λ, so that (almost) all of the light is absorbed. A few microns is a typical absorbing-region thickness.

When a photon is absorbed in a semiconductor, its energy is converted into an electron–hole pair. In the absence of an electric field, the electron–hole pair will simply recombine. If a strong enough electric field is applied to the crystal, the electron and hole can separate (instead of recombining) and be collected at opposite electrodes, resulting in a small current. One of the easiest ways to create a strong electric field on a micron scale is with a p–n junction. This explains why photodiodes can be the cheapest type of photodetector, and why they are so popular for optical data communication applications.

Inexpensive commercial photodiodes come in two types: the PIN photodiode and the avalanche photodiode (APD). Other types of photodetectors are also available, including phototransistors, photoconductors, and MSM detectors. The PIN photodiode is the most popular detector for data

communication applications; we will discuss it first. The APD will then be discussed briefly. Finally, the MSM photodetector will be discussed, as it has some important advantages over a PIN diode.

3.6.1 The PIN Photodiode

For the energy of a photon to be collected as a current, the absorption must take place in an electric field. As we have noted earlier, the depletion region of a p–n junction is where there is a strong electric field. Therefore, we want the absorption to take place in the depletion region. But the depletion region (the absorbing region) needs to be at least a few absorption lengths thick for good efficiency. This is thicker than the depletion region in a typical general purpose p–n diode.

An even more important reason to have a relatively thick depletion region is to decrease the junction capacitance, C_j (see Eq. (3.15)). To make a relatively thick (several microns) depletion region, a p-i-n structure is used. The depletion region will start in the p-type region, extend over the entire intrinsic (i-type) region, and end in the n-type region. Figure 3.15a illustrates the structure of a PIN diode. Unlike other detectors, a PIN diode will work with no applied bias (thanks to the built-in field of the junction), but a few volts reverse bias is usually applied to improve performance.

The bandwidth of a PIN diode is limited by two factors, either of which may dominate. One limiting factor is that the carriers take time to drift across the depletion layer from where they are created to the electrodes. If the depletion layer width is W and the velocity of the carriers is v, the time can be estimated as $t = W/v$, where the velocity v is equal to $v = \mu E$ (μ is the mobility and E is the electric field) for weak fields, or $v = v_{sat}$ for strong fields.

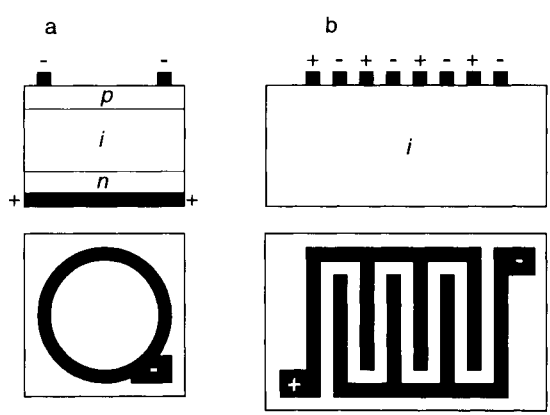


FIGURE 3.15 Comparison of the construction of (a) a typical PIN photodiode and (b) a typical MSM photodetector. Both cross-sectional and top-down views are shown. Metal contacts are shown in black.

For most cases, it is reasonable to assume saturation velocity is attained in a PIN diode (or APD).

Another limiting factor in detector bandwidth is the RC time constant, $\tau = RC_{tot}$. The relevant resistance R is the resistance of the amplifier bias circuit to which the PIN diode is attached. The relevant capacitance C_{tot} is the capacitance of the PIN diode itself, C_j, in parallel with the capacitance of the input stage of the amplifier C_{amp} and any package capacitance C_{pkg}. The PIN diode capacitance C_j is given by Eq. (3.15).

The PIN diode bandwidth is therefore approximately

$$B = \frac{1}{2\pi\,[(W/v_{sat}) + R(C_j + C_{amp} + C_{pkg})]} \tag{3.19}$$

The PIN photodiode, compared to the APD, has a low dark current and low noise, and is able to operate at low applied voltages. Its maximum quantum efficiency is 1.

3.6.2 The Avalanche Photodiode

The avalanche photodiode works by taking advantage of the avalanching effects described in section 3.5.3. As in the PIN diode, light absorbed in the depletion region causes the formation of electrons and holes, which separate as a result of the electric field. But in an APD, the field is made strong enough that at least one of the carriers (usually electrons) gains enough energy from traversing the field that it can cause additional electron-hole pairs to form. These electrons may in turn cause another generation of pairs. In this way, the APD has gain, which the PIN diode does not. The maximum quantum efficiency of the PIN diode is 1; but APDs can obtain quantum efficiencies approaching 100.

Avalanche photodiodes must be operated at a relatively high reverse bias (typically 20 V). For data communication applications, such relatively high voltage sources are not easily available. Primarily for this reason, APDs are not often used in data communication applications.

3.6.3 MSM Photodetectors

Another type of detector that is gaining popularity is the metal-semiconductor-metal (MSM) detector. The MSM is semiconductor-based, but is not based on the p–n junction. Figure 3.15b illustrates the structure of an MSM detector. To create an MSM, metal fingers are deposited on the surface of an intrinsic semiconductor crystal. The contact between the metal and the surface of the semiconductor is fabricated to be what is called a Schottky contact. The physics behind Schottky contacts will not be discussed here; it is sufficient to say that a Schottky contact acts like a diode. Alternating fingers are biased alternately, creating an electric field between the fingers. The

fingers are made quite close together to allow a reasonably high field to be created; high enough to form a space-charge region between the fingers and to some depth below the surface of the detector. Photons striking the semiconductor between the fingers create electron–hole pairs that are collected at the finger contacts, thereby creating a photocurrent.

A disadvantage of the MSM is that the area taken up by the metal fingers blocks some of the incoming light. However, MSMs have some interesting advantages over PIN diodes. It turns out that an MSM has a lower capacitance than a PIN diode of equal area, because of geometric factors. This means that an MSM of equal capacitance (and therefore equal speed) can have a larger area than an equivalent PIN diode. This is important at the next level of assembly, because *packaging* (with the requirement of precise alignment of the optoelectronic device to the fiber) is one of the most expensive steps in the manufacture of a finished device. A detector with a larger active area requires less accuracy in alignment of the package. Another advantage of the MSM is that unlike the PIN photodiode and the APD, no back contact is required for an MSM, as both the positive and negative contacts are on the surface of the device; this greatly simplifies device fabrication.

3.6.4 Choosing a Photodetector for the Application

Choosing a photodiode for a given application, like any other engineering choice, involves trade-offs. Among the considerations:

1. *Responsivity.* One would like high responsivity at the desired wavelength. This is primarily governed by the material. Other factors can be surface reflectivity (antireflection coating improves device responsivity) and light blockage resulting from top–surface contacts.
2. *Dark current.* Dark current is also strongly dependent on material. Dark current results in a nonzero "off" level that must be dealt with at the circuit level. Poor crystallographic quality of a junction will result in higher dark current.
3. *Noise.* Noise on the signal degrades link performance. The frequency range of the photodetector noise relative to the system data rate is important; low-frequency noise and very-high-frequency noise can be filtered out of the signal, but noise near the system data rate is difficult to handle. Noise (from all sources, not just detectors) near the operating frequency of the system can be the limiting factor in link performance. Noise is usually specified in terms of NEP.
4. *Capacitance.* For a given device, minimum capacitance is desired if speed is important (which it is in data communication applications).
5. *Package.* As mentioned in the discussion of the MSM detector, packaging is an important consideration for optoelectronic devices. (This subject is addressed in Ch. 4.)

6. *Cost.* Because data communication tends to be a very cost-sensitive area, the dollar cost of each design decision must be considered at every level of assembly. The total system cost is what is most important; sometimes a more expensive component will result in a less expensive overall system.

The PIN and MSM detectors are the most likely to be used in data communication applications today because of their small size, low required bias voltages, and relatively modest cost.

3.7 Light Emitting Diodes

Light emitting diodes or LEDs are incoherent optical sources that emit light based on spontaneous emission. LEDs have found their way into almost any common electronic device that we use: stereo systems, TV sets, VCRs, and telephone answering machines are a few examples. LEDs are typically infrared, red, orange, or green in color. Although blue LEDs have been commercially available for more than a decade, they are more expensive than other colors and are less frequently used. With their increasing intensity and their long life, LEDs are finding their way into signs and even traffic lights.

We start this section with a look at appropriate materials for LEDs. Then we look at the operation of LEDs and some key characteristics of these devices. The device structure of an LED can dramatically change its emission beam pattern and shape.

3.7.1 LED Materials

In section 3.3, we looked at optical processes in semiconductor materials. The difference in band structure between direct and indirect bandgap semiconductors were discussed. As was seen, efficient photon emission occurs in direct bandgap material. This is why the material of choice for efficient LEDs tends to possess direct bandgaps.[2]

In Table 3.1, the wavelength corresponding to the bandgap transition, λ_g, was shown for several direct bandgap semiconductors, for example, 0.87 μm for GaAs. Therefore, if emission of about 0.87 μm is required, a GaAs *active region* is selected for the LED. This active region tends to be sandwiched between AlGaAs layers—with higher bandgaps than GaAs—to permit confinement of carriers (electrons and holes) within this region. Carrier confinement gives the electrons and holes greater opportunity to recombine

[2] In some cases, the desired wavelength—for example, green—may not be readily obtained by well-understood direct bandgap semiconductors. In these situations, an indirect bandgap semiconductor, such as GaP, may be heavily doped with an impurity, such as N. The impurity gives rise to levels in the band structure, allowing radiative transitions to occur, although at a lower efficiency than direct bandgap semiconductors.

radiatively in the active region. The AlGaAs layers are doped with n- and p-type impurities so that a p–n junction (see section 3.5) is formed, allowing *pumping* of the structure with electric current.

Figure 3.16a shows the case where the p–n junction is under no external pumping or $V_{\text{applied}} = 0$ V. By applying a voltage $V_{\text{applied}} \approx E_g(\text{GaAs})/q$, the bandgap is changed to that of Figure 3.16b. In this case, electrons and holes are available for direct bandgap recombination to generate photons of frequency v_0, where $hv_0 \approx E_g(\text{GaAs})$ and the corresponding wavelength is $\lambda_0 = c/v_0 \approx \lambda_g$.

3.7.2 LED Structures

Light emitting diodes tend to be made in two varieties: surface-emitting LEDs and edge-emitting LEDs (ELEDs). Figure 3.17 shows a typical surface-emitting LED with an optical fiber attached to it. A GaAs substrate that is n-doped is used to deposit epitaxially the LED structure shown in Figure 3.17. These layers are the n-doped AlGaAs (or n-AlGaAs) lower carrier-confinement layer, the p-GaAs active layer, the p-AlGaAs upper carrier confinement layer, and a highly p-doped GaAs (or p^+-GaAs) layer for the p-electrode. A dielectric is deposited on the p^+-GaAs layer. A small circular window is etched out in this dielectric so that the p^+-GaAs is exposed there. The p-electrode is deposited so that it makes contact only through the

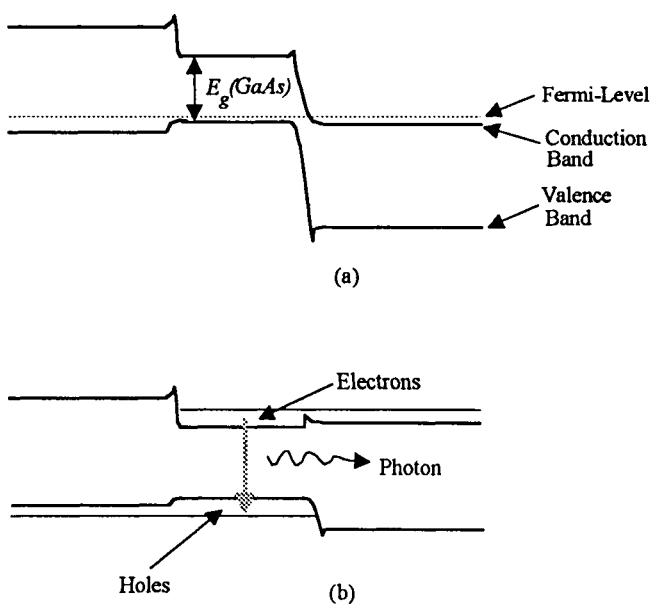


FIGURE 3.16 p–n junction under (a) no bias and (b) bias.

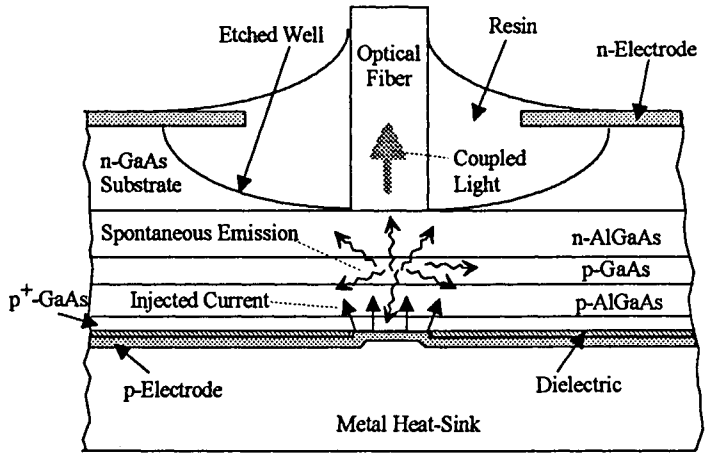


FIGURE 3.17 Surface-emitting LED.

circular opening to the p^+-GaAs. The deposition of the n-electrode is performed with large circular openings to accommodate the optical fiber. The n-GaAs substrate is etched to create a well that extends to the surface of the n-AlGaAs layer. The metal heat sink is attached to the p-electrode. The optical fiber is inserted and is fixed in position with resin.

As Figure 3.17 illustrates, the active layer is very close to the large heat sink, so the p–n junction is prevented from heating up too much (heat generally has an adverse effect on emission properties of semiconductors and leads to lower efficiencies). A small percentage of the light generated in the active layer is coupled into the fiber.

Figure 3.18 shows a possible structure for an ELED. The difference here in the material structure compared with the surface emitter in Figure 3.17 is that the AlGaAs layers not only act as carrier confinement layers, but also as claddings that allow *optical* confinement (see section 3.2.3). Careful selection

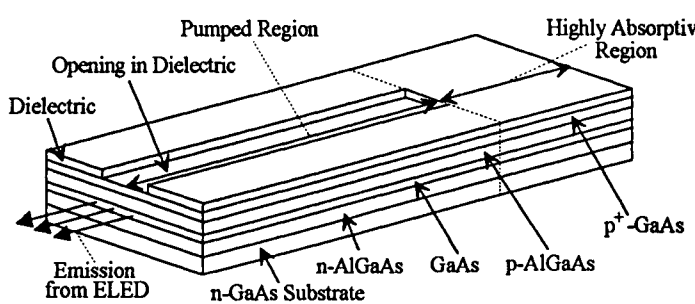


FIGURE 3.18 Edge-emitting ELED.

of the thickness of these AlGaAs claddings is required in conjunction with the choice of their composition (composition x in $Al_xGa_{1-x}As$ will determine its index of refraction). The GaAs active layer thickness will also be important in determining the optical confinement.

In order to maintain spontaneous emission from an ELED and to prevent it from achieving stimulated emission (see section 3.8.2), there has to be sufficient loss in the ELED. In some ELED structures, the facet reflectivity is reduced to maximize transmission loss from the facet. In this case, because the facet reflects less light, more light leaves the facet and is transmitted out, rather than being reflected back into the ELED cavity. Another technique used in ELEDs to prevent stimulated emission from kicking in is to have a thin active region which causes the confinement factor (see section 3.2.3) to be small.

The structure shown in Figure 3.18 uses yet another method for preventing stimulated emission. Here, a section of the ELED cavity is unpumped (no electric current is permitted to flow) so that this region remains highly absorptive. This highly absorptive region absorbs any light that the cavity sends toward it and, therefore, no significant light is reflected back into the pumped section of the ELED to allow stimulated emission to occur. The pumped region of the ELED cavity allows spontaneous emission to be amplified before it leaves the structure. At high levels of pumping, the ELED gives out a high level of optical output, known as superluminescence.

The ELED in Figure 3.18 is fabricated using n-GaAs substrate that has the following layers epitaxially deposited: n-AlGaAs lower cladding layer, undoped GaAs active layer, p-AlGaAs upper cladding layer, and p-GaAs contact layer. Then, a dielectric such as Si_3N_4 is deposited on the structure and a window is opened up in this dielectric to expose the p^+-GaAs surface. The p- and n-electrodes are deposited (not shown in Fig. 3.18 for clarity). Individual ELED devices are formed by using the cleavage planes of the semiconductor crystal to cleave the four sides (more on cleaving is given in section 3.9 on semiconductor lasers).

3.7.3 LED Characteristics

We described the far-field in section 3.2.4. A key difference between surface-emitting LEDs and ELEDs is in their far-field pattern. Figure 3.19 shows the far-field patterns for (a) a typical surface-emitting LED and (b) a typical ELED. Both the vertical far-field (VFF) and the horizontal far-field (HFF) curves are shown. The FWHM of these curves are used as the values for the VFF and HFF. For low numerical aperture (see Chapter 2) optical fibers, ELEDs can provide significant improvement in coupling efficiency over surface-emitting LEDs.

Another key characteristics is the optical power versus current (or P vs. I) characteristics of LEDs. Typical surface-emitting LEDs give higher optical

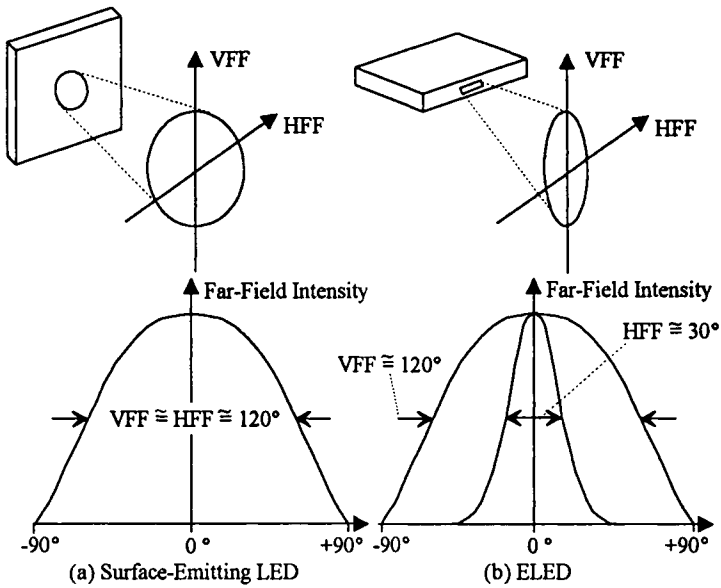


FIGURE 3.19 Far-field characteristics.

output power at low current levels when compared to ELEDs, as ELEDs tend to be effected by heating to a greater extent than surface-emitting LEDs. This is because surface-emitting LEDs have their junctions very close to their heat sinks and therefore operate "cooler" than ELEDs. With careful design of the ELED structure, however, most of the optical field can be "positioned" in the cladding layers. In this case, the ELED can be made to go superluminescent and high optical output power can be obtained at high current levels. This is achieved through the amplification of spontaneous emission along the ELED cavity. Furthermore, the surface-emitting LED still gives higher optical output than the ELED at low current levels, as the active region of the ELED causes self-absorption at these levels. As the pumping current increases, the active region becomes more transparent (less self-absorbing) and larger optical output is obtained.

In general, LEDs are excellent optical emitters for optical fiber systems requiring:

- low power (0.1 to 3 mW)
- large spectral width (50 to 100 nm)
- slow rise time (>1 ns)
- insensitivity to optical reflections
- low temperature dependence
- high reliability ($\sim 10^8$ h)
- low cost

3.8 Lasers

Earlier in this chapter we saw the difference between stimulated emission and spontaneous emission. The word *laser* is an acronym for light amplification by stimulated emission of radiation. Laser action was first demonstrated in a ruby crystal by Maiman [5] back in 1960. Maiman's work allowed coherent light to be generated for the first time.

Lasers have come a long way since those early days. Lasers can be found in different sizes ranging from the smallest laser diode which easily fits on the tip of a finger to ones that easily fill several rooms. Wavelengths spanning the spectrum from x-rays [6] to the deep infrared are now possible as are powers of gigawatts and beyond used in experimental work for fusion reactors. Lasers have found many applications that have revolutionized a great many fields, ranging from medicine to telecommunications.

This brief section introduces the essential components of a laser and the conditions necessary to obtain laser action.

3.8.1 Three Elements of a Laser

As is shown in Figure 3.20, there are three essential laser elements: (a) means for feedback, (b) a gain medium, and (c) a pump. These three elements are the fundamental building blocks of a laser cavity. Like any oscillator, laser oscillators need feedback and gain to keep the oscillation going. Because we have light that we are trying to feed back into the cavity, a mirror or a partially transmitting mirror is ideal for achieving this optical feedback. The gain medium is made up of a material that can provide the necessary population inversion of electrons and allow stimulated emission to take place. In section 3.3.2.2 on emission, we saw that a material that is in an excited state may emit light as the excited electron falls to a lower energy state. Figure 3.21 shows a material in thermal equilibrium using our party-goers model. Once the material is pumped, it is no longer in thermal equilibrium and population inversion is observed, as illustrated schematically in Figure 3.21b. Good gain media can maintain their population inversion (they do not quickly loose their population inversion through spontaneous emission transitions).

Population inversions can be created in a number of different ways [7]: ruby lasers use a xenon flash lamp to pump the ruby crystal; HeNe lasers use gas

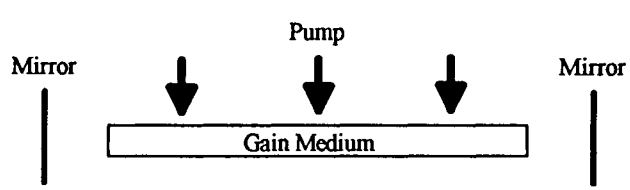


FIGURE 3.20 The three elements of a laser.

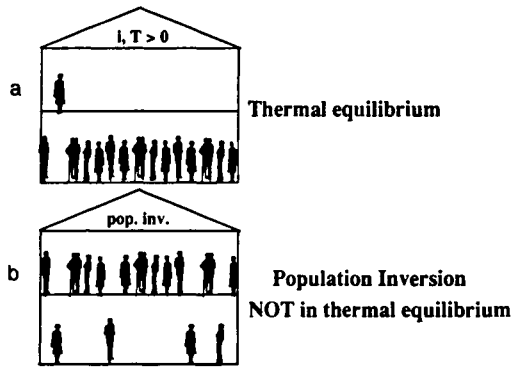


FIGURE 3.21 Illustration of population inversion in a laser.

discharge to create population inversion; and semiconductor lasers use an applied voltage to inject carriers and create the inversion.

As we saw in section 3.3.2.2, photons of particular wavelength and direction can stimulate an electron to make a transition from a high state to a low one, known as stimulated emission. This process causes photons to gain strength as they traverse—or move along—the laser cavity, and hence the name "gain medium." Now, as we steal away electrons from the high state and reduce the inversion, the gain medium loses its strength to the photons. To maintain the laser oscillation, we need to ensure that the inversion is kept in place. This is done by a pump that introduces energy to the gain medium in such a way that the inversion is maintained at an adequate level.

3.8.2 Condition for Laser Oscillation

We have all been at some function where the microphone picks up sound from the speakers and generates a horible noise that wakes everybody up. This is a typical example of oscillation in which there are the prerequisite ingredients of amplification and feedback. However, have you every wondered what gets the oscillation going in the first place. It is that ever-present noise. In lasers, spontaneous emission is ever-present in our gain medium and actually gets the oscillation going when the conditions are just right in the cavity.

Figure 3.22 has a medium of gain G and loss a, and two partially reflecting mirrors of reflectivity R_1 and R_2 separated by a distance L. We can use Figure 3.22 to see when the conditions would be suitable for laser oscillation to take place. Assume a beam of light of intensity I_0 in the cavity incident on mirror R_1. After reflecting off the mirror, the intensity is $R_1 I_0$ with $(1 - R_1)I_0$ being transmitted through the mirror. The gain medium amplifies light exponentially so that just as it reaches mirror R_2, the intensity is $R_1 I_0 \exp[(G - a)L]$. Following the light back through the cavity to the point where we started, we

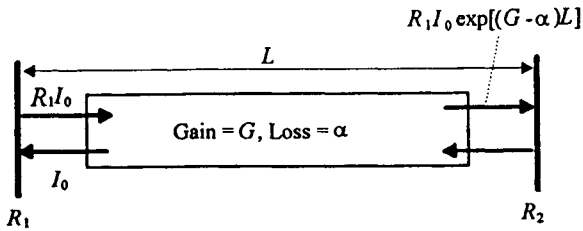


FIGURE 3.22 Intensity as light travels in a laser cavity.

find that the intensity will be $R_2R_1I_0\exp[2(G-a)L]$. For the oscillation to take place, the intensity of the light after one complete trip around the cavity must be at least equal to its starting intensity, I_0. This minimum requirement is called the threshold condition:

$$R_1R_2\exp[2(G_{th}-a)L]=1 \tag{3.20}$$

or

$$G_{th}=a-\frac{1}{2L}\ln(R_1R_2) \tag{3.21}$$

where G_{th} is G at threshold.

3.9 Semiconductor Laser Diodes

Dioide lasers convert electric current into coherent light. In fact, they are the most efficient method for converting electricity into light with conversion efficiencies running as high as 90% in some of the most advanced versions of laser diodes.

Efficient laser action in a semiconductor requires:

- optical confinement
- optical reflection or feedback
- carrier confinement
- gain

Fortunately, all of these are possible in some semiconductor materials and this is why we have such efficient semiconductor lasers available to us.

In 1962, stimulated emission was demonstrated for the first time in semiconductors using a homojunction by Hall [8]. Heterojunction lasers followed homojunctions and by 1970, continuous wave (CW) operation of a semiconductor laser was achieved at room temperature (for a detailed description of the early work on semiconductor lasers and homojunctions see Casey and Panish [9]). Almost all commercially available semi conductor lasers are

now based on heterojunctions. We will limit our discussion here to heterojunction laser material and will review some of the most common semiconductor laser cavities that are fabricated with such structures.

3.9.1 Heterojunction Lasers

The word *hetero* means "different" and a heterojunction laser is one that has dissimilar materials for its junction, particularly of different bandgaps. Here, we shall look at double heterostructure (DH) lasers and graded-index separate confinement heterostructure single-quantum-well (GRINSCH-SQW) lasers (GRINSCH is pronounced "grinch").

3.9.1.1 DH Lasers

Semiconductor lasers were able to make the tremendous leaps in performance mainly on the basis of one semiconductor material system: $Al_xGa_{1-x}As$. This semiconductor system is remarkable because the lattice constants of GaAs and AlAs are almost identical, and, therefore, $Al_xGa_{1-x}As$ semiconductor crystals of extremely high quality can be grown on a GaAs substrate.

DH lasers are formed by sandwiching a layer of GaAs—the core—between two layers of AlGaAs—the claddings. As we will see later in this section, optical feedback can be provided for the laser cavity through various means. Figure 3.23 illustrates optical confinement, carrier confinement, and gain in the DH structure. Optical confinement occurs because AlGaAs has a lower index of refraction than GaAs. As we saw in section 3.2.3, sandwiching a material between two lower index of refraction materials can lead to waveguiding or optical confinement. Carrier confinement results from AlGaAs having a larger bandwidth than GaAs, as well as favorable band offsets between GaAs and AlGaAs. Figure 3.23 shows how the conduction band and valence bands are able to confine the carriers (i.e., electrons and holes) in the active region (GaAs layer). A p–n junction is usually formed to allow injection of carriers with an applied voltage, V_{applied}. When $V_{\text{applied}} \approx E_g(\text{GaAs})/q$, injected electrons and holes overlap each other in the GaAs active region. This leads to population inversion or gain.

DH lasers are the most common material structure for a semiconductor laser and is found in many applications such as CD players, CD-ROM drives, laser printers, and sources for optical fibers.

As discussed in section 3.2.3, the confinement factor is given by the shaded region under the curve divided by the total area under the curve in Figure 3.23 and is usually approximately 50% for DH lasers. The confinement factor is a measure of how much the stimulated emission-generating region of the laser structure is coupled with the total optical mode that the structure supports.

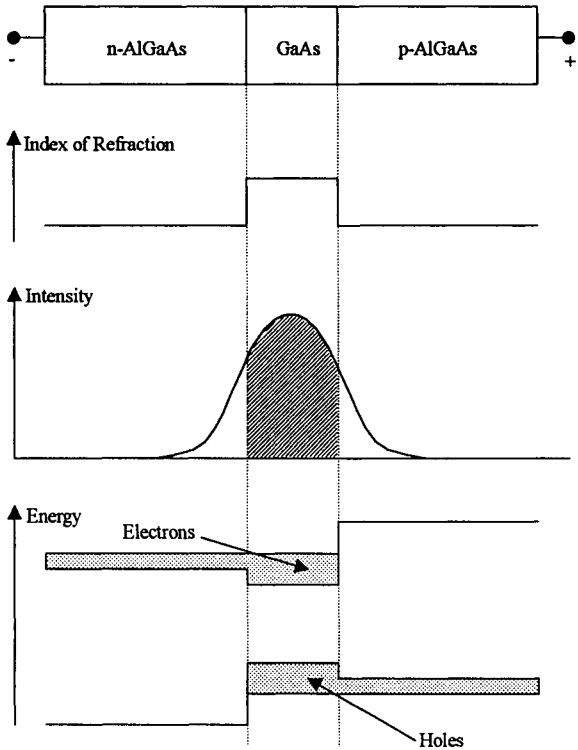


FIGURE 3.23 DH laser.

3.9.1.2 GRINSCH-SQW Lasers

Some of the most advanced lasers make use of a structure known as GRINSCH-SQW. Figure 3.24 shows an example of such a structure using AlGaAs. This structure provides better optical and carrier confinement, and because of the use of quantum wells, higher gains can be obtained compared to DH structures. Some typical thickness values are shown in Figure 3.24a together with dopings to form a p–n junction. The optical mode associated with the structure appears in Figure 3.24b.

A typical GRINSCH-SQW structure has the following layers (from left to right in Fig. 3.24a):

- substrate
- lower cladding
- lower graded region
- quantum well
- upper graded region
- upper cladding
- contact layer

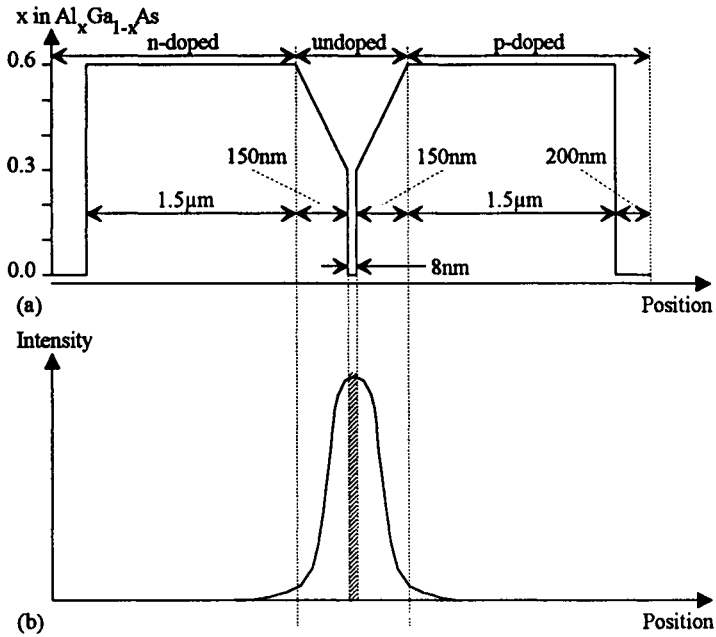


FIGURE 3.24 GRINSCH-SQW laser.

The confinement factor is very small in GRINSCH-SQW laser structures and is typically about 2%. Quantum wells have large optical gains associated with them that compensate for their small confinement factors.

The quantum well thickness and material composition determines the lasing wavelength of the semiconductor laser. By using a *strained* quantum well—such as the InGaAs quantum well in the preceding structure—wavelengths not available through DH structures can be obtained. Although the quantum well is not lattice-matched to the rest of the laser structure, as the typical thickness of a quantum well is very small (4 to 15 nm), the lattice-mismatch may be accommodated elastically. Growing a thick layer of lattice-mismatched material on the same structure would result in misfit dislocations that would rapidly degrade the performance of the laser.

3.9.2 The Semiconductor Laser Cavity

In order to turn a semiconductor structure into a working laser device, we need to fabricate a laser cavity. Here, we look at several different cavity types, from the simplest to the most complex.

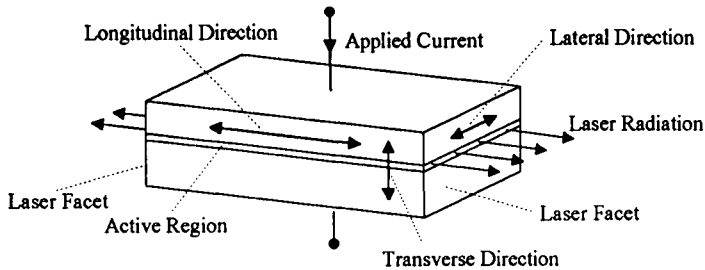


FIGURE 3.25 Broad-area laser cavity.

3.9.2.1 Broad-Area Lasers

Broad-area lasers are the simplest kind of semiconductor laser cavities. These laser cavities were the first ones formed. Figure 3.25 shows this structure and the way in which electric current is applied.

Early on, it was observed that facets—or semiconductor laser mirrors—could be formed through the cleaving of the semiconducting crystal. The cleavage planes of the semiconductor crystal is ideal for this purpose. As we saw in section 3.2.1, when light travels from a medium of index n_1 to n_2, Eq. (3.1) gives the reflectivity for perpendicular incidence. So, if in Figure 3.1 we have GaAs for medium 1 ($n_1 = 3.66$) and air for medium 2 ($n_2 \approx 1$), then

$$R = \left(\frac{3.66 - 1}{3.66 + 1}\right)^2 = \left(\frac{2.66}{4.66}\right)^2 = 0.326 = 32.6\%$$

One of the problems, however, encountered with cleaving was that the sides perpendicular to the facets also behaved as facets and internal circulating modes were observed (see Fig. 3.26). Internal circulating modes cause loss of efficiency in the semiconductor laser [10]. The solution to this problem was to have rough sides and thereby suppress the internal circulating modes. The roughening was obtained by sawing the nonfacet sides—instead of cleaving. This process is still used in manufacturing broad-area lasers.

Figure 3.25 shows the longitudinal, transverse, and lateral directions in the laser cavity. The cavity length effects the longitudinal modes of the laser seen

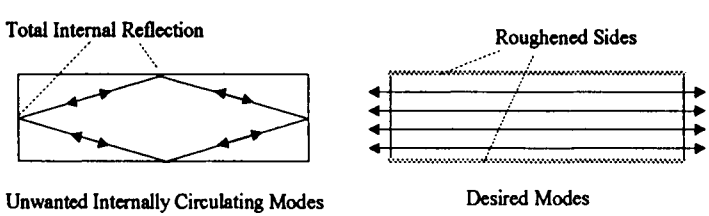


FIGURE 3.26 Illustration of problem with cleaving broad-area laser sides.

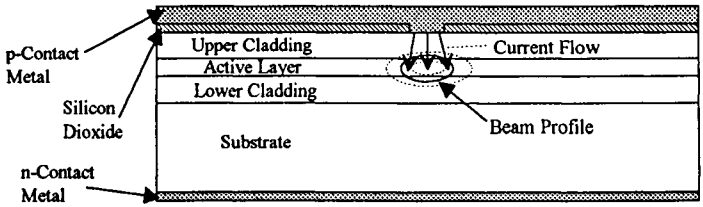


FIGURE 3.27 Stripe laser.

in the laser spectra. The epitaxial growth of the semiconductor structure determines the transverse mode. Because no attempt is made to confine light in the lateral direction, the structure in Figure 3.25 is known as a broad-area laser. In the next section we look at one of the first ways devised to confine light laterally.

3.9.2.2 Stripe Lasers

Stripe lasers provide lateral confinement of the optical laser beam by preventing current flow from everywhere except from a small defined region, as shown in Figure 3.27. Because current is needed to maintain the population inversion in the semiconductor laser, only regions pumped by current flow experience gain and may achieve laser action. There are a large number of ways semiconductor lasers can be fabricated to allow this lateral current confinement. The simplest method, known as oxide isolation, allows the top electrode to make contact with the semiconductor in a small stripe along the length of the laser (Fig. 3.27 shows the cross-sectional view). Other techniques such as proton-bombardment or oxygen implantation, zinc diffusion, and buried heterostructure can be used to obtain similar lateral current confinements [11].

The beam profile or shape of these lasers tends to change as the pumped current increases (beam profiles with dotted lines in Fig. 3.27). The laser structure presented in the next section overcomes this problem by having the beam shape defined by the index of refraction rather than by the pumped current.

3.9.2.3 Ridge Lasers

Ridge lasers are another type of semiconductor laser that offers lateral confinement of laser light. The lateral confinement is obtained by lowering the index of refraction in regions adjacent to the ridge. This lowering of the index allows lateral waveguiding, similar to what we saw earlier in section 3.2.5. Figure 3.28 shows a ridge laser and its associated beam profile.

By etching away regions outside of a ridge and covering those areas with a low dielectric material such as silicon-dioxide or silicon-nitride, the index of refraction is lowered and leads to lateral optical confinement.

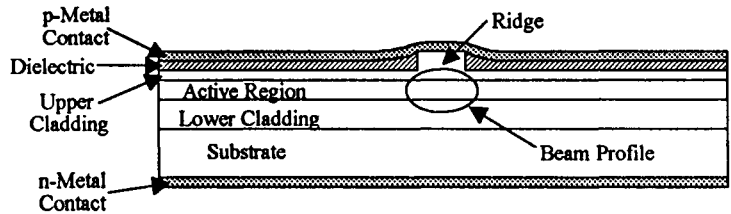


FIGURE 3.28 Ridge laser.

Example

Use the effective index technique (section 3.2.5) to predict how the horizontal near-field (that is, the near-field profile in the lateral direction) for a ridge laser varies as a function of the parameters T and w (shown in Fig. 3.5). Assume only fundamental (or lowest order) mode operation.

Because we have only studied the effective index technique in a qualitative manner, we will answer this question in the same manner. We can use Figure 3.29 to show the effect of the variation of T on the horizontal near-field, with w held constant.

Referring back to section 3.2.5, we find that the effective index is lower for smaller T values (note that the maximum value of T is d_1 in Fig. 3.6). So as T gets smaller, the difference between the effective index under the ridge and the regions adjacent to the ridge becomes larger. This results in tighter optical confinement and, therefore, smaller horizontal near-fields.

For the case of T constant and w varying, the situation gets a little more complicated. Initially, as expected, as the ridge gets smaller; we find tighter lateral confinement until the ridge becomes too small to contain the mode and the horizontal near-field increases (Fig. 3.30). The optical mode requires a certain physical region and if it is *squeezed* too much with the ridge it will tend to expand outside the ridge.

3.9.2.4 Vertical Cavity Lasers

Vertical cavity lasers or microlasers are some of the latest and smallest semiconductor laser structures that have ever been built. Unlike the lasers we have looked at so far, they do not have cleaved facets and obtain their optical feedback though Bragg reflectors (section 3.2.2). Because these lasers have

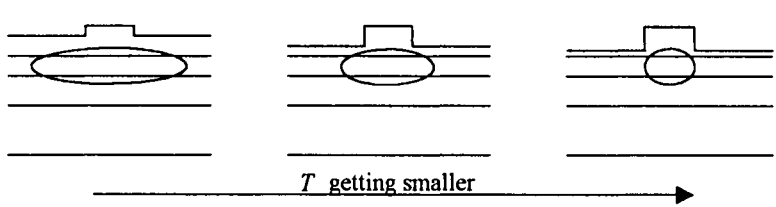


FIGURE 3.29 Horizontal near-field variation with T.

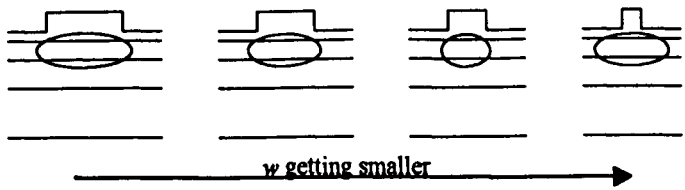


FIGURE 3.30 Horizontal near-field variation with w.

their reflectors built-in, a large number of microlasers can be fabricated side-by-side [12]. A cross-sectional view is shown in Figure 3.31 where the active region and the reflectors are labeled.

Because microlasers are typically only a few microns in length, the reflectivity of the Bragg reflectors have to be 99% and more. This allows sufficient optical feedback to compensate for the short cavity length of the microlaser.

Another unique aspect of microlasers is that they have a cylindrical shape. This avoids the need for a separate mechanism to provide lateral guiding. The laser cavity is like a miniature optical fiber that guides the beam and prevents its escape from the sides of the cylinder. With small enough diameters, these lasers will have one mode that oscillates between the mirrors.

3.9.2.5 Distributed Feedback Lasers

Distributed feedback (DFB) lasers are also semiconductor lasers that use Bragg reflectors instead of cleaved facets to obtain optical feedback. They have excellent wavelength selectivity and tend to operate in a single longitudinal mode. The reason for such selectivity arises from the structure of this laser, which contains a first-order Bragg reflector along the length of its cavity [13]. This reflector is formed by etching gratings in the semiconductor laser structure, as shown in Figure 3.32. The entire length of the cavity is pumped by applying current through the p-contact metal.

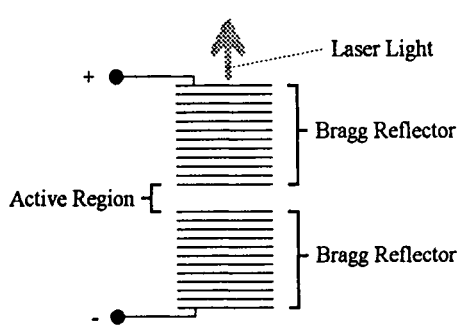


FIGURE 3.31 Vertical cavity laser.

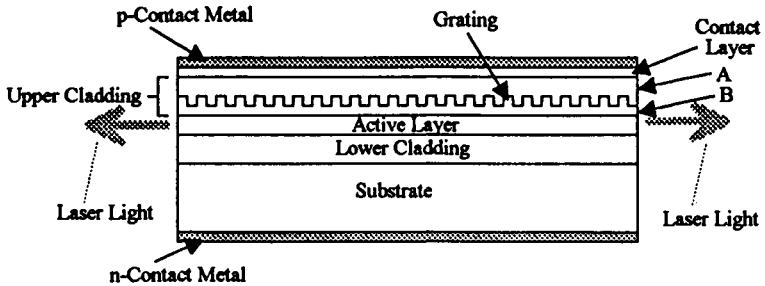


FIGURE 3.32 DFB laser.

In Figure 3.32, the upper cladding layer contains the grating and is made up of two different semiconductors, labeled A and B. For example, if the substrate was GaAs, a possible candidate for A could be $Al_{0.1}Ga_{0.9}As$, and $Al_{0.3}Ga_{0.7}As$ for B. The difference in the index of refraction between A and B and the distance of the grating from the active region determines the extent of the optical feedback.

3.9.3 Laser Characteristics

A semiconductor laser is an excellent choice for optical fiber applications requiring:

- high power (3 to 100 mW)
- efficient coupling
- very narrow spectral width
- very fast rise time

The downside is that semiconductor lasers that are used in optical communication tend to be expensive. Temperature fluctuations of the laser can shift their operating wavelength and power output level. Lasers are also very susceptible to optical reflections from the optical fiber or other optics in the system back into the laser.

3.9.3.1 Power versus Current Characteristics

Two key parameters associated with any semiconductor laser is its efficiency of converting electrical current into laser light and the amount of current required before the laser reaches stimulated emission, known as the *threshold current*, I_{th}. The power versus current (P vs. I) characteristics allows this information to be extracted from a semiconductor laser. Figure 3.33 shows a typical example of P versus I characteristics. The slope of the linear segment above threshold is the *slope efficiency* (dP/dI) that is related to *external differential quantum efficiency*, η_{diff}, through the following equation:

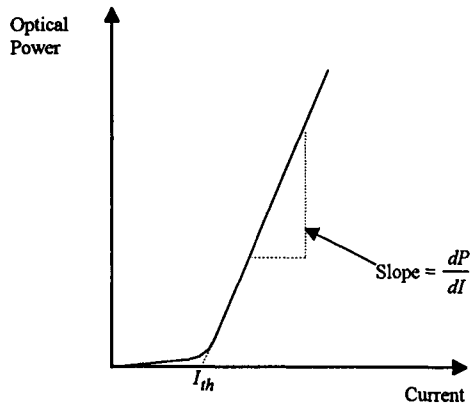


FIGURE 3.33 Power versus current characteristics for a semiconductor laser.

$$\eta_{\text{diff}} = \frac{\lambda_0 q}{hc}\left(\frac{dP}{dI}\right) \tag{3.22}$$

The threshold current I_{th} is found by extrapolating the linear segment down to $P = 0$ (intersection with the I axis). Low threshold currents and high efficiencies are desirable in semiconductor lasers.

The applied current initially gives rise to spontaneous emission from the laser device. As this current is increased, the cavity and mirror losses are overcome (section 3.8.2) and the laser achieves threshold and starts emitting stimulated emission. The optical power from the laser above I_{th} is given by:

$$P = (I - I_{th})\frac{dP}{dI} \tag{3.23}$$

where I is the operating current.

3.9.3.2 Threshold Current versus Temperature Characteristics

Semiconductor lasers are generally very sensitive to temperature changes. The I_{th} as a function of laser-junction operating temperature θ is given by:

$$I_{th}(\theta) = I_{th}(\theta_1) \times \exp\left(\frac{\theta - \theta_1}{\theta_0}\right) \tag{3.24}$$

where θ_0 is the characteristic temperature and θ_1 is usually room temperature (300 K). A typical I_{th} versus θ plot appears in Figure 3.34, from which the parameter θ_0 can be extracted.

A large value for θ_0 is usually desirable in semiconductor lasers because it translates into a lower sensitivity of the threshold current to the operating temperature of the laser-junction.

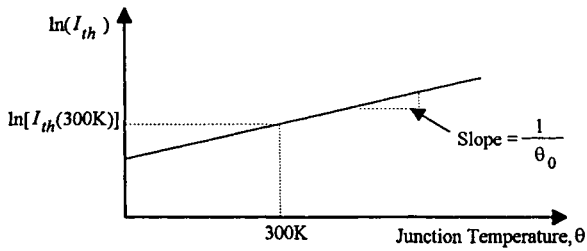


FIGURE 3.34 Natural logarithm of threshold current versus junction temperature.

3.9.3.3 Spectral Characteristics

A plot of laser intensity wavelength gives rise to the spectral characteristics or *spectra.* The spectra gives information about the linewidth of the laser and whether it operates in single-longitudinal mode or multilongitudinal mode. A DFB laser typically operates in single-longitudinal mode, in which its spectra reveals only one "line" at the lasing wavelength. A Fabry-Perot laser, on the other hand, tends to operate in multilongitudinal mode and has several lines in its spectra.

3.9.3.4 Near-Field and Far-Field

We looked at near-field and far-field in section 3.2.4. These characteristics allow us to determine whether the laser is operating in fundamental mode or whether non-fundamental mode behavior is creeping in. They also allow us to determine how efficiently laser light can be coupled into an optical fiber.

General References

G. E. Neudeck and R. F. Pierret (1983). The PN junction diode. *Modular Series on Solid State Devices* (Vol. 2). Reading, MA: Addison-Wesley.

C. Hentschel (1988). *Fiber Optics Handbook,* 2nd ed. Germany: Hewlett Packard.

M. Fukuda (1991). *Reliability and Degradation of Semiconductor Lasers and LEDs.* Boston: Artech House.

G. P. Agrawal and N. K. Dutta (1986). *Long-Wavelength Semiconductor Lasers.* New York: Van Nostrand Reinhold.

S. R. Forrest (1988). Optical detectors for lightwave communication. In S. E. Miller and I. P. Kaminow (eds.), *Optical Fiber Telecommunications II.* New York: Academic Press.

References

1. H. A. Macleod (1986). *Thin-Film Optical Filters.* New York: Macmillan, pp. 17–27.
2. M. V. Klein and T. E. Furtak (1986). *Optics,* 2nd ed. New York: John Wiley & Sons, pp. 71–86.

3. H. C. Casey and M. B. Panish (1978). *Heterostructure Lasers: Part A.* New York: Academic Press, pp. 20–109.
4. J. Buus (1982). *IEEE J. Quantum Electron.* **QE-18,** 1083.
5. T. H. Maiman (1960). *Nature* **187,** 493.
6. D. L. Matthews and M. D. Rosen (1988). *Scientific American* **259**(6), 86.
7. J. T. Verdeyen (1981). *Laser Electronics.* Englewood Cliffs, NJ: Prentice-Hall Inc., pp. 252–299.
8. R. N. Hall (1976). *IEEE Trans. Electron. Dev.* **ED-23,** 700.
9. H. C. Casey and M. B. Panish (1978). *Heterostructure Lasers: Part A.* New York: Academic Press, pp. 1–19.
10. M. Ettenberg, H. F. Lockwood, and H. S. Sommers, Jr. (1972). *J. Appl. Phys.* **43,** 5047.
11. H. C. Casey and M. B. Panish (1978). *Heterostructure Lasers: Part B.* New York: Academic Press, pp. 207–217.
12. J. L. Jewell, J. P. Harbison, and A. Schere (1991). *Scientific American* **265**(5), 86.
13. A. Yariv (1984). *Optical Waves in Crystals.* New York: John Wiley & Sons, pp. 439–447.

Chapter 4

Integrated Circuits, Transceiver Modules, and Packaging

Jerry Radcliffe
BroadBand Technologies
Research Triangle Park
North Carolina

Carolyn Paddock
IBM Corporation
Essex Junction, Vermont

Ronald Lasky
Universal Instruments Corporation
Binghamton, New York

4.1 Basic Link Overview

The two main considerations in the design of fiber-optic links are the system loss budget and the link bandwidth. Because of the typical modest distances of a data communication system, the loss budget will typically be rather small compared to telecommunication standards. However, because of the multiple manufacturer interoperability requirements, the loss budget must be met by a wide variety of components from a large number of manufacturers. This situation does not allow for a high degree of optimization in the link design. However, it does allow for advantageous use of statistical design.

The link bandwidth is often constrained by the cost pressures of data communication. There is great pressure to use the lowest cost cable plant, which usually implies the lowest bandwidth fiber. The fiber choice is often not driven by the fiber characteristics but by the cost of

Optoelectronics for Data Communication

connectorization. For instance, large core step-index fibers are perceived to have significantly lower connector costs resulting in a lower overall system cost. And while this situation may be true in a specific instance, care must also be taken to consider future system upgrades which may require the same cable plant to carry higher data rate traffic or to be patched in as part of a longer distance link. Hence in the overall view, the fiber may be the least expensive part of the system. In the long run, the greatest cost savings will be realized by putting in place the highest bandwidth system.

In a data communication system, the requirements are usually stated in terms of a communication distance requirement. However, individual installations vary greatly in their configurations and the number of connectors, patch panels, splices, and fiber runs are not initially known. The solution to this problem is to do a link design based on a reasonable link configuration to meet the distance using conservative fiber, connector, and splice specifications. This aproach allows the majority of system users to achieve the link distance with economical industry grade components. The few users who require performance in excess of these minimum requirements may use higher grade fiber or connectors and meet their requirements at a slight cost increase.

4.1.1 Industry Standards

To achieve the greatest utility, a data communication link design should rely on generally available industry standards. A number of fiber specifications are in use. The most common are those of Bellcore and the Electronics Industries Association (EIA). These bodies publish standards for fibers for both telecommunication and data communication applications [1]. In addition to these industry specifications, additional specifications may be found in some standards, such as the FDDI standard, which specifies a fiber with greater bandwidth than the industry standard. In addition, some system manufacturers, such as IBM ESCON, specify fibers with higher bandwidth or lower losses than the standards.

These standards are usually not a problem. The process of arriving at standards almost ensures that they are some time interval behind the industry's most advanced practice. Fibers of superior performance may be easily had, although at some additional cost. Using the standard specifications ensures ease of multiple sourcing. If a design is carried out with specifications meeting the standards, extraordinary situations and specific applications may be easily met by the use of relatively easily obtained better grade fibers.

4.1.2 Loss Limited Design

The most common link description is in terms of a loss budget. This design approach considers the permissible optical power loss in the transmission system and is the arithmetic difference between the transmitted optical power

and the receiver sensitivity. It is composed of both the attenuation of the fiber and the losses in the connectors and splices. In the past, this design has often been done using worst-case design methodologies. However, this method may lead to excessively restrictive and expensive designs. The fiber-optics industry has matured to the extent that common standards exist that allow the use of statistical design instead of a worse-case approach.

Using a statistical approach, the system loss budget constraint requires that the loss budget exceed the mean loss of the cable plant by at least three standard deviations.

$$LB > \mu_L + 3\sigma_L \tag{4.1}$$

where: LB is the loss budget, μ_L is the mean link loss, and σ_L is the standard deviation of link loss.

The link losses are given by:

$$\mu_L = l_t\mu_c + N_s\mu_s + N_{CON}\mu_{CON} + \mu_{cl} \tag{4.2}$$

and

$$\sigma_L^2 = l_R l_t \sigma_c^2 + N_s\sigma_s^2 + N_{CON}\sigma_{CON}^2 + \sigma_{cl}^2 \tag{4.3}$$

Where:

l_t = Total length of the cable plant
l_R = Average reel length of the cable
μ_c = Mean cable loss
σ_c = Standard deviation of cable loss
N_s = Number of splices
μ_s = Mean splice loss
σ_s = Standard deviation of splice loss
N_{CON} = Number of connectors
μ_{CON} = Mean connector loss
σ_{CON} = Standard deviation of connector loss
μ_{cl} = Mean excess coupling loss due to mode field mismatch in the cable plant
σ_{cl} = Standard deviation of excess coupling loss

A cable splice should be assumed to be located at each end of the trunk and at one per kilometer over the length of the trunk. In addition, two splices are assumed for future repair and change activity. This results in an effective reel length of 1 km. The number of splices in the link is given by:

$$N_s = 3 + \frac{l_t}{l_R} \tag{4.4}$$

where N_s is an integer rounded upward.

An example of a loss limit analysis may be found in the Fiber Channel Standard (FCS) [2]. This standard addresses four data rates and numerous distances and communications technologies. It is strongly desirable to have installed portions of the fiber system be appropriate for future upgrades. As an example, we examine the requirements for single-mode (SM) links. These

TABLE 4.1 Examples of single-mode cable plant parameters used for calculating loss budgets

Parameter	Symbol	Value
Cable	μ_c	0.50 dB/km
	σ_c	0.0 dB/km
Splice	μ_s	0.15 dB/spl
	σ_s	0.1 dB/spl
Connector	μ_{CON}	0.60 dB/con
	σ_{CON}	0.20 dB/con
Additional coupling loss	μ_{cl}	0.40 dB
	σ_{cl}	0.0 dB

links operate using 1300 nm light from MLM lasers with distance requirements of 2 and 10 km. The components have the conservative values listed in Table 4.1.

Substituting values from Table 4.1 into Eqs. (4.2) and (4.3) gives

$$\begin{aligned}\mu_L &= 0.50l_t + 0.15(3 + l_t) + 0.60N_{CON} + 0.40 \\ &= 0.65l_t + 0.85 + 0.60N_{CON}\end{aligned}$$

and

$$\begin{aligned}\sigma_L^2 &= I_R l_t(0)^2 + (3 + l_t)(0.10)^2 + N_{CON}(0.20)^2 + (0)^2 \\ &= 0.01l_t + 0.04N_{CON} + 0.03 \qquad (4.5)\end{aligned}$$

The standards committee believed that four and eight connectors were reasonable for the distances of 2 and 10 km. The resulting calculations for the two distances are shown in Table 4.2.

4.1.3 Bandwidth Limited Design

In a fiber-optic link, the distance may be either loss limited or bandwidth limited. The bandwidth is usually the limiting case for LED (light emitting diode) data links utilizing multimode (MM) fiber.

For MM LED data links, an effective cable bandwidth is calculated from the modal bandwidth and chromatic dispersion (intramodal) effects. The length limit is then calculated from this effective cable bandwidth. The

TABLE 4.2 Loss budgets for Fiber Channel 2- and 10-km single-mode links

Distance	No. Connectors	$\mu + 3\sigma$	LB
2 km	4	5.92 dB	6 dB
10 km	8	14.16 dB	14 dB

effective system bandwidth can be represented by the relationship of intramodal and modal bandwidth of a given span as shown:

$$\frac{1}{BW_{effec}^2} = \frac{1}{BW_{modal}^2} + \frac{1}{BW_{intra}^2} \tag{4.6}$$

where BW_{modal} is the bandwidth of a fiber limited by the pulse broadening due to optical power traveling via different waveguide modes. This bandwidth is specified for the cable. BW_{intra} is the bandwidth of a fiber limited by the pulse broadening due to chromatic dispersion (the speed of an optical pulse changes as its wavelength changes).

The chromatic dispersion of a fiber can be calculated by using the equation:

$$D(\lambda) = \frac{S_0}{4}\left[\lambda - \frac{\lambda_0^4}{\lambda^3}\right] \quad \text{ps/nm} \cdot \text{km} \tag{4.7}$$

where $D(\lambda)$ is the chromatic dispersion in ps/nm·km, S_0 is the dispersion function slope at λ_0, λ_0 is the zero dispersion wavelength, and λ is the wavelength of operation.

For LEDs operating near 1300 nm:

$$BW_{intra} = \frac{18\,700}{\sigma_{intra}} \tag{4.8}$$

$$\sigma_{intra} = L\sqrt{D(\lambda)^2(\Delta\lambda \times K)^2 + \frac{S_0^2(\Delta\lambda \times k)^4}{2}} \tag{4.9}$$

where σ_{intra} is the RMS pulse broadening in psec, k is 0.425 (the conversion factor from FWHM to RMS), $\Delta\lambda$ is the FWHM source spectral width in nanometers, L is length in kilometers, and S_0 is the dispersion slope in ps/nm^2·km.

Once again using the Fiber Channel Standard as an example, the 138-Mbit, 1300-mn LED link specifies a spectral width of 250 nm FWHM. Using preceding equations, this will support a distance of 1.5 km.

4.2 Cable/Connector

As mentioned previously, one of the significant differences between opto-data communications and opto-telecommunications is the fact that most cable-to-transceiver connections will be made by untrained personnel. Therefore it is necessary for the insertion and proper connection of the connector into the transceiver to be user friendly and foolproof. These requirements have resulted in the development of a user friendly, keyed, duplex connector discussed in section 2.8. The SC (super convex) duplex connector in both SM and MM is one of the most popular because of its strong performance and lower cost. Figure 6.5 in Chapter 6 shows an example

of this type of duplex connector. Recently, several organizations have proposed standards for duplex connectors, the details of which are discussed in Chapter 2. Even with these standards, the issues of multisourcing can create difficulties as the intermateability of vendor A's transceiver and vendor B's duplex connector may be a hit-or-miss proposition with regard to mechanical plugability and coupling of light. These mateability issues are discussed in greater detail in Chapter 5.

4.2.1 Optical Sources

The physics and technical details of optical sources are given in Chapter 3. In this section, the implementation issues relating to optical sources are discussed. The choice of optical source depends on the data rate, optical medium, link length, bit error rate (BER), laser safety requirements, and link cost. Today, the optical source to choose from include the following.

Surface emitting LEDs (SE-LEDs) in the visible light or at 1.3 μm (infrared) are available. The visible light LEDs can be used as a source for 0.5 to 1.0-mm plastic fiber. The large modal dispersion of this plastic fiber limits the bandwidth of this link. Therefore, plastic fiber MM configurations are useful for short distances at speeds up to 200 MHz or for longer links at low transmission rates in the 20 to 40 MHz range. The large fiber core allows direct coupling to the optical source, providing an inexpensive optical subassembly. Infrared LEDs at 1.3 μm are useful sources for 50- or 62.5 μm MM fiber. Because the source is incoherent, modal noise is not a concern. The emission pattern from an SE-LED is lambertian, hence a ball lens is often used to couple the light into MM fibers. As discussed in Chapter 3, the modulation bandwidth of low cost, SE-LED is limited to approximately 200 MHz [3]. Electrical pulse shaping in the drive circuit design can achieve some improvement in the bandwidth. LEDs have the advantage of low modulation current requirements and higher reliability than semiconductor lasers.

Edge emitting LEDs (EE-LEDs) at 1.3 μm have higher modulation speeds and increased coupling efficiency over SE-LEDs. A 350-MHz bandwidth has been demonstrated for edge emitting LEDs [4]. With special drive circuitry, modulation speeds up to 1.2 Gb/s have been demonstrated in SM fiber. The spectrum of the EE-LED is narrower than the SE-LED, but still much broader than a laser because the enhanced waveguiding property of the device [4]. Optical coupled power of the EE-LED is 2 to 3 dB greater than the coupled power from an SE-LED. As a comparison, the coupled power into a SM fiber was 30% and 2% for an EE-LED and an SE-LED, respectively [4].

Compact disc (CD) lasers (780 to 850 nm) typically are designed with self-pulsation frequencies between 1 and 2 GHz. The self-pulsation broadens the spectral width of the laser and decreases the sensitivity to back-reflections.

This property is highly useful for CD players. The broadened spectral width is also useful to minimize modal noise in a MM optical link. To avoid excessive ringing on the falling edge of the laser pulse, the relaxation frequency of the laser should typically be 2.5 to 3 times the modulation frequency. This requirement increases the cost of CD lasers for modulation frequencies greater than 400 Mb/s. The relaxation frequency of a laser increases with bias current, but higher bias currents decrease laser reliability. Recently, several manufacturers have demonstrated short-wavelength optical links at 1 Gb/s using non-self-pulsating lasers. These lasers are still low cost compared with 1.3 μm lasers because of the high volume used for CDs and laser printers. Optical links with these sources may still be susceptible to modal noise, therefore, BER concerns versus link cost will govern whether they should be implemented. The reliability of the CD laser is lower than the long-wavelength 1.3-μm laser [5]. To improve this situation, device structure and burn-in techniques can be used to improve reliability to meet data communication product needs.

The optical coupled power is determined by laser safety concerns and packaging cost allowance. To increase coupling efficiency, ball lenses can be used to provide 30 to 50% coupling, but add the cost of a lens and active alignment. Commercial transistor outline (TO) cans with ball lenses have been introduced to reduce package cost but may not provide adequate control over the coupled power variations because of laser-to-lens and lens-to-fiber distance variations. The Fiber Channel Standard limits the maximum coupled power in a short-wavelength laser link to 1.3 dBm. An open fiber control circuit is incorporated to ensure compliance with world-wide laser safety standards [6]. This circuit actively samples the link to determine whether the link has been disrupted. The laser is turned off if the connection between the transmitter and receiver is broken, thereby minimizing the possible exposure to laser light.

Lasers of 1.3 μm are usually used in SM fiber links. At moderate data rates, such as the ESCON 200 Mb/s link, this source could be used for 10 to 20 km links. At higher modulation frequencies, such as 1 Gb/s, the link distance is limited to 2 km by the Fiber Channel Standard to maintain a 10^{-12} BER [6]. Optical coupling into a 9-μm core fiber is usually accomplished with a GRIN lens, a ball lens in the TO can, or a lensed fiber. The packaging usually requires two or three "active" (laser turned on, coupled light measured) alignments, which are expensive and time consuming. Laser safety can be accomplished with an open fiber control circuit or by minimal coupling into the fiber. An example of this type of package is a transmitter optical subassembly (TOSA), as seen in Figure 4.1. The assembly of this package will be discussed in greater detail in section 4.5.

Highly coherent distributed feedback (DFB) or distributed Bragg reflection lasers at 1.55 μm light are used for very long distance links or links incorporating a wavelength division multiplexing scheme. Typically these

sources have been avoided in data communication products because of the high cost and complexity of the links. These sources are highly sensitive to intensity noise and therefore optical coupling assemblies must be designed to minimize back reflections into the laser. An example of this DFB laser is shown in Figure 3.32 in Chapter 3.

4.2.2 Optical Detectors

The choice of detector usually depends on the incident wavelength and intensity of light, the modulation frequency, bias voltage, coupling mechanism, and back reflection. For data communications, the typical choice of

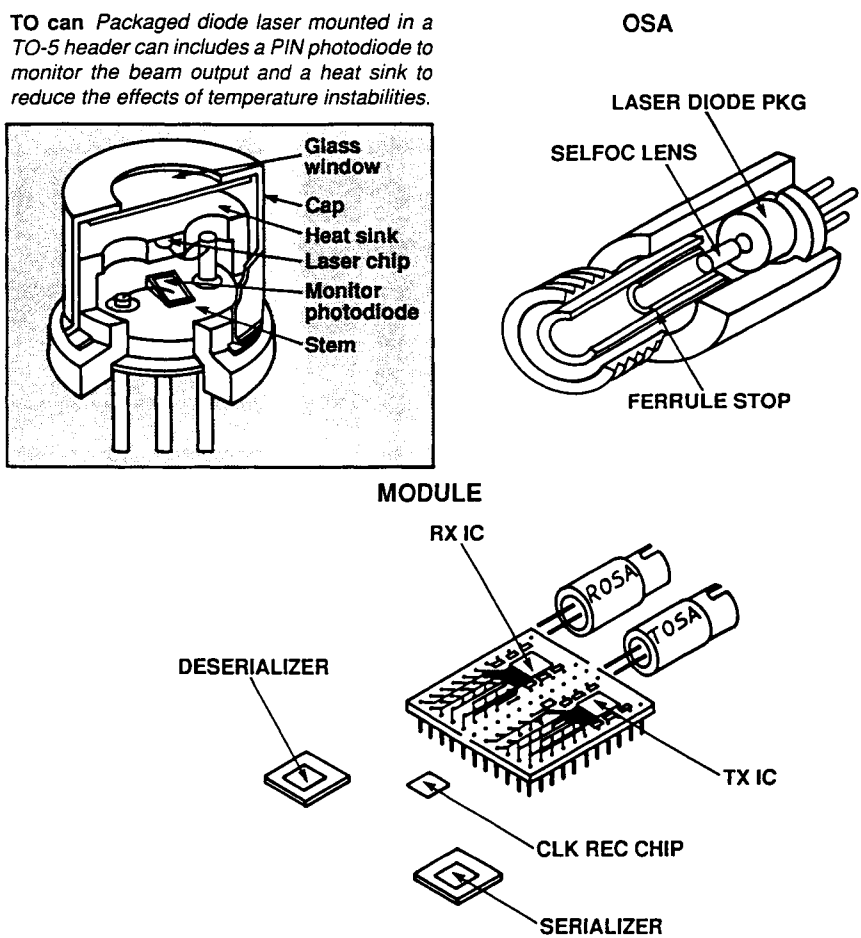


FIGURE 4.1 Common solution today: Laser TO can → optical subassembly (OSA) → module. Three levels of packaging; several active alignments.

detector is a semiconductor pin (p-type/intrinsic/n-type) photodiode in either silicon or gallium arsenide. An avalanche photodiode is often considered too expensive for this application. An important measurement of a photodiode is the responsivity, which is defined as

$$R = \frac{I_p}{P_o} \text{ A/W} \tag{4.10}$$

where I_p is the photocurrent generated, P_o is the incident optical power, and A/W is the amps/watt.

The responsivity depends on the incident wavelength of light and bandgap energy of the photodiode. The responsivity of a receiver optical subassembly (ROSA), the receiver version of a TOSA shown in Figure 4.1, is typically less than the ideal responsivity of the photodiode. It is degraded by the coupling efficiency of the fiber to the detector and any antireflective coatings applied to the window or photodiode surface. Silicon has a responsivity of 0.4 to 0.6 A/W in the short-wavelength region of 780 to 850 nm. The responsivity is slightly lower in the visible part of the spectra and has a sharp drop-off above 1.1 μm. Hence, gallium arsenide detectors are preferred for 1.3 μm or 1.55 μm wavelength links. The responsivity of a gallium arsenide p-i-n detector is 0.8 to 0.9 A/W at 1.3 μm. If a dual source detector is desired for both short- and long-wavelength links, a gallium arsenide detector would be suitable. The typical responsivity of GaAs photodiodes at 850 nm is 0.1 to 0.2 A/W. The devices can be designed for improved responsivity in the short-wavelength region up to approximately 0.4 A/W. In this case, the responsivity at the long wavelength is decreased. In a dual source detector, the transmissivity of any AR coating applied to either the TO can window or the device itself must also be adequate for both wavelength regions. If an optical link has a demanding requirement for receiver sensitivity, an integrated pre-amp may be incorporated in the TO package. The pre-amp may be either silicon or GaAs and may be incorporated with either integrated or hybrid techniques.

The active area of a photodiode is important for determining the method of optical coupling to the device. GaAs detectors in TO cans for communications typically have small active areas, often 100 μm^2. Efficient coupling is achieved with a lens, typically a GRIN or ball lens. Several manufacturers provide lensed caps to the TO cans thereby reducing the number of elements in the ROSA. Silicon photodiodes have larger active areas, typically 200 to 1000 μm^2. Direct coupling from MM fiber can be quite efficient for these devices. The capacitance of a photodiode is a concern for both performance and noise considerations. The device capacitance is directly related to the area and thickness of the active region. High speed silicon photodiodes are available with active regions between 200 and 400 μm^2. The speed of the device increases as the wavelength increases below the cutoff frequency because of the increased absorption in the active region. Increasing the bias

voltage voltage on the device also improves performance by decreasing the transport time of the electrons from the active region and decreasing the capacitance. A significant improvement in the rise and fall time of short-wavelength Gb/s pulses is exhibited as the bias voltage is increased. The 20 to 80% rise time can decrease from 1100 ps to 300 ps by increasing the voltage from 3.6 V to 12 V. GaAs detectors with smaller active areas have lower capacitance and correspondingly faster rise times. The 20 to 80% rise time of a 100 μm^2 device with 3.6 V reverse bias is less than 200 ps.

4.2.3 Transceiver Electronics

Figure 4.2 shows a block diagram of the typical electronic components necessary for transceiver function. These components can be integrated or discrete. Parallel electronic data is often encoded to provide balanced code to the transceiver. The data are then sent to a multiplexer, often called a

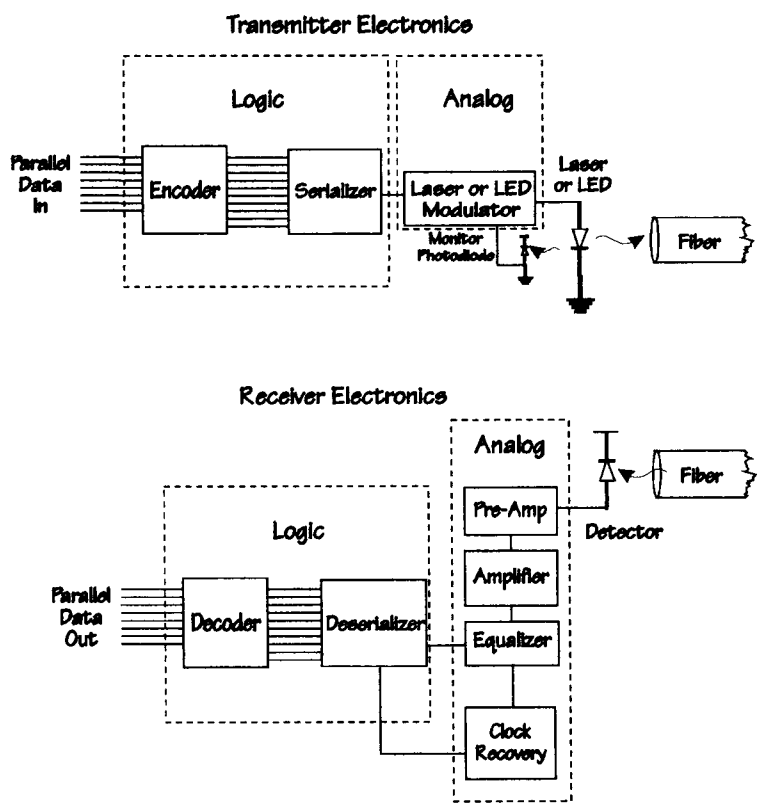


FIGURE 4.2 Block diagram of transceiver electronics.

serializer, to produce a serial data stream. In optoelectronic transceivers, this stream is next sent to the modulation circuit, which is tailored to drive either a laser or an LED. The receiving electronics receives a current from the optical detector. The current is amplified and, if necesary, an equalization filter is used to reduce signal distortions. The clock is recovered from the signal and used to time the demultiplexer, often called a deserializer, to produce parallel data. The data are decoded, if necessary, and the parallel output is sent out for further processing by the computer.

Transceiver electronics can be designed in either GaAs or silicon. The choice of electronic material depends on the speed, cost, power dissipation, and degree of integration required for the optoelectronic transceiver. GaAs is often used for high performance, high speed circuits because of the lower power, low noise transistors available in that technology. Silicon circuits, used in transceivers, can be designed in either bipolar, analog CMOS, or BiCMOS technologies. Bipolar silicon technology usually contains high speed NPNs, lateral PNPs, possibly high speed vertical PNPs, resistors, and capacitors. Analog CMOS usually consists of MOSFETs with analog models for the linear region, resistors, capacitors and a parasitic bipolar device. BiCMOS silicon contains high speed NPNs, lateral PNPs, possibly a high speed vertical PNP, resistors, capacitors, and MOSFETs. Currently, the transmitter and receiver electronics in modulation speeds below 600 Mb/s tend to be in silicon. Bipolar designs tend to be higher speed, up to the low gigahertz range, but also higher power. For example, one bipolar chip set for 1 Gb/s FCS, incorporating the laser modulator, serializer, receiver amplifier, clock recovery and deserializer dissipated over 4 W of heat. This thermal load can lead to difficult challenges on the package and environment to provide sufficient cooling mechanisms for the module to function and not impact the reliability of the transceiver. Digital CMOS circuits dissipate power only when switching, therefore lower power and possibly lower cost circuits can be achieved in analog CMOS. This technology is suitable for the amplifier, clock recovery, and both the serializer and deserializer circuits, although discrete bipolar devices would probably be needed in the modulation circuit. Recently, high performance BiCMOS technologies have been introduced that are suitable for modulation speeds in the low gigahertz range [7]. This technology allows the highest degree of integration, which should minimize cost and increase reliability.

The degree of integration that can be achieved in the optoelectronic transceiver circuits depends on the technology used, circuit design, cost, and receiver sensitivity. Integration leads to higher reliability and lower cost at the expense of flexibility for design changes. The laser drive circuit and serializer have been integrated. The encoder may or may not be included. This integrated circuit could still be used for other electronic applications, if necessary. On the receiver, the amplifier, clock recovery, and deserializer can be integrated, provided the power dissipation does not exceed the packaging

requirements. A decoder circuit can also be integrated, if necessary. Integration of the transmission and receiver circuits on the same chip has been achieved with sufficient noise immunity if proper isolation is used in the circuit design. In BiCMOS or analog CMOS technologies, a p^- substrate can be used to improve isolation between digital and analog circuits, as well as between the transmit and receive circuits [7]. Another technique is to use backside contacts for low resistivity p^+ substrates.

4.2.4 Receiver and Transmitter Electronics

The transceiver is composed of a transmitter and receiver, and each requires electronics to support its function. The receiver usually has an amplifier and the transmitter a modulator. These functions will be discussed briefly in this section.

4.2.4.1 Amplifiers

The optical detector generates a current pulse that needs to be amplified. High performance receivers may have a low noise pre-amplifier located close to or integrated with the detector. A second amplifier is often included for high sensitivity receivers. The first amplifier is most sensitive to noise. The noise amplitude in subsequent amplifiers is reduced from the signal amplitude by the gain factor of the first stage. Several different amplification schemes are used depending on the cost and bandwidth of the link. A low impedance amplifier is the simplest and least expensive. There is a trade-off in this amplifier between bandwidth and sensitivity. The low impedance allows good bandwidth, but sensitivity is limited by thermal noise [3]. Thermal noise is reduced in a high impedance amplifier, but this amplifier is limited in frequency response by the high load resistance and dynamic range. If both high bandwidth and high sensitivity are required, a transimpedance amplifier, although more complicated, is the best choice. In high sensitivity receivers, an equalizer may follow the amplifier to provide pulse shaping to invert the frequency response of the optical fiber. This approach may be important to reduce signal degradation caused by intersymbol interference, modulation, or amplifier distortions. Equalization may also be required to boost the high frequency components after a high impedance amplifier.

4.2.4.2 Modulators

Modulators will not be covered in this text.

4.2.5 Clock Recovery

The business of data communications could be most efficiently carried out if only the actual data of interest had to be transmitted. However, raw data has

several very poor transmission characteristics. One main drawback is that it does not contain timing information concerning the location of the separation between the bits. Another is that it does not have a determined set of frequency spectral characteristics to ease the design of the transmitters and receivers.

All data transmission requires some form of timing information. For serial data, there is only one channel of communication. Therefore the timing information must be combined with, and carried along with, the data over this one channel. This process is performed by scrambling or encoding the data discussed in the next section.

At the receiving end of the link, the timing information must be extracted. This process is termed *clock recovery.* The digital information is then synchronized to the recovered clock to remove any timing distortions it picked up during transmission and to allow for a clean transfer of the information to the remainder of the system. This process is termed *regeneration.*

In this section, we discuss only the transmission characteristics of the serial data and the process that lead to accurate recovery of the individual bits.

The usual data communication has several characteristics that ease the implementation of the data recovery process.

- The links are point-to-point. There are no optical splitters in the path. This configuration allows for continuous communications without the requirement of adapting rapidly to the signal strength and timing of multiple sources.
- Each end of the link has a separate independent timing source. There is no requirement that the clock be recovered in sufficient purity so that it is capable of acting as a timing source for transmission for another link. This approach greatly eases the design of the clock recovery and simplifies the link design by eliminating "jitter accumulation."
- The communication is usually symmetrical and uses nominally equal data rates in both directions. The clock from a transmitter may be used to aid the clock acquisition of the receiver.

The data is typically transmitted using some form of non-return to zero (NRZ) format. This formatting produces the minimum receiver bandwidth requirement for a given data rate. However, examination of the frequency spectral components of the signal shows that the shape is of the form $\sin(f/B)/(f/B)$, where B is the bit rate on the link. This frequency spectrum has no energy at the data rate. Therefore, some form of nonlinear processing is required to extract the timing information. These nonlinear processes fall into broad categories: resonate techniques and phase locked loops.

Resonant techniques use nonlinear signal processing, such as squaring, to produce frequency components at the data rate. The frequency components are then separated using narrow band filters. Most of these techniques require physically large components and do not lend themselves to an integrated

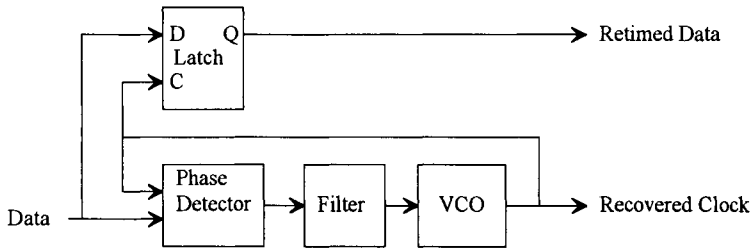


FIGURE 4.3 Block diagram of phase locked loop clock recovery and regeneration system.

solution. One exception to this are the surface acoustic wave (SAW) filter techniques. However, SAW filters, although very precise, tend to be expensive. Therefore, data communication systems tend to use phase-locked loop (PLL) techniques.

The block diagram of a phase locked loop clock recovery and regeneration system is shown in Figure 4.3. The incoming data is compared to the output of a voltage controlled oscillator (VCO) using a phase detector. This device puts out an error signal that is a function of the phase difference between the data and the local oscillator. This error signal is filtered and used to produce the control voltage for the VCO. The result is that one edge of the signal from the clock oscillator is aligned with the edges of the data. The other edge of the clock is located in the center of the bit period of the data. This edge is used to latch the data into the retiming latch forming the regeneration function. The dynamic response of the PLL is mainly controlled by the loop filter. It may be made as fast or slow as desired for a particular application.

The heart of the PLL, and quite often the trickiest part, is the phase detector. There are two broad categories of phase detectors: multipliers and sequential detectors. Multiplier-type phase detectors require frequency components at the data rate and employ techniques similar to those used by resonant clock recovery methods. These solutions tend to be pysically large and do not lend themselves to integration. Therefore, the preferred phase detector for data communication systems is the sequential detector.

Sequential detectors are constructed of logic blocks. They examine the sequence of logic transition edges between the data and the VCO clock to determine phase relationships. To use logic blocks, the data and clock signals must be digitized before processing. The sequential nature of this process requires the use of logic circuits that are very fast compared with the data rate. This approach may produce limitations in extremely fast systems.

A phase detector that produces a linear output signal proportional to the phase difference, and whose response is limited by logic block delays, is shown in Figure 4.4. The phase detector is a two-bit shift register that shifts the data on alternate clock phases. The latch inputs and outputs are connected to an exclusive or gate. This design produces an output pulse whenever the latch

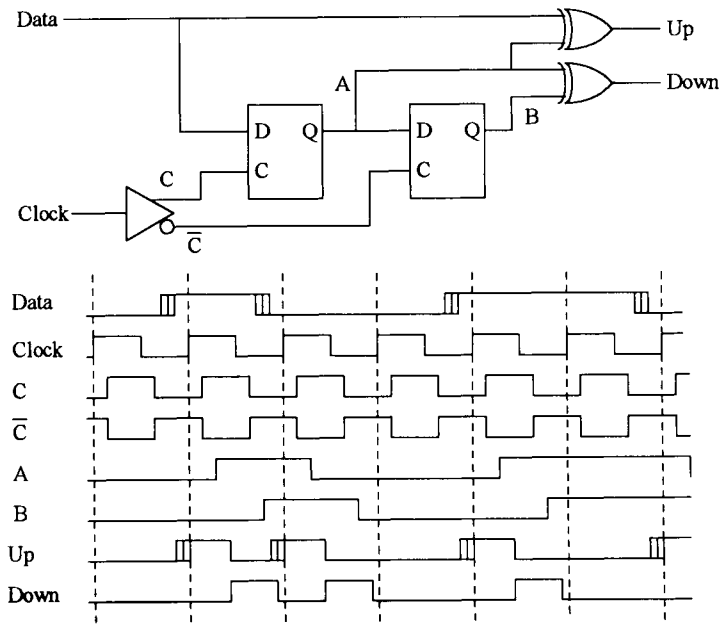


FIGURE 4.4 Phase detector.

changes state. The pulse width is proportional to the time differences between the data and clock signals at the latch input. The Up and Down signals are filtered and fed to the VCO through low pass filters to increase or decrease the frequency of the VCO. The timing diagrams in the figure have logic block delays included.

The data input is shown in three slightly different timings with respect to the ideal timing to the clock input. Note that signals A and B are retimed versions of the data with the input time variation removed by the process of clocking the data into the latch. The pulse produced at B always has the width of one-half of a clock pulse. However, the width of the pulse at A varies with the timing of the data input. If the data are late, the Up pulse is narrow. This situation results in the Down output being dominant in controlling the speed of the VCO, thus slowing down the clock input and bringing the data into alignment with the clock. Data phasing will stabilize at the center of the locations shown in the figure.

The important thing to note about this detector is that the stable data and clock phasing are not the ideal phasing for retiming the data. Ideally, a clock edge should be located in the center between the data edges. In this case, it is late by two logic block delays. This retardation must be compensated for by two block delays between the clock and the retiming latch. This approach may be difficult in high speed systems. Particularly in light of the fact that this

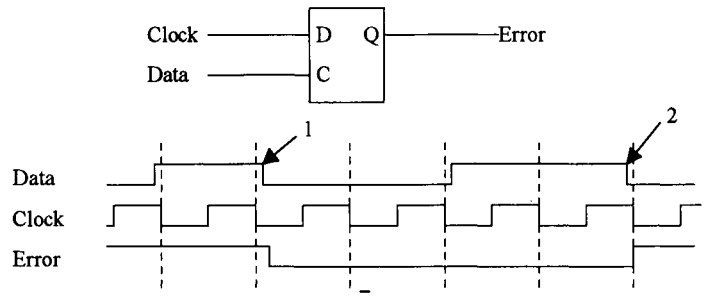


FIGURE 4.5 A "bang-bang" detector.

part of a clock recovery system is at the highest speed and will gate the technology, and hence the cost, of the entire function. There will be an economic drive to trim the design speed for this section as close as possible. An additional factor to be considered is that this type of detector is sensitive to the clock duty cycle. However, it does have a phase error output that is linear with the input phase difference.

Another commonly found detector is shown in Figure 4.5. It does not have a linear output. In fact, it is often called a "bang-bang" detector as the error signal is always fully high or low. This fact makes this type of detector difficult to analyze by conventional PLL analysis methods.

The detector uses a conventional latch. However, the clock is placed on the data input and the data is on the clock input. The data samples the clock in its negative edges. In the figure, edge 1 is late with respect to the clock and places a zero on the error output. On edge 2, the data are early and the clock places a one on the error output. A one would speed up the VCO and a zero would slow it down. Note that there is no stable state. The control is either in full speed or fully slowed down. However, if the loop filter is set up properly, excellent data clock alignment may be obtained.

A factor to be considered in the analysis of the PLL for clock recovery is the variability of the phase detector gain. Because these detectors sense data edge locations, the gain is a function of the transition density in the data. If there are periods of missing transitions, no correction signal will be generated. This shortcoming can be overcome to some extent by slowing down the control loop to "ride over" the period of missing error information. However, if the loop is required to be agile, this approach may not be acceptable. In this case, the use of a coding technique guaranteeing a high transition density is required.

False lock is a significant problem in data recovery systems, especially in those employing sequential detectors. The PLL can lock to harmonics of the signal near the data rate. This error is a particular problem because lock is often required to occur when the received signal is a short repetitive pattern,

such as an idle pattern. A 10-bit idle pattern will usually allow stable phase lock at 0.9, 1.0, and 1.1 times the data rate. Either the tuning range of the PLL must be constrained so as to not allow a wide range or some form of external frequency control must be used.

Many forms of VCOs are found in the industry. To some extent, the preferred form is influenced by the packaging available. Fully integrated oscillators tend to use either ring oscillators or multivibrators. These oscillators have a wide tuning range, allowing the PLL to be used over a number of data rates. However, the resulting high gain from control voltage to frequency makes them very sensitive to noises on the control line. If control loop filtering external to the chip or package is required, the filtering layout must be handled with great care.

There are two other design considerations for these types of oscillators. First, they require careful design to eliminate a tendency to high intrinsic phase noise. Second, their wide tuning range makes PLLs designed with them subject to false lock problems.

If package options allow for some form of resonate element, such as a tank circuit or shorted stub, a narrow tuning range LC oscillator may be developed. This method may be used to advantage when the chip is mounted on a fixed substrate such as a ceramic carrier or epoxy glass module. This approach allows the geometry control required to make a tightly controlled resonate circuit. These oscillators have the advantage of a narrow tuning range and low gain. These factors minimize false lock and noise coupling exposures.

Initial lock acquire may be a much greater problem with clock recovery systems than for PLL locking to constant frequency signals. The phase frequency detectors commonly employed in constant frequency systems will not work with data signals because of their sensitivity to missing transitions. The VCO frequency must be brought close enough to the data frequency to allow the loop to pull in directly. Ideally, the frequency difference should be less than the loop bandwidth. Fortunately, the precision of most data communication systems allows for wide loop bandwidths.

One common method of getting the VCO frequency near the data frequency makes use of a training frequency near the desired data rate. Because data communication systems are usually symmetrical, a signal at the nominal data rate is available for use by the transmitter. This signal may be used as a training frequency. A typical system is shown in Figure 4.6.

When the PLL is out of lock, the control directs the output of the phase frequency detector to control the output of the VCO. This type of detector will pull the loop into lock from any frequency difference. Once the loop is in lock with the training frequency, control of the loop is switched to the phase detector. At this time, the loop should be close to the data frequency. Ideally, within the bandwidth of the control loop. The loop will then quickly lock to the data. (General references for these topics are Refs. 8–10.)

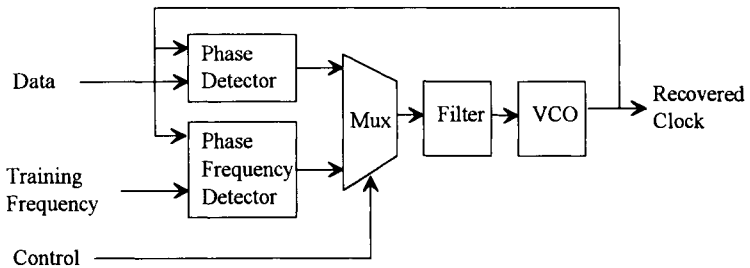


FIGURE 4.6 Training frequency generator.

4.2.6 Encode/Decode

For the maximum transmission efficiency, it would be best if only the information of interest were to be transmitted. However, real data usually has poor transmission characteristics. It seldom has good balance between the number of ones and zeros required for processing by AC-coupled circuits. The data transitions needed for clock synchronization are not guaranteed. These shortcomings are corrected by encoding the data into a more suitable format and decoding at the receiving end. For telecommunication systems scrambling is usually used. The byte-oriented content of data communications systems lends itself to various block coding formats with the 4B5B of FDDI (fiber-distributed data interface) and the 8B10B of Fiber Channel being the most popular. For an introduction to the mathematics of this interesting field, please see Reference 8.

4.3 Building the Link

4.3.1 Initial Sizing

The initial sizing of the data link is a critical task. As stated earlier, the critical issue that has delayed the proliferation of opto-data communication is cost. Hence, the link must be carefully sized to minimize cost. In the sizing, the first consideration is usually the length of the link. Many designers will assume that a long link length is needed, yet several analyses performed by IBM have shown that the great majority of data links are less than 125 m. It would therefore be very foolish indeed to assume that a link needed to have the performance characteristics for 15 km without some very serious analysis; the reason being that the technology to support this link would have to be SM and would be several times the cost of MM technology in the mid 1990s. If the entire network has a mix of distances, it may be advantageous also to mix the technologies to minimize total cost. A similar concern holds for data rate. Many people will want 1 Gbit/s because it sounds impressive, yet there are

few applications today that require this data rate. There is a rule of thumb that is helpful in determining which technology will be required. It is If data rate x distance > 500 Mbits-km/s, SM technology is required. Hence, if this metric < 500 Mbits–km/s, MM technology should probably be used to reduce cost.

One must also consider the usage environment. If the vast majority of the link is indoors, the installation will be simplified and the cost of the components will be low. On the other hand, if the installation is outdoors, hardened cable and hardened active components will be required. This situation will increase the difficulty of installation and increase the cost.

4.3.2 Optical Issues

There are several optical issues that need to be addressed when designing the optical link. These issues include:

1. Reliability of the link, lifetime of the laser
2. Laser safety requirements
3. Reflection-induced intensity noise (RIIN)
4. Self-pulsation or relaxation oscillation frequency
5. Modal or speckle noise
6. Receiver sensitivity and rise time as a function of wavelength.

The first three issues are related to the optical power in the link. Typically, long distance, high performance optical links have very high reliability targets. The failure rate of CD-type lasers is generally significantly higher than that of 1.3-μm semiconductor lasers because of the increased energy of the photons. As semiconductor lasers age, the current required to maintain constant output power increases [11]. Often, the failure mechanism for a laser in an optical link is a result of this increased current requirement exceeding the drive current capability. Accelerated life tests are used by many vendors to predict laser lifetimes. The lasers are operated at higher power and temperature while measuring the operating current required to maintain the elevated constant output power. An upper limit on the operating temperature is usually in the range of 75 to 80°C. Above these temperatures, the laser may not lase or the damage mechanisms may not accurately reflect those encountered during the normal operating life of the laser. The definition of laser failure is usually defined as the operating current required exceeding the original operating current by 30 to 50%. The laser lifetime, τ, may be approximated by the Arrhenius equation [11, 12]:

$$\tau \approx A_p P^{-n} \exp(E_a/kT) \tag{4.11}$$

where A_p is constant, P is the light output power density, $n > 0$, E_a is the activation energy, k is a Boltzmann's constant, and T is the temperature in degrees kelvin. Both n and E_a are determined empirically. Typical values for

n are $2 \leq n \leq 3$ for CD-type lasers. Another model for reliability relates lifetime to the current density of the laser [11, 12]:

$$\tau \propto \mathcal{J}^{-n} \tag{4.12}$$

Laser lifetimes can be improved by special laser cavity designs, process control during fabrication, and burn-in [13, 14]. As shown in the preceding equation, the operating lifetime is dependent on the intensity in the laser cavity; therefore, the lifetime increases if the laser is driven at lower powers.

A second concern that affects the optical power in the link is laser safety. Often, optical links are set up with class 1 laser safety requirements in order to allow an untrained operator to set up and maintain the link. This safety requirement puts an upper limit on the optical power exiting the TOSA.

Relative intensity noise (RIN) is the intrinsic noise profile of the laser that originates from the amplified spontaneous emission in the cavity. RIN can be defined in terms of the noise optical power (δP^2) and mean optical power as follows [12, 15, 16]:

$$\text{RIN} = \frac{(\delta P)^2)}{(P)^2} \tag{4.13}$$

From this equation it is obvious that the intensity noise is greater for an optical digit 1 than an optical digital 0.

Optical surfaces in the link produce reflections that can be coupled back into the laser cavity to dominate the noise profile. These reflections produce reflection-induced intensity noise (RIIN.) RIIN can degrade the S/N ratio of the link. This noise can be minimized by operating the laser at a greater mean optical power. There is a trade-off here as this high power operation might possibly violate safety or reliability requirements. Another means of decreasing RIIN is to minimize the intensity of back-reflections with antireflective coatings and physical contact connectors. The physical contact connectors, discussed in detail in Chapter 2, eliminate a surface in fiber-to-fiber coupling and, in addition, minimize reflections. The laser can also be optically isolated from the link with an absorptive layer with an antireflective coating. This approach would prevent violation of the safety requirements but may still degrade laser reliability.

The self-pulsation or relaxation frequency of the semiconductor laser is proportional to $\sqrt{P^6}$. The modulation frequency should be less than 2 to 3 times the self-pulsation or relaxation oscillation frequency to avoid excessive ringing on the falling edge of the pulse. Therefore, optical links with high modulation frequencies can improve the S/N ratio by increasing the average power out of the laser. This phenomenon becomes important when CD-type lasers are used for gigabit/second optical links. Currently, most commercially available CD-type lasers have self-pulsation or relaxation oscillation frequencies in the 1 to 2 gHz range [17].

Speckle or modal noise of a laser occurs as a result of variations in the intensity and distribution of the longitudinal modes of the laser diode [12, 13]. On optical surfaces within the coherence length of the laser, a speckle pattern is generated by interference between the longitudinal modes of the light. The coarseness of the speckle pattern depends on the number of different longitudinal modes lasing at any one time. These variations in the frequency distribution of the light occur without affecting the average optical power of the laser. Self-pulsating lasers with very short coherence lengths have many longitudinal modes lasing at once, minimizing the speckle and RIN. Modal noise in an optical link is produced by speckle patterns generated at optical interfaces and frequency-dependent coupling coefficients in MM optical connectors [12–14]. This noise may be minimized by using self-pulsating lasers to reduce the speckle pattern, by using physical contact connectors, and minimizing lateral offsets in connectors.

Finally, for high frequency modulation, the wavelength of the incident light can have a marked effect on the link performance by affecting the rise and fall time and sensitivity of the detector, as well as the bandwidth and attenuation of the fiber [18]. The responsivity of semiconductor PIN detectors increases with wavelength for frequencies above the bandgap. In silicon PIN detectors the active wavelengths are between 400 to 1100 nm. The optical absorption length also decreases as the wavelength increases within this region. For CD-type lasers, there is a marked difference in rise and fall times and responsivity as the wavelength is varied from 780 to 850 nm.

4.3.3 Electrical Issues

In the design of optoelectronic modules, careful consideration must be given to electrical issues as well as to the optical and mechanical issues. These issues are numerous and include:

1. Signal distribution between drivers and photoemitters and between photodetectors and amplifiers
2. Power distribution
3. Signal distribution from the terminal pins to the using circuits
4. Noise susceptibility to externally generated noise inducers
5. Cross talk in modules containing both transmitter and receiver functions.

The relative importance of these factors and the measures required to control them is strongly influenced by the module and circuit structure. It cannot be overemphasized that both the circuit and the module must be designed with consideration of the other needs and the environment. Too often, the circuit chip is designed then "thrown over the wall" to the module designer or package. This route is a sure one to disaster.

The first optoelectronic modules were simple transducers that converted an electrical logic signal to an optical output, or an optical data stream to electical logic levels. There are many of these still on the market and they provide a cost-effective solution to many applications. In spite of the relative simplicity of these structures, the electrical design may in fact be exceedingly complex. To a large extent, this complexity is due to the simplicity of the supporting module structure. There is a strong desire to use the simplest, lowest cost substrate and module packaging. When this approach is used, it must be done carefully or excessive noise and signal distortions may occur. In this type of structure it is often difficult to achieve proper power supply decoupling. On the other hand, the single transducer approach does offer the advantage of isolating the transmitter from the receiver.

In recent years, the use of duplex optical connectors has lead to a drive for duplex modules. In many cases, these modules are combined transducers. The most commercial modules for FDDI fall into this class. Care must be exercised to not allow the mutual support structure to couple noise from one side to the other.

In other cases, additional function has been included in the duplex module. These "full function" modules often include clock recover and data retiming, as well as the parallel-to-serial conversion on the transmit side and the serial-to-parallel conversion on the receive side. The additional circuitry, and its special requirements, creates additional challenges in signal distribution, noise control, and packaging.

There are four interrelated areas to be considered in the electrical design of an optoelectronic package. Three of these are the power distribution and the internal and external signal distributions. The fourth area is noise control. In many ways, noise control is the most difficult of the four areas. This challenge exists because in the other three areas, the designer is working to achieve a specific goal with the necessary tools and materials at hand; often the resulting design affects noise. Noise control crosses design disciplines. Hence, effective noise control requires good communication between all members of a design team. The factors that lead to good noise control will be mainly covered in the discussions on power and signal distribution. These areas are where the noise sensitivities and noise generation occur, along with the isolation and reduction of interferences.

The power supplied to the module and circuits is an often overlooked complication in the design. The first major choice is which power supply voltage to use. To increase the ease of use of the module in a system, the optoelectronic device should use the same power supplies as the logic in the system. This choice includes both the same tolerance as well as the same nominal values. However, although this approach creates a module that is easy to integrate into a system and that also enjoys a good marketing position, it exposes the module to the high noise levels that logic switching induces in the power supply.

The primary function of power distribution is to get the voltage supplied at the module leads to the using circuits within their voltage tolerances. For the DC distribution component, is usually not difficult—it is keeping the packaged circuits within their AC voltage requirements that is the major challenge. If the module shares power supplies with the logic in the using system, there is likely to be a considerable amount of noise on the input voltages. In fact, the power distribution is the primary means by which environmental noise gets into a module. Because it is generally the higher frequency components of the noises that cause problems, the usual method is to employ a low pass filter in the power leads. This noise suppression takes the form of power supply decoupling. RC low pass filters are seldom employed except in low current areas because of the impact of the resistance on the DC power distribution. One area where they may be used to advantage is in the power distribution to the PIN diode in a receiver. This area is one of the most noise susceptible in a design, but it also has low current requirements.

Usually LC filters are employed to reduce noise. However, it is seldom that discrete inductors are used. The inductance of power supply leads or sections of wire in the module may be combined with capacitors to form very effective filters. Often, additional filtering may be had by supplying the power leads to the module through a portion of board wiring. The trace inductance produces an effective filter when combined with a chip capacitor. However, at higher frequencies, the capacitor becomes inductive, which occasionally causes problems. The package inductance of a typical chip inductor is in the range of 1 nH. A typical 50-ohm trace in an epoxy glass printed circuit board has about 14 nH/in. of inductance. Combining these inductances produces an inductive voltage divider with 23 dB of rejection.

Cross talk becomes a major consideration when designing a duplex or full-function module. Common power distribution between the transmit and receive portions of a module can be a primary means of conducting signal between the two halves. It is usually good design practice to provide separate power feeds to the transmit and receive functions. This is easily done if the module is only duplex, which is only a transmitter and receiver connected to a duplex connector.

The ground connections within a module are a different story. The construction of the module often connects the grounds between the transmitter and receiver. This effect will often occur at a metal connector shell, if at no other point. Ground currents may flow from a disturbance within a module, such as the transmitter switching to all connected ground pins. The distribution of currents between the module leads is a complex interaction of the high frequency inductances of the package. A ground current from a switching transmitter may flow directly past a high sensitivity area of the receiver resulting in high noise coupling. These couplings are difficult to diagnose. The best plan in this situation is to provide a row of ground pins

between the transmitter and receiver to provide the lowest possible inductance path to ground. Then, the path that leads past the receiver will present a higher inductance, and hence will carry less current. Another advantage of numerous ground pins between the transmitter and receiver is that the net ground inductance is reduced, minimizing the voltage drop across the ground connection when there are changes in ground current.

Occasionally, advantage may be taken of splitting the ground between the transmitter and receiver. However, if there is any connection between the transmitter and receiver circuits this approach will probably result in a ground loop and increased noise. Connections of this type are almost a certainty in full-function modules. In this case, a single point ground should be used, a big one that is connected everywhere.

In the transmitter, one connection critical to the optical operation of the module is the connection between the driver and the laser or LED. Both lasers and LEDs offer challenging packaging constraints. In a transmitter module or in the transmitter function of a full-function module, one of the most important connections is between the driver chip and the laser or LED. Both lasers and LEDs are current control devices and careful control of the current waveform is important to the successful control of the optical waveform. Both lasers and LEDs are sensitive to package inductance. Lasers are usually used in high data rate systems and employ fast current rise times. On the other hand, the slower data rates and rise times of LED-based systems are offset by the much higher modulation currents required. In many cases, the dI/dt for the two are similar and the noise generation and inductance sensitivities of the two are similar.

The best practice is the most obvious; keep the connections short. The inductance of the modulation path is influenced by both the signal connection and the return path of the current. It does very little good to keep the connection between a driver and a laser cathode short and have the anode current travel through a module lead, through the supporting printed circuit card, and back through another module lead to complete the connection to the driver chip.

In addition to current waveshape control, inductance minimization through careful attention to the return path can reap benefits in noise reduction. Structures that have high inductance tend to be open and to form excellent transmitting antennas. It is therefore beneficial to close down the loops by routing return lines close to the outbound lines. Parallel connections are good and the use of ground planes adjacent to the signal lines is an even better approach to reduce noise if the module structure is of sufficient complexity to provide this option.

Lasers are often driven by current switch-type drivers. One needs to be sure that the current from the leg that does not drive the laser is treated with the same care as the current from the laser drive leg. This current can be a significant source of noise if not treated properly.

In the case of laser transmitters operating at high data rates, the transmission delay of the connection must often be considered. For gigabit data rates, current rise times in the 200 psec range are required. On a ceramic module, this is the round trip time of a 1-cm connection. Signal distortions resulting from transmission delays are common. In fact, advantage may be taken of the transmission lines in current waveshape control by the use of transmission line filters. However, even though these methods may allow very good current rise times, this benefit is usually at the expense of current lost in the terminators and matching networks. This effect requires a higher current driver, which may be difficult at the highest speeds. Great advantage can be gotten from the use of balanced signal design. For current switch drivers to the optical emitter, less noise is generated by routing the drive and return lines close together and terminating the return line close to the laser or LED. This approach has the effect of reducing the noise injected into the power supply.

In a receiver, the most important packaging consideration is the photodetector. This device is where the weakest signals occur followed by the highest gain. In data communication systems, the sensitivity requirements are usually modest and the receivers usually employ PIN diode detectors followed by transimpedance amplifiers. Amplifiers of this type are sensitive to the capacitance of the connections between the PIN and the input node of the amplifier. Capacitance at this point degrades receiver sensitivity, reduces bandwidth, and increases receiver noise.

From the preceding discussion, it would appear that the connections should be kept as short as possible and everything else as far away as possible. The short connection approach is always good. However, keeping other connections away from the input mode may cause other problems. The PIN diode acts as a current source whose magnitude is proportional to the signal. This signal current must flow in a closed loop. The other, usually AC ground, lead of the PIN is also an important circuit element. Although this lead has a low impedance to ground and is not capacitance sensitive, it forms a loop with the lead connected to the receiving amplifier. If the ground is far away, the connection could be opened to noise coupling. The effort to minimize capacitance is often at conflict with the desire for noise shielding.

In a high noise environment, it is desirable to use a differential connection between the PIN and the receiver. This connection is more complicated but has the advantage of very high noise immunity. However, the physical connection must be symmetrical along with the electrical design and schematic.

One often overlooked noise coupling in receiver modules is between the high sensitivity front end and the data output drivers. These signals are logic levels and usually generate a lot of noise. The use of differential ECL drivers is highly recommended here for all but the most modest data rates. Alternatively, if the receiver module is being designed in conjunction with the following module, an advantage may be gained with nonstandard

interconnection levels. Reductions in voltage swing may reap great rewards in overall noise generation; however, in this situation even greater care should be taken with the routing and lead dress of the interconnection lines to prevent them from receiving the ambient noise.

In the preceding discussions, we have considered only single transmitter or receiver modules. In duplex or full-function modules, the potential for interaction is greatly increased and additional care must be taken. The typically high noise photoemitter drivers are near to, and often sharing power and ground connections with, the highly sensitive receiver front ends. In full-function modules, the additional functions of clock synthesis and recovery, retiming, and mux/demux functions provide additional opportunities for noise trouble.

However, all is not lost. The majority of the design methods that are good practice for simplex modules are still appropriate for duplex modules. Usually, the only additional precautions that need to be addressed have to do with shared power and ground connections. It is the usual practice to provide separate power connections to the two sides of the module. In addition to eliminating cross talk through the common power connections, this approach provides the opportunity for external decoupling to reduce the possibility of the noise from the transmitter to the receiver taking a path through the external power connections.

Although the power connections may be separated between transmitter and receiver, there is often no choice with the ground connections. Often the package material is required to be metallic and to be connected to both the transmitter and receiver. This requirement may be for mechanical reasons associated with taking up the plugging forces associated with the duplex optical connector or for reasons of shielding. The need for shielding is critical and should not be overlooked. Often, a shield for external noises acts as a conduction path for noises between circuits enclosed within the shield. The noise coupling mechanism is usually induced ground currents. Recall that all currents into circuits induce image currents into surrounding conductors, then consider the routes these currents must take and note if any of these paths would be located to couple into the receiver. Often a circuit connection to the common ground or shield at a strategic point will shunt the currents harmlessly out of the way.

Full-function modules provide both advantages and disadvantages in electrical design. The most obvious potential disadvantages is all of the additional circuits, with their noise generation and susceptibility, near to the transmitter and receiver. However, this concern can be offset by potential reductions in noise from the interconnections.

The full data rate communications are usually confined to the module where the distances are short and I/O leads do not need to be traversed. This approach allows for the possibility of low noise interconnection and circuit designs. Voltage swings may be greatly reduced. The distances are short

enough that transmission line effects are usually not significant. This condition allows the use of unterminated nets or high termination impedances. If differential communications are used and the lines are routed close together, the effective noise voltage may be nearly eliminated. Another method that may be used to advantage is current mode communications. A low value current source may be switched between a pair of differential lines communicating with a low impedance receiver on the destination chip. An arrangement of this type has nearly no voltage swing and will not couple noise into voltage-sensitive circuits. Both of the preceding methods may be used depending on whether the surrounding circuits are primarily sensitive to voltage or current noises.

The frequency requirement of the transmitter and receiver are similar. That is, the transmitter tends to produce noises in the most sensitive frequency band of the receiver. However, with a full-function module, the noise of the off module signal communications is in a lower frequency band because of the mux/demux action. This allows the signal transitions to be slowed down to reduce noise generation. However, simultaneous switching of the drivers can eliminate this advantage. It is highly recommended that the internal circuit design be structured to ripple the output drivers to minimize simultaneous switching.

There are two other factors associated with the parallel output busses. First, the large number of I/O pins required for the parallel interfaces precludes the use of different input and output signals. This removes one of the options for noise control. Second, these slower data rate signals often communicate with relatively slow-speed external logic, which often communicates using high voltage TTL or CMOS signal swings. This removes the other major option of noise control.

A special consideration concerning full-function modules is the clock synthesis of the transmit side and the PLLs used in the receive clock recovery. These PLLs are typically very sensitive to near frequency noises. As the transmit and receive portions of the modules are running at the same nominal data rates, this can be a significant problem. Although a portion of this problem may be handled by careful attention to packaging, which should always be done, this problem is best addressed at the circuit and chip level. One needs to be careful of the power feeds to PLLs and in particular to the phase detectors and the VCO. One technique that has been successfully used at gigabit data rates is to operate the VCO from a regulated power supply developed on chip from a bandgap reference. The resulting series regulator had greater than 60 dB of rejection at the data frequency. This not only prevented the clock recovery from becoming an injection locked oscillator at the wrong frequency, but also provided enough rejection at the circuit level that the design of the next level package was greatly eased.

In general, the power and ground distribution considerations of a full-function module are similar to those of a simple duplex module. The only

significant additional consideration is the power distribution to the off chip drive circuits and the containment of their noises, which has been discussed previously.

The external signal distribution has conflicting requirements. First, it must be compatible with the external support circuits. This limits the signal swings to one of several commercial logic families. The noise generation characteristics of these signals are often not compatible with sensitive receivers. In general, the circuit and package designer has some control over the output signals but not the input. There is a desire on the part of the receiver designer to slow down the off chip drivers to reduce noise generation. However, the drivers cannot be slowed down too much as this may result in jitter and timing problems at the system level. For slow speed communications, most logic circuits families employ high voltage swing, unterminated signals such as TTL or rail-to-rail CMOS. These logic families need careful handling. However, they are usually encountered at the slower data rates where traditional module noise filtering may be used.

An exception to the preceding situation is encountered in full-function modules in which a high data rate may be combined with relatively slow off chip drive in the parallel interfaces. Two design practices may be conveniently used in this situation. First, the noise of the off module communications, which is associated with the data content, is in a frequency band that is the data rate divided by the parallel interface width. The higher frequency noises are mostly edge rate dependent and may be minimized by slowing down the drivers as far as is consistent with the timing and jitter requirements of the interface. The second practice is to stagger the switching times of the drivers to minimize simultaneous switching.

Most high speed serial communications into and out of data communications modules is done using ECL levels. These connections are usually differential, terminated connections. Of the presently commonly available options this is by far the best choice. A consideration which must be kept in mind is that any signal distortions at these interfaces directly impacts the quality of the communications. Of particular importance is the connection between the receiver and the circuits that perform the clock recovery.

These considerations make the termination of these connections critical. On the transmitter side, the ideal location for the terminator would be in the transmitter module itself. However, this is not often done for heat dissipation reasons except at the very highest data rates. These data rates are not often encountered in data communications. The terminators are usually located on the printed circuit card mounting the module. The connections to the module should be run to the module, connect to the input pin, and continue to the terminator. This will ensure the minimum reflection distortion of the signal. The common practice of running the connection to the terminator and then continuining to the input pin produces a capacitive stub on the transmission line that may cause significant signal distortion.

4.3.4 Mechanical Issues

There are numerous mechanical issues related to link performance. The first consideration is ensuring that the connector will plug into the transceiver module. This concern is discussed in detail in Chapter 5. The importance of this issue cannot be overstressed, as intermateability, especially between different component vendors, has been shown to be a significant problem in early work at IBM.

To have a reliable link, one needs to ensure that the individual components and the configured link will withstand mechanical shock, vibration, thermal cycling, and hostile environments (dust, relative humidity, etc.). The use condition will define how extreme the testing needs to be. A military application will require much more stringent specifications than an office environment. It is critical to not overspecify a requirement, as the cost can increase dramatically with more stringent specifications. Typical specifications have been produced by the United States military and Bellcore.

4.3.5 Thermal Issues

Thermal issues can be critical in the design and cost of a transceiver module. This is true for several reasons. SM lasers typically need to be operated at temperatures below 60°C to have a life long enough (100 000 h) for reliability specifications. When one considers that the ambient temperature inside a computer or a peripheral is approximately 40°C and that the drive circuitry inside the transceiver module generates up to 6 W of heat, it is a considerable challenge to keep the temperature of the laser below 60°C. In some applications, where cost is less important, thermal-electric coolers are used. In typical opto-data communications, this cost is prohibitive and passive cooling techniques such as cooling fins and thermal grease technologies are used.[1]

4.4 Industry Examples of Packaging Solutions

Over the last few years the optoelectronics industry has progressed to the point where there are a wide variety of commercial modules available. In this section, we discuss a number of these as they apply to general packaging solutions. Although many designs will not require an industry packaging solution, a study of the methods used in these designs may aid in the design of a specific application.

[1] It is beyond the scope of this text to give a detailed discussion on this topic. The following references are given for those interested in further study: *Principles of Electronic Packaging*, Seraphim *et al.*, McGraw-Hill, 1989, Chapter 5 and *Microelectronics Packaging Handbook*, Tummala *et al.*, Van Norstrand Reinhold, 1989, Chapter 4.

The electronic packaging materials that are used are many and varied. For the commercial environment, the predominant encapsulation material is plastic. This material may enclose a circuit support structure of epoxy glass or ceramic depending on the requirement of the circuits. In some cases, usually concerned with more extreme environmental conditions, higher reliability, or longer life, metal-ceramic packages are employed.

The most common modules on the market are the simplex transducers. These modules perform the conversion from logic level electrical signals to modulated optical data streams and back to logic level signals on the receiving end. The electrical interface levels are usually ECL logic for minimum noise coupling. In many cases, additional features may be included, such as current monitors on laser-based transmitters and loss of signal monitors on receivers. These modules are available from many vendors and may be obtained with mating connectors covering the range from ST to FC to SC, as discussed in Chapter 2. Most of these modules are of similar construction based on a variation of the dual in-line pin (DIP) logic module layout.

There are also a number of duplex transducers on the market, usually referred to as transceiver modules. The design for these devices is usually driven by the requirements of a standard calling out a duplex connector. The alignment requirements of these connectors are usually such that it is not practical to align separate transmitters and receivers to the required tolerances, hence a single transceiver module is developed. A first example of this type of module is the IBM ESCON module marketed by Siemens and AT&T. In this module, two simplex modules have been joined by the connector housing to make a duplex module. Additional manufacturers have since combined these two functions into one molded part. As this module is made from two separate simplex modules, the resulting pin-outs are in the form of two DIP packages.

There are a number of FDDI duplex modules on the market. These structures typically use a common housing for both the transmitter and receiver and are designed to accept the FDDI MIC (media interface connector). The AMP 124002 module shown in Figure 4.7 is an example of a module of this type. These modules show their evolution from simplex modules in their shape and DIP configuration.

A variation on the FDDI module is the Sumitomo SMD3101-XF, which was originally developed by Sumitomo as a packaging alternative to reduce molding cost. The package was the first major variation on the dual in-line package and has the pins in line across the rear of the module. It can be packaged with one molding operation. This package is presently being sourced by several other suppliers.

Fiber Channel and the Low Cost FDDI initiative are migrating to the duplex SC connector for reduced cost. A version of the Sumitomo FDDI for this application (SDM4153-XC) is shown in Figure 4.8. This low-cost module form factor is currently finding much favor in the industry and is

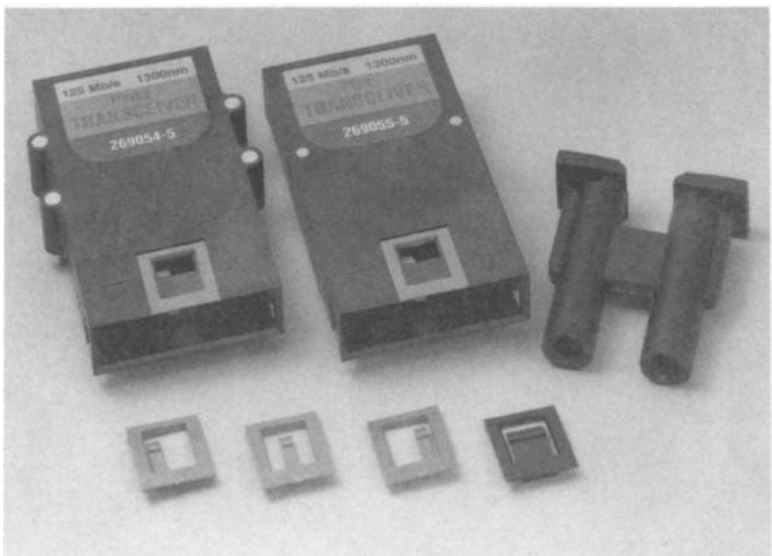


FIGURE 4.7 AMP FDDI transceiver.

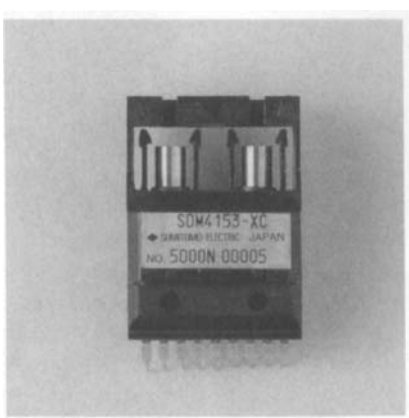


FIGURE 4.8 Sumitomo SDM4153 FCS module with SC duplex interface.

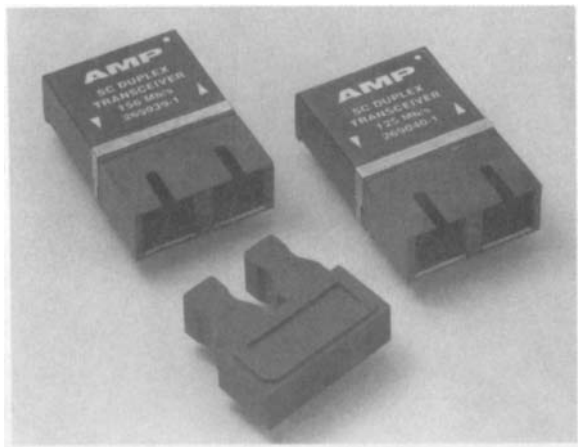


FIGURE 4.9 AMP ATM (left) and FDDI modules with SC duplex interface.

a number of data rates. Two versions of this module from AMP for 125 Mbit/s and 156 Mbit/s are shown in Figure 4.9. Industry agreements have lead to a number of suppliers using common intermateability designs. This approach permits module interchangeability and allows for multiple sourcing of components.

Industry agreements for common packaging designs and interchangeable components has accelerated the acceptance of commercial modules, rather than custom designs, in the user community. These agreements have in turn decreased cost through the economy of scale of volume manufacturing. This acknowledgment of the industry advantages of cooperation is being carried forward with recent agreements concerning common packages for SONET OC-3 data rate modules using 1300-nm laser-based optics. The package is an elongated version of the 9-pin SC duplex module.

Some modules have been designed to variations of standards to meet specific market needs. An example of this is the Hewlett-Packard HFBR-5104/5203 variation of the FDDI standard. This module communicates on low-cost, 850-nm LED optics, but is otherwise compatible with the FDDI standard. In some applications this may reduce system cost in which compatibility with modules using the 1300-nm FDDI standard is not required.

A large number of proprietary optical links exists in equipment on the market. These links tend to be designed for specific applications and usually employ some of the wide variety of commercial optical devices and support chips packaged on the circuit board of the using system. However, in some cases, portions of these boards have been lifted and developed into separate products.

An example of this approach is in the extensive use of the Advanced Micro Devices TAXI chip set as the system interface to simplex transducers. This chip set has led AMD and BT&D to combine the TAXI chip with 850-nm LED optics in a plastic molded package, resulting in the DLT6000-ST transmitter and the DLR6000-ST receiver. Although these modules are not compatible with any standards, they may provide a clean solution to a full-function interface in many applications.

There are only a few full-function duplex modules on the market at this time. One design is marketed by Hewlett-Packard (HOLC-0266) and IBM for the 265.625 Mbit Fiber Channel Standard. This module is a pluggable printed circuit card with 10-bit parallel TTL data interface for both input and output. It is constructed with surface mount components on epoxy/glass circuit board material. The transmit side of the module contains a transmit clock synthesis, parallel to serial conversion, a 780-nm laser along with a laser driver and laser power control circuits. The receiver side contains a silicon PIN diode receiver, loss of light circuits, clock recovery, retiming, serial-to-parallel conversion, and byte alignment. The wavelength of light, in this application, is 780 nm using a CD-type laser and in a duplex SC connector assembly. It is intended for operation with 50/125 μm fiber. Also included is a safety system called Open Fiber Control which automatically shuts off the laser in case there is a break in the fiber which might allow laser light to escape. A version of this card addressing the 531.25-Mbit/s Fiber Channel data rate is also available from IBM.

The 1.0625-Mbit/s data rate of the high speed version of the Fiber Channel Standard is presently the highest data rate full duplex standard. The Siemens V23806-A5-C1 module addressed this market in a single full-function module. The module employs a 1300-nm laser operating with a 9/125-mm fiber. It operates with a 20-bit parallel interface and includes all required parallel to serial and back conversion functions. Also included is an Open Fiber Control System. The packaging is in a peripheral leaded metal ceramic package with additional strain reliefs to take up the strain of plugging the duplex SC connector. Siemens also markets a version of this function packaged to be interchangeable with the IBM/Hewlet-Packard optical link cards as the V2306-6120-X1.

4.5 Future Trends

Fiber-optic communications is an area of technology that is of crucial importance to U.S. high-technology industries. The 1990 Council on Competitiveness identified fiber optics as one of the few key technological areas in which U.S. industry is still competitive in the world marketplace [19].

At present, the telecommunications industry is the largest user and developer of fiber-optic products. However, the data communications industry is

beginning to convert to fiber, and it is in data communications that the most growth potential is expected over the next 20 years.

As discussed before, data communications applications have very different requirements from telecommunications applications. The most striking differences are that data communication requires many more transceiver modules and duplex connectors per user, whereas telecommunications has much greater lengths of cable. These differences place a tremendous cost pressure for optical data communications (opto datacom) on modules and connectors, whereas optical telecommunications (opto telecom) is concerned more with the cost of cable [20]. The opto telecom industry responded to the long cable length requirement by dramatically improving cable price/performance. However, real breakthroughs in low-cost design and manufacturing technologies for opto datacom transceiver modules and connectors are still needed for significant proliferation of this technology to occur. These significant cost reductions, to less than $150 per Gbit/s transceiver modules, are needed for future opto datacom technologies to reach their full potential. A transceiver module such as this will be an enabling technology for rapid growth of the electronic super highway, ATM (asynchronous transfer mode), FCS (Fiber Channel Standard), and fiber to the home.

The optoelectronic telecommunications industry developed the fundamental technologies to facilitate the emergence of the optoelectronic data communications industry. There are, however, considerably different user requirements in these two industries that have driven opto datacom product developers to expand on the telecommunication-based technologies. One of the most significant differences is that opto datacom is transceiver module/connector intensive, whereas opto telecom is cable length intensive [20]. Currently, opto datacom is emerging more slowly than expected because of the resistance for transceiver modules and connectors. Recently, the cost of opto datacom cables with duplex connectors has dropped to below $100 with the introduction of the SC duplex connector [21]. It is expected that current efforts by cable vendors will continue this cost reduction to the $50 range, especially with the recent introduction of plastic ferrules. The transceiver module is another story. Projected costs are much too high to allow proliferation of this technology. Discussions that the authors have had with several optoelectronic suppliers [22] indicate that a price of $100 to $150 for a Gbit/s transceiver module is needed for wide application of opto datacom and eventually fiber to the home. Currently, these modules are in the $600 range and economies of scale can only reduce this price range to $300 to $400. This $150 price objective is unobtainable with today's technologies, primarily because of packaging. Figure 4.1 depicts this situation. The laser or detector die is first actively aligned in the TO packaging. The TO package is actively aligned in the optical subassembly (OSA) and the OSAs are finally packaged again in the transceiver module package. Many of the piece parts are also expensively machined stainless steel or ceramic. A concerted effort to reduce

the levels of packaging and to develop significantly less expensive opto packaging designs, materials, and processes is needed for opto datacom to proliferate widely. It is especially important to eliminate the active alignments, as they are a major part of the assembly cost.

As previously mentioned, the technological elements of a data link are the transceiver module, the jumper cable and connector, and the part of the circuit board that the transceiver is mounted on. Between 40 and 80% of the cost of the data link is in the transceiver module [20]. Circuit board technology at the required level of complexity is a mature technology and one would not expect to see significant cost reductions in this area. Similarly with the increasing acceptance of NTT's SC connector, in both simplex and duplex versions, we see this excellent connector as a price/performance leader and see little reason to try to improve on it. The transceiver module is another story (see Figure 4.1). The module consists of a transmitter optical subassembly (TOSA) for launching light into a fiber, a receiver optical subassembly (ROSA) for receiving the light from another module, transmitter (TIC) and receiver (RIC) drive circuitry, serialize/deserialize (SERDES) and bias/safety control (BSC) circuitry, a mounting substrate, and electrostatic discharge (ESD) shielding. As mentioned in the introduction to this section, almost every aspect of the transceiver module design has considerable potential for dramatic cost reduction, although the TOSA and ROSA, representing 50% or more of the module cost [20], should receive the most attention. It is important to mention that it is not expected that cost will be significantly reduced with increasing production volumes with current designs; new design concepts and processes are needed.

A proposed very low cost transceiver design concept is shown in Figure 4.10. A molded completely plastic optical platform (POP) is the heart of the design. It houses the light source or detector, lens, drive or preamp, and the monitor diode (MD) circuitry. The POP must go beyond the state of the art in plastic molding tolerances, achieving 1-μm tolerances over several

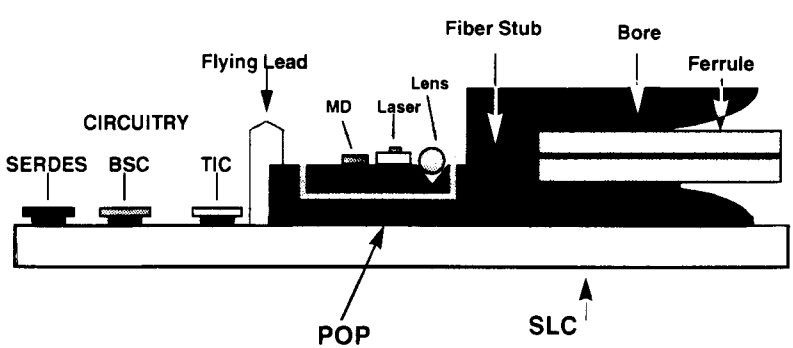


FIGURE 4.10 Proposed very low cost opto-electronic transceiver module.

millimeter distances. To achieve this precision, nascent technologies such as LIGA (lithography/galvanometry) [23] will be needed to produce the plastic injection mold for the POP and to perform materials investigations on the polymer material to minimize shrinkage and warping. The polymer should have a high modulus to resist wear and stresses from the ferrules in the duplex connector. This is especially a concern because ferrules are often made of hard zirconia ceramic. Liquid crystal polymers (LCPs) may be a candidate material for the POP as they have a modulus of about 2 Mpsi (13.8 GPa). However, the crystalline nature of LCPs could be a problem in precision molding, therefore other high modulus polymer materials should also be considered. Because long-term creep at low stresses from the connector ferrules will be a concern, the POP material must also not creep more than 1 μm at stresses of 50 psi (0.3 MPa) over the use life at a temperature extreme of 50°C. It may not be possible or beneficial from a tolerance perspective to have the bore that receives the ferrule from the mating connector share the mold with the POP. Therefore it is likely that the bore would be molded separately in a second mold that is also made with a LIGA process or by precision machining. The joining of the POP and the bore will pose a significant challenge from a tolerance perspective. To avoid an active alignment of the POP and the bore, they will either be joined by butting to a precise feature or by using the "C4" solder alignment techniques [24]. The POP will also have some circuitization on it, requiring an effort to ensure circuitization adhesion to the POP material. The coupling of light into the optical receiving fiber in the ferrule may be guided by a fiber stub. A tolerance/coupling analysis has shown that a total worst-case error of 4 to 5 μm in the laser spot on the fiber stub will give acceptable performance for the link in SM technology. The mechanism of choice for this optical alignment is to butt the active devices and lenses to precise features on the POP, although there are other passive alignment schemes using C4 solder ball alignment that may work. Any of these approaches requires precise knowledge of the location of the active devices with respect to a die mechanical feature. Because this information is not required in current active alignment approaches to better than 25 μm, the dicing, or cutting, of the device from the wafer is currently performed to this type of tolerance. New work needs to be done to improve the precision of this dicing. One approach will be to photolithographically create precise etched lines for dicing. This approach should be capable of better than 1 μm precision. The circuitized substrate on which the transceiver may be built would need to be a low-cost, organic substrate with fine circuitry features, such as IBM's SLC technology. This card will take surface mount, plated through hole, or direct chip attach components. The POP and bore would be mechanically mounted onto the SLC. The SLC will also carry the TIC, RIC, SERDES, and other circuitry. This design may present some difficulty in cooling the module as the circuitry generates 6 W of heat and the laser 0.1 W. An effective cooling approach is to use standard heat sinks with thermal grease

or radial finger contacts to the devices. The functional specifications of an SM transceiver like this would need to be BER 10^{-15}, distance 3 km, reliability 10^6 h mean time to fail (MTTF), class I laser certification, and a 20-bit parallel TTL/CMOS interface.

The two most critical cost reduction elements of the proposed approach are the passive alignment of the optical elements and the use of an ultra-precision molded plastic unibody. This method is in contrast to the present use of active alignment and expensive, machined stainless steel and ceramic piece parts. The passive alignment also requires precision dicing of the laser and receiver dies. Dicing precision would be a major technical leap in that it would be improving on current methods by a factor of 25. This low-cost approach also reduces the levels of packaging from three (TO can, OSAs, module) to two (POP, module). The hermetic sealing used today is expensive and may not be needed. Therefore, stress and humidity testing of unprotected devices could be performed to determine whether hermiticity is an absolute requirement or whether low-cost sealants could be used.

The expected technical hurdles in this approach track closely to the most critical cost reduction elements. Achieving micron tolerances in plastic will be a considerable challenge. This challenge can only be met by improving the mold-making technology and optimizing material properties. There is great hope for LIGA to achieve the required mold tolerances, but material optimization studies are needed to minimize polymer shrinkage and anisotropy enough to achieve micron level tolerances. Micron level dicing accuracy is most likely the other great hurdle. To achieve passive alignment, the dicing of the laser and detector dies must be precise to within a fraction of a micron. By using photolithographically defined etch lines, it is possible to achieve such precision. However, it must be understood that the precision objectives for the POP and die dicing and the resulting passive alignments would have considerable risk and take the state of the art beyond what some people believe is possible. The plan to eliminate hermetic sealing for the semiconductor lasers and devices would carry risk also. We would expect some difficulties here as our use environments cover a broad range and some are quite hostile. Although this plan cannot claim that all aspects of this design are workable, design concepts along these lines are needed to achieve the low costs required to allow opto datacom to proliferate to the consumer level.

Other low-cost approaches may compete with or replace these ideas, but any concepts that are successful must reduce levels of packaging or eliminate active alignments.

References

1. EIA/TIA-492BAAA. *Detailed Specifications for Dispersion Unshifted Class IVa Single-Mode Optical Fiber.*

2. *Fiber Channel Physical and Signaling Interface* (FC-PH). Working draft proposed by American National Standard for Information Systems, Rev. 4.1, September 17, 1993.
3. J. M. Senior (1985). *Optical Fiber Communications Principles and Practice.* Englewood Cliffs, NJ: Prentice-Hall International.
4. T. Ohtsuka, N. Fujimoto, K. Yamaguchi, A. Taniguchi, H. Naitou, and Y. Nabeshima (October 1987). Gigabit single-mode fiber transmission using 1.3 μm edge-emitting LEDs for broad-band subscriber loops. *Journal of Lightwave Technology,* **LT-5**(10), 1534–1541.
5. R. L. Soderstrom, T. R. Block, D. L. Karst, and T. Lu (1990). Optical components and electronic packaging for high performance optical data links, *Proceedings of the First International Workshop on Photonic Networks, Components and Applications.* Montebello, Quebec, Canada, October 11–13, pp. 19–28.
6. *Fiber Channel Physical and Signaling Interface* (FC-PH). Working draft proposed by American National Standard for Information Systems, August 12, 1993.
7. J. Gross, H. Chuang, and T. Schmerbec (1993). *State-of-the-Art Analog BiCMOS.* IEEE GaAs IC Symposium, San Jose, CA, October 10–13.
8. Y. Takasak (1991). *Digital Transmission Design and Jitter Analysis.* Boston: Artech House.
9. F. Gardner (1979). *Phaselock Techniques.* 2nd ed. New York: John Wiley & Sons.
10. J. Crawford (1994). *Frequency Synthesizer Design Handbook.* Boston: Artech House.
11. M. Fukuda (1991). *Reliability and Degradation of Semiconductor Lasers and LED.* Boston: Artech House.
12. G. Keiser (1983). *Optical Fiber Communications.* New York: McGraw-Hill.
13. R. Soderstrom, T. Block, D. Karst, and T. Lu (1991). An optical data link using a CD laser. *SPIE Proceedings, Vol. 1577. High-Speed Fiber Networks and Channels.* Boston, September 1991, pp. 163–173.
14. T. Lu (1991). Reliability of a self-pulsating CD laser for fiber-optic data links. *Proceedings SPIE, Vol. 1620. Laser Testing and Reliability.* San Jose, CA, November 1991, pp. 49–59.
15. K. Petermann (1991). *Laser Diode Modulation and Noise.* Boston: Kluwer Academic Publishers.
16. G. P. Agrawal, and N. K. Dutta (1986). *Long-Wavelength Semiconductor Lasers.* New York: Van Nostrand Reinhold.
17. E. Foster, K. C. Jen, C. Paddock, J. Radcliffe, and W. Vetter (1993). High performance packaging of gigabit data communication optical modules. *Proceedings of the SPIE Conference on Processing and Packaging of Semiconductor Lasers and Optoelectronic Devices.* Los Angeles, CA, January 20–21, 1993, Vol. 1851, pp. 54–63.
18. M. J. Hackert, S. M. Sibley, and P. Sendelbach (1992). Performance of telecommunication grade multimode fiber at 780 nm. *Journal of Lightwave Technology* **10**(6), 712–719.
19. U.S. Council on Competitiveness, 4/91.
20. R. K. Warren (1991). *Proceedings of the 41st Electronic Components Technology Conference.* Atlanta, GA, May 11–16, 1991, pp. 161–168, IEEE Pubs.
21. IBM FCS Duplex Connector Literature. Optical Link Products, IBM, Endicott, NY.
22. Discussion among Bell Northern Research, GTE, and IBM. Fall 1992.
23. H. Guckel (1992). *J. Micromech. Microeng.* **2**, 225–228.
24. J. McGroarty, P. Borgesen, B. Yost, and C.-Y. Li (1993). Statistics of solder alignment for optoelectronic components. *IEEE-CHMT* **16**, 527–529.

Chapter 5

Connector/ Module Interface

Darrin P. Clement
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire

Ronald C. Lasky
Universal Instruments Corporation
Binghamton, New York

Lawrence Brehm
IBM Corporation
Microelectronics Division
Endicott, New York

5.1 Introduction

5.1.1 Foreword

The previous four chapters provided an insightful look at many of the components that make up an optical data communication system. Although a basic understanding of each of these components is necessary for the engineer, it is also important to have an integrated knowledge of how these components must be assembled to achieve a statistically reliable system. Much of the communication industry is focused on exceeding "fiber to the curb" capabilities to systems in which the user will have direct access to a fiber-optic line ("fiber to the home/office"). Before these optical local area networks (LANs) can be implemented, however, manufacturers must devise user-friendly systems that provide robust performance. This chapter will attempt to

Optoelectronics for Data Communication

explain the challenges of mating connectorized fiber cables with optical subassemblies (OSAs). Some terminology will be borrowed from the OSA telecommunication industry. However, data communication requires some unique engineering solutions and therefore some unfamiliar terminology will be introduced. A review of Chapter 2 would be helpful.

In addition to existing markets for LANs, the future of fiber-optic signal transmission may very well include many applications in libraries, schools, and private homes. As the data revolution evolves, connector-to-module interfacing will be critical in the development of video conferencing, interactive learning, and interactive television. These designs will require numerous pluggings, unpluggings, and repluggings with no significant mechanical or optical degradation. Although this book has been written as a text on data communication, it is clear that a fundamental understanding of this material could also aid any voice/data/video communication engineer.

The total available bandwidth of optical fiber is approximately 25,000 GHz in both the 1.3-μm and 1.55-μm regions [11]. The telecommunication industry has wisely chosen to make use of this huge bandwidth by developing a system in which many people use the same line. The available bandwidth is divided into small sections of 64 kHz for each user. (64 kHz represents 16 times oversampling of the inherent 4-kHz bandwidth of the human voice [21].) Although this telephony technology has many endusers, they are usually connected electrically after an optoelectronic signal conversion. Hence, there are few interconnections in the *optical* link. However, many *new* technologies (and most likely many that have not yet been conceived) will involve more optical users per link, each requiring more bandwidth when compared with a simple telephone call. Accomodating more users per link requires more sources and receivers per link. This means that more effort should be spent in reducing the costs of optical sub assembly modules while improving performance. Because of the increase in both number of and exposure to users, the transmitter and receiver will play a large role in data communication. Although many different types of connectors and optical modules are currently in use, the fundamental issues regarding system performance and safety remain consistent. The connector-to-module interface must be durable, small in size, and environmentally secure while providing simple connection (successful first attempt). Additionally, because of user accessibility, the links must be safe. These considerations place both optical and mechanical constraints on design. Each connection must be mechanically secure so that inadvertent unplugging does not occur. Furthermore, if a user fails to take the necessary precautions when plugging and unplugging (e.g., by turning off the optical power), the module design itself must ensure user safety by either blocking the light with an opaque barrier or by killing the signal (see Ch. 2). It is also possible that an optical cable may become severed somewhere along its length, therby exposing the user to radiation. Because of this complication,

international safety standards have arisen that require the optical power at *any* point in the line to be less than 0.600 mW [14]. In the United States, the suggested limit is 1 sec of exposure at 1 mW of optical power [1]. The use of lasers with multimode fiber may compound the safety problem because of the relative ease with which, compared with single mode, the radiation can be coupled into the fiber.

Recent developments have provided a different perspective on safety concerns. The alignment of source to fiber need not be controlled so as to keep the launched power below a certain value. Automatic power control is permissible in IEC standards. A thorough understanding of the relationship between all optomechanical parameters and coupling efficiency is necessary in order to design hardware that is durable and that will hold power to within limits compatible with the dynamic range of even an automatic power control. Equipment must be able to withstand stresses from handling by operators, thermal changes, mechanical vibrations, and even clumsy users. An understanding of all of the optomechanical elements of coupling efficiency is important in this regard.

Generally, power in the fiber needs to be held to roughly 1 mW to receive a class 1 safety rating. Such a class rating allows universal handling by persons with no special training, which would encompass the typical user of a workstation or personal computer. The exact level of power deemed to be safe depends on the wavelength. The human cornea is essentially opaque to the 1500 nm radiation used in long-distance telecommunication links. This corneal opacity protects the retina, but the cornea, that is, the outermost surface of the eyeball, must be protected from the beam. At the shorter wavelength windows of 800 and 1300 nm, the cornea can focus the radiation onto the retina. Therefore, this additional biologic optical interface must be considered in anticipating the possible mechanisms for ocular damage at the shorter wavelengths [31].

In the United States, laser safety limits are written into law by Congress and administered through the Center for Devices and Radiological Health (CDRH) section of the Food and Drug Administration [30]. In Europe, the International Electrotechnical Commission (IEC) has power to set legally enforceable limits and practices. The American National Standards Institute (ANSI) sets rules that function as standards for appropriate and safe professional practice. ANSI also has the authority to represent the IEC standards to companies in the United States, for example, to those intending to manufacture hardware for sale and use overseas.

5.1.2 Contrasting Single-Mode and Multimode Fiber

We should at this point contrast the cases of single-mode (SM) and multimode (MM) fiber. In doing so, we will necessarily have to compare the situation of laser sources with that of LED sources: although a laser can be

used with either fiber type, LEDs are for all practical purposes suitable only for MM fiber.

The differences between SM and MM present the potential user with some choices that may be straightforward or uncertain, depending on the application. It may generally be said that MM is lower in cost, whereas SM offers higher performance; indeed, SM is physically capable of delivering orders of magnitude higher performance but at proportionately higher cost. One consideration over which the user may have no control is the available fiber installation. Optical fiber may have already been installed in anticipation of the need for it at some future time. This so-called "dark fiber" will be SM or MM depending on whatever future direction for the technology was foreseen at the time of installation.

Multimode fiber used in conjunction with LED sources can be very low cost. There are today a number of commercial offerings of fiber-optic transceivers designed for use with MM fiber that provide distance-bandwidth products in the hundreds of megabit-kilometers. Multimode coupling is more tolerant to physical misalignments, as explained in Chapter 2. Requirements for precision in the critical dimensions of the MM connector interface are readily achievable with good- to high-quality machining. Indeed, in certain applications, a plastic fiber with a core diameter of 100 μm may suffice.

What then would be the advantage of resorting to a laser source in any situation in which MM fiber would seem to offer capability? The answer is in the raw data rate. Modulation of an LED at anything approaching a 1 Gb data rate remains a laboratory excercise. Even if the circuits are capable of such performance, the higher drive current requirements of the LED may create more electrical noise than a laser because of the higher amounts of drive current being fed to the source. Such "delta I" noise can create a thorny electronic packaging problem. If not isolated, it can measurably degrade hardware performance.

As of this writing, it appears certain that links with data much beyond 200 Mbits will have laser sources. When it is necessary to move large amounts of data rapidly but over a short distance, a laser source with MM fiber provides a way to accomplish this while avoiding the mechanical tolerance requirements of SM. Such a need is likely to arise in a number of conceivable data communication applications, including connections between linked processors, between computational processors and terminal screens displaying real-time high-resolution, full motion videos, or between certain server units and client stations. Data pattern distortion due to dispersion, whether modal or chromatic, remains small because the total path traveled is small. The relaxed mechanical tolerances afforded by MM hardware reduce the hardware cost considerably from what it would be with SM equipment. These relaxed tolerances are more forgiving to repeated physical handling of the interconnections, a consideration that may be quite important with office workstations as compared with dedicated telephone equipment.

The employment of a MM capability in relatively short distance, high data rate data communication applications will be bound by certain signal-degrading effects, including chromatic and modal dispersion, material dispersion, attenuation, and modal noise. These concerns have been addressed in other chapters. To attain a low cost for the source, the laser used for short distance MM applications will likely have a relatively large spectral width. Roughly speaking, its spectral width will be midway, in orders of magnitude, between that of an LED and that of a distributed feedback (DFB) laser intended for long distance telecommunication applications.

Spectral width taken alone will contribute chromatic dispersion in both SM and MM cases. Spectral width will also give rise to modal dispersion in MM fiber only. These dispersive effects are small when the fiber distance is small, at least in relative terms, compared with long-haul SM applications; however, they can set some limits on the utility of MM.

A common parameter for comparison between systems is the *distance bandwidth product.* For example, a system with a distance-bandwidth product of 10 km-Gb could send information at a data rate of 10 Gb/s over a distance of 1 km; or, if the data was reduced for this same system to, say, 1 Gb/s, the information could travel 10 km without undue signal degradation. For low bit rates, systems are typically attenuation limited, whereas dispersion is the limiting factor for the higher bit rates that are of concern here. Typical maximum values for dispersion-limited systems are 2 km-Gb for MM and up to 250 km-Gb for idealized SM applications, although these cases are difficult to achieve in the field [27]. (The reader should be aware that the per second (/s) notation is often suppressed when it is understood that one is referring to a bandwidth or a data rate.)

For MM fiber, a dispersion factor of -80 ps/km/nm can be expected to limit distance to approximately 1 km for a 500-Mb data rate and to half that distance for 1 Gb. The specific distance-bandwidth product depends on such factors as achievable bit error rate, source extinction ratio, and so on. This limit will be somewhat smaller for 62.5-μm diameter fiber compared with 50-μm fiber. Modal dispersion will limit the distance-bandwidth product. The consideration of modal noise is somewhat more problematic and will be discussed further later on.

We must bear in mind that this discussion has emphasized the need for *short* distance data transmission at *high* data rates. Thus it has been possible to reduce cost with MM solutions. For much larger distances, SM fiber must be used to ensure high bit rates in excess of 1 Gb/s. For this reason, the strategy may be to use SM fiber "to the curb," supplanted by MM fiber "to the home or office." Although this chapter is not intended to detail the inherent dispersive/attenuation characteristics of SM or MM fibers themselves, we must account for both situations when developing estimates of tolerances on coupling interfaces.

5.2 Interface Definition and Importance

An *optical interface* is defined by a connection between some type of optical transmitter and some type of photonic receiver. Looked at in more detail, the interface includes several interconnection elements, specifically, all those between the electric circuitry and the fiber. The cases considered here involve the mating of a source (transmitter) to an optical fiber (receiver) and/or the coupling between a fiber (transmitter) and a photodetector (receiver). In each case, there is an electro-optic element: in the transmitter, the laser diode or LED; in the receiver, the photodetector. An optical element, generally a lens, is situated between the electro-optic element and the fiber waveguide. An electronic packaging structure has been devised that is appropriate for data communication hardware needs. It contains all of these elements and provides the required degree of precise optomechanical alignment among all elements. These have come to be known as optical subassemblies, or OSAs, and we distinguish those for transmitter and receiver ends of the link as TOSAs and ROSAs, respectively.

The OSA makes use of a transistor outline (TO) packaging configuration for the electro-optic element. The electro-optic TO can was refined for the compact disc and video disk laser markets and so provides a well-established, reliable, and inexpensive way to contain the electro-optic semiconductor element. The TO can may also contain, on the transmitter end, a monitor photodiode for feedback control of power in laser TOSAs, and may contain, on the receiver end, a preamplifier circuit chip. Leads for electrical connection, a hermetically sealed window for exit/entrance of optical radiation to and from the fiber, and a small, easy-to-handle cylindrical canister are all part of such a package design. With the optoelectronic elements packaged in this way, the problem becomes an optimechanical one of aligning the TO can to a fiber so that the optical radiation is properly conveyed between the two.

For a TO can within an OSA configuration, the source is necessarily some distance away from the fiber, sealed behind a window at the top of the TO can. Because of this setback, "butt coupling" is not possible, that is, when the fiber butt end is brought up as close as is mechanically possible to the source. Thus, we are led to consider various lens options. The simplest lens would be a simple sphere of glass, plastic, or sapphire. The refractive index and the radius determine the focusing and imaging characteristics. A simple lens with spherical curvature is, of course, an option; a more sophisticated variant is to employ an aspheric lens (molded, not ground, for low cost). Perfect image quality per se is not a requirement; rather, we wish to minimize any distortions that might affect either coupling efficiency or tolerance to misalignment.

Another effective option is a graded index (GRIN) lens. GRIN lenses can be made with all of the requisite optical properties, and they tend to have less sperical abberation than a conventional lens made of uniform material. Their

optical index gradient properties are formed by a diffusion of replacement ions into the lens body from the surface. Because they are most easily made in a small size, they are compatible with fiber dimensions. Other lens possibilities include compound lens systems and even prismatic or cylindrical lenses to shape the elliptical radiation from the source (if laser) into a more cylindrically symmetric pattern. For a low cost data communication application, however, a single lens will most likely be employed.

A popular design scheme for data communication is the connectorized module. Connectorizing in this context is the provision of an optomechanical construction at the module that can provide secure alignment and retention of the fiber end face ferrule while also allowing simple manual detachment and reconnection. For source-to-fiber or fiber-to-detector modules, both interfaces must be outfitted with a connectorized coupling mechanism. A ferrule must be attached to the fiber end and the source/receiver must incorporate a bore with which to mate the ferrule.

Many applications also use pig tailed packages in which a length of fiber "pigtail" is aligned and fastened in an immobile and permanent connection to the electro-optic element. In a pigtail, the connector is moved out to the fiber end opposite the electro-optic element, and thus is implemented as a fiber-to-fiber connector interface, rather than a source-to-fiber interface.

When using an LED, the fiber into which light is coupled would be MM, whereas SM fibers require LD sources. In either case, the fibers used must be ferruled in order to provide durability and accuracy in optomechanical alignment. These cylindrically ferruled fibers are joined to the TOSA module by "plugging" them into either a split-sleeve or solid bore located in the TOSA. Similar methods are used when mating the transmitting fiber to the ROSA, although the optomechanical constraints are less severe. As will be described later, each bore method has its advantages and disadvantages. For SM systems, we show that solid bores provide the best performance.

Generally, the trend in data communication hardware is to employ "duplex" connectors/modules, meaning that the connector hardware accommodates a pair of fibers, one each for the sending and receiving paths. In the case of pigtails, a duplex type of connector could be configured at the far end of a fiber pigtail pair. Duplex connectorizing at the module simply imparts convenience in handling and greater simplicity in the geometry of the physical components. Thus, at the transceiver module, the TOSA and ROSA are situated alongside each other and both fibers can be connected or detached in one and the same manual operation. Figure 5.1 is a picture of the IBM ESCON™ series and Fiber Channel Series (FCS) duplex modules.

The duplex connector makes use of "floating" ferrules, that is, the ferrules within the connector are allowed a certain freedom of motion and are not fixed in a rigid position. During mating, this lets the ferrules "find" the

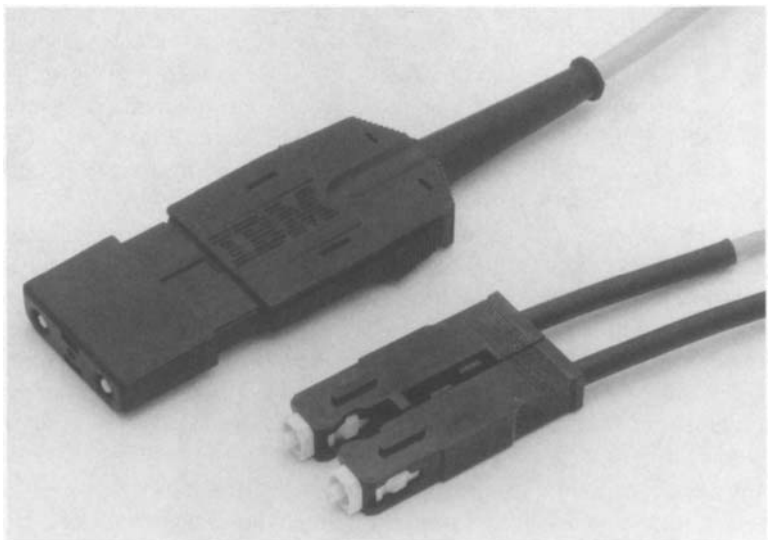


FIGURE 5.1 IBM's ESCON™ and FCS modules.

bores (which are housed within the subassembly module) and relaxes the manufacturing tolerances that would otherwise be required to achieve secure matings. These mechanical floats ease plugging by reducing the plugging forces. This approach results in less mechanical stress and therefore increased repeatability.

Two other connector types presently in common use are the FC and ST®. However, these types of connectors employ a screw-thread/retaining collar to hold the fiber ferrule secure and so do not lend themselves to the duplex design, which would prefer a push–pull action for the connection. The SC design and the IBM ESCON™ series of duplex connectors/modules are two examples that embody this design.

For the connectors at the module, coupling to the TOSA and the ROSA requires different considerations, mainly because of the large active diameter of a typical photodetector (60 to 300 μm) and the smaller fiber core diameter (~10 μm for SM and ~62.5 μm for MM). These parameters result in the TOSA/fiber interface having much more stringent optomechanical constraints than the fiber/ROSA interface. The TOSA/fiber interface (where both components are of similar optical dimensions) will be emphasized here because this is the most difficult interface to optimize. Furthermore, the optomechanical requirements for SM are much more demanding than for MM and require a correspondingly greater degree of analysis in order to understand the critical factors affecting coupling and launch. Hence, the discussion of the 12 elements of coupling repeatability later in the chapter

applies chiefly to SM. This is not to say that they are not operative in MM. Rather, optomechanical dimensions and tolerances, once satisfied for SM will also have more than satisfied MM requirements. Situations in which variations or degradation in optomechanical dimensions and tolerances can have debilitating effects in the SM case may have negligible effects for MM.

For comparison, assume there are two hardware sets, one for SM and one for MM that have the same absolute tolerances on dimensions, angles, and distances. A misalignment in the SM hardware has caused a reduction in coupling efficiency that is barely tolerable. For example, a transverse displacement of fiber so that the launched power falls off by 1 dB, or about 20%, from the maximum value when all components are perfectly aligned. A loss of this amount could occur with a transverse displacement of fiber core by about 1 μm. Now compare the MM case. It takes about 15μm of similar MM displacement (50-μm diameter core) to give rise to the same fractional loss of power [15]. The *analysis* of coupling geometry and efficiency is more difficult for the MM case; however, MM hardware design is superfluous once hardware has been developed for the MM case. Connector hardware parts that fail optomechanical tolerances for SM are often quite serviceable for MM applications.

The proper conveyance of the optical radiation from the optical source, whether laser or LED, into the fiber so that a guided mode is established is called *launch*. As we examine some important operative factors in the launch, some significant differences between SM and MM will become apparent. The most obvious difference is that launch into SM must be more difficult and probably less efficient than with MM simply because of the smaller fiber size in the SM case. (The end face area of the SM core is at least 25 times smaller than the MM core.)

It should be understood that there is a problem in the launching of laser light into a fiber that is fundamentally different from the LED case, apart from concerns about the relative difference in size between SM and MM fiber. It is natural to think of launch as akin to directing a somewhat divergent light beam into a narrow dielectric (i.e., glass) tube. For the LED case, this situation is largely the extent of the problem. A laser on the other hand is not simply a source of light; rather, it should properly be thought of as an electromagnetic cavity resonator from which energy in the cavity leaks out as radiation from the ends. (Depending on the mirror reflectivity some light emanates from the end opposite the fiber, where it is usually intercepted by a monitor photodiode.) The problem in the launch of laser radiation is one of arranging a good alignment or matching so that radiative output is established as a guided mode within the fiber optic waveguide.

The launch and the sustaining of wave propagation should be done with as little reflection back to the laser as possible. This condition is not simply a matter of efficiency (after all, any light reflected back is necessarily removed from the forward direction). Reflections back to the laser are disruptive to the

very oscillatory behavior that creates laser radiation in the first place. (Dithering or self-pulsating may work to reduce this reflection sensitivity.) LEDs, even edge-emitting types, are lacking the characteristic of laser cavity resonance and thus are not affected by reflection. Reflections can and do occur at each interface of an abrupt refractive index change. These interfaces include the TO can window, surfaces of lens elements, fiber endface surfaces, and the detector surface. All of these surfaces can, in principle, be anti-reflection coated, and this coating is applied to the window and lens surfaces as well as to the detector surface. In fiber connections, it is far better to rely on a good physical contact of the abutting glass endfaces. This approach leaves the initial entrance and final exit endfaces at the two ends of the fiber link. In principle, these two surfaces could be antireflection coated or polished at an angle large enough so that reflected light would be deflected from the nominal optic axis of the system. However to do so would complicate the hardware set, as these fiber ends would be distinct from those throughout the rest of the link.

An effective way to reduce the effect of endface reflections at the laser is to *intentionally* defocus the optical arrangement at launch. The laser, lens, and entrance endface should be arranged so that an image of the laser endface is formed some distance (roughly 10 to 100 μm) before the fiber endface. At the plane of the fiber endface is an enlarged defocused spot. The loss of launched power resulting from the reduction in coupling efficiency can be made up for by increasing the laser drive current, which has its own benefit of generally tending to reduce laser intensity noise. Here is a case in which we can deal specifically with the characteristics and requirements unique to data communication applications as opposed to telecommunication. Because of the relatively shorter data communication links, with their correspondingly smaller attenuation, one is not necessarily interested in attaining maximum launched power. For safety concerns, defocus should be arranged so that the laser endface facet is imaged *infront* of the endface of the receiving fiber, not within it. In this way, the light launched in the fiber will be at its maximum value when the connection is fully seated. Were defocus to be such that the laser endface image was inside the fiber, loosening the fiber connection a bit could result in an increase in power into the fiber—a possible safety concern.

5.2.1 Understanding Light Coupling

In considering the coupling problem, perhaps with a source whose emitting area is greater than that of the fiber cross-section, people often imagine the utility of something that functions as a "light funnel." For example, they might ask why a lens cannot collect and focus (in effect, funnel) the light from the source down to a fine point whose dimensions are smaller than the cross-section of the fiber, or, alternatively, whether a glass taper can be constructed that would be large enough at the entrance to accept all of the

light from the source and direct it into a narrow pencil, again with an exit cross-section smaller than that of the core of the accepting fiber.

Unfortunately, the "light funnel" approach is generally not possible [12, 20]. For MM systems, the accepting fiber (and any taper coupler as well) necessarily has a certain maximum numerical aperture, NA. If the path of light exceeds the acceptance angle, θ_a, the light will not be guided (remembering that $NA = \sin \theta_a$). The behavior of the light in this situation is governed by some fundamental principles of physical systems which show that attempts to constrict the cross-sectional area of a beam of light will expand the range of directions of the rays within the beam. The fact that rays within a light beam extend throughout a large angular range is of little concern if one wants to illuminate a small region, as in focusing light with a condenser to illuminate a sample in a microscope. To bring about launch at a fiber, however, it is necessary to do more than illuminate the endface of the fiber. The light must fall within the acceptance cone of the fiber (see Ref. 28 for discussion and representations of ray paths after successive internal reflections within light funnels). Multimode technology almost exclusively uses graded index fiber, so the NA is maximum at the core center and decreases radially [4]. The fact emphasizes the value of good radial coupling alignment.

In the MM case, coupled power can be calculated through a surface integral performed over the extent of the core endface which accounts for all rays from the source that fall within the NA of the fiber. Using Snell's law and considering the refractive index values for the core and cladding, one can compute whether or not a given ray meets the conditions for total internal reflection at the core-to-cladding interface. If the interface has a coupling lens, the emission pattern of the source will have to be transformed appropriately to account for the redirection of the source rays by the lens. The integration becomes yet more complicated if the source (laser) lacks cylindrical symmetry or if the source and fiber axes are not aligned. Fairly readable presentations of integration for the simpler cases can be found in References 5 and 16.

For SM fiber, it is not really correct to use the concept of a NA. Numerical aperture considerations arise in ray models of source-to-fiber coupling and launch (which do not apply to SM theory). The proper treatment of SM fiber proceeds differently. Efficiency of coupling is determined through the calculation of an overlap integral. The argument of this overlap integral is essentially a product of the field distribution of the beam incident to the fiber endface and the field distribution of the fundamental mode of the fiber. The field of the incident beam is determined by the angular and spatial characteristics of the source emission and possibly by the effects on its geometry of any intervening lens. The shape of the fiber's fundamental mode is determined by Maxwell's equations for a waveguide, given the physical dimensions and reflective index values of the guide and the wavelength of light. For a detailed

discussion of the determination of coupling efficiency through calculation of the overlap integral, see References 8, 17, 22, and 23.

5.2.2 Minimum Coupled Power

Any time two mechanical/optical components must be brought into contact with each other there will be statistical variations. Inasmuch as we are concerned with the pluggability of fiber optic connectors into TOSAs, both mechanical and optical characteristics must be considered when determining the amount of coupling deviation that is acceptable. As explained previously, the amount of light that can be coupled into a fiber from a source is reduced by mechanical misalignments and optical mismatches. To achieve successful data transmission, criteria must be established that put a lower limit on what is considered "enough" coupled power. The *coupled power range*, CPR, is a measure of how much variation in light transfer is allowed when mating any given fiber with any given transmitter OSA to ensure adequate system performance and yet not violate laser safety requirements.

To account for all of the losses along a system's length, a concept referred to as the *link budget* must be understood. A system loses power due to variations in connector/module matings, attenuation, and because of the splices and connectors themselves. After launching light from the source into the fiber (which is a variable emphasized in this chapter), the light is attenuated according to the cable length. Each fiber-to-fiber splice and connector along the path will further reduce the signal power. Finally, when the light emerges at its destination, it must be strong enough for detection. By "strong enough" we mean that each bit of information must still be distinguishable. The degree to which the information must be accurate is referred to as the bit error rate (BER). The BER is the probability that a 1 is mistaken for a 0, or vice versa. (What in fact determines a true 1 or 0 depends on the specific protocol used by that system.) Data communication requires a BER of 10^{-15}, which means that only one in 10^{15} can be erroneous. (Telecommunication requirements are more relaxed and can be as low as BER = 10^{-9}.)

In any system of signal transmission, the frequency of errors is determined by the signal-to-noise ratio. In a fiber-optic digital data transmission system, the signal strength is essentially the difference in optical intensity in the beam between the bright and dim levels, corresponding to the logical 1 and 0 levels, or the corresponding voltage levels at the receiver output. The ratio of the output power for a 0 to the power for a 1 is called the *extinction ratio.* The required minimum coupled power is determined by the target BER, along with the attenuation of the link, the extinction ratio of the modulation, and the noise accompanying the signal (on both logic levels). The noise in the output of a laser diode is, among other factors, a function of the output power relative to threshold. Each light level, bright and dim, corresponding to the two digital logic states also carries its own noise level. By increasing the

difference in intensity between the two states, the signal strength is increased; hence, there is an increase in the sensitivity of the system to extinction ratio. Perhaps another way to put it is this: given a certain amount of noise in the hardware and certain attenuation characteristics of the link, and once a tolerable frequency of errors in transmission is specified, then the optical intensity of the signal is necessarily dictated. If, on the other hand, one is faced with certain prescribed limits on optical power, the allowable intensity difference between 1 and 0 will be limited and whatever finite amount of noise accompanies the signal will limit the attainable BER

In data communication applications, the standard BER of 10^{-15} represents no degradation in performance from what has been established over the years using copper wires. Voice communication in telecommunication is much more tolerant of errors because aural perception is much more tolerant of lower bandwidth and more noise. Some low performance fiber links tout specifications midway between the two values.

Figure 5.2 shows a typical star coupler link for a SM gigabit system operating with light of wavelength 1.3 μm. Roughly speaking, today's SM connectors are manufactured such that each connector interface contributes an average of 0.5 dB of loss. Fusion splicing can be easily performed with losses less than 0.05 dB per slice. Additionally, attenuation degrades the signal at about 0.3 dB/km (λ = 1.3 μm). Given a 10-km line with eight connectors and four splices, the losses resulting from attenuation and splices/connectors are about 7.2 dB. If this link is part of a mini-network that passes through a 4 × 4 star coupler, this coupler could contribute additional losses, resulting in roughly 14 dB total loss (for transmission from source to detector as shown). The minimum light that a typical photodetector can receive for a successful BER of 10^{-15} is about −23 dBm as discussed in Chapter 3. This leaves a minimum launching power of −9 dBm. Thus, the TOSA and connector must achieve at least −9 dBm of coupling *every time* (say,

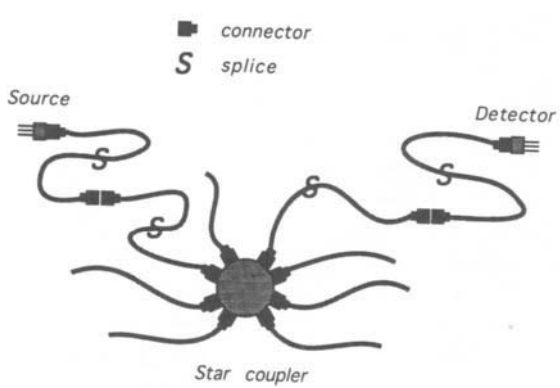


FIGURE 5.2 Link budget example.

within a 3-sigma limit or less than 27 faulty connections per 10,000). These numbers will, of course, vary depending on particular system implementations.

There is a distinction between dBm and dB: dBm is relative to 1 mW, whereas dB is always a comparison of two entities, both of which must be specified (or understood from context). Absolute receiver sensitivity is in units of actual received signal power and dBm is therefore an acceptable unit of power. Whenever a receiver is said to have a sensitivity of so many dBm, it must also be understood that this is with respect to an attainable BER, which again is either specified or understood (e.g., 10^{-15}). There is opportunity here for sloppiness if one is not careful, like gas mileage claims for auto engines.

5.2.3 Maximum Coupled Power

Maximum coupled power considerations are more prevalent in laser systems (compared with LED) simply because of the higher available intensities. Some LANs will be user-accessible, meaning that nontechnicians may be able to connect/disconnect optical fiber cables between computers just as (electrical) serial ports have been used for the past 20 years. This access provides robustness, but also presents potential safety hazards to the user (in laser applications). One successful method of user protection in the event of a simple disconnect between connector and module is implemented in IBM's ESCON™ series interconnect products [3]. A two-tiered shield system effectively blocks light emerging from a TOSA using spring-activated physical barriers that close when the connector is removed. Still, special precautions are necessary in case someone accidentally cuts a cable along its length and then exposes herself or someone else to the light. Additionally, there is concern that technicians may be exposed to focused/amplified light when inspecting an optical link for problems [1]. As previously mentioned, international laser safety standards place a limit of less than 0.600 mW of optical power in the link [14] and the ANSI standard is less than 100 s of exposure to 1 mW of light at $\lambda = 1.3$ μm or 1.55 μm [1]. Applying this to our link budget example, the most coupled power allowed would then be −2.2 dBm (0.600 mW), giving a CPR of 6.8 dB (−2.2 − −9.0 = 6.8). Additional "guard-banding" may reduce the maximum coupled power to −4.5 dBm, resulting in a CPR of 4.5 dB.

This situation is unique to data communication because the telecommunication industry typically has user-inaccessible optical fibers. Thus, if the telecommunication power received at a detector is too weak, the laser output can simply be increased with much less concern for safety. The limit of maximum coupled power is more likely to be dictated by the characteristics of the components: exceeding threshold power limits can result in reduced/

incorrect responses. Undesirable nonlinear effects in fibers can also arise at high power levels.

As mentioned earlier, telecommunication usually involves many users for a single fiber-optic cable, whereas in data communication, each user potentially has his or her own cable and transceiver module that he or she would be able to patch into personally. Thus, the size and expense of each module must be reduced, which precludes the use of isolators, attenuators, and so on that could reduce the safety risk to the user. Therefore, it is the goal of data communication engineering to produce precise efficient fiber-optic systems that guarantee a small CPR such that extra components will be unnecessary. This keeps the components small and inexpensive.

Most considerations of power output limits have to do with the possible hazard to the eyes from the radiation output of laser diode sources. Another reason to have limits on output power, even with LED sources, is receiver satuaration. In MM systems, the coupling efficiency of light exiting the module can be expected to be high, and yet allowance must be made for variations in the strength of source output. Receiver sensitivity can be very high, approaching -40 dBm or better in these systems, yet the dynamic range of the receiver may be limited. If the power at the detector is too high and saturation occurs, response speed can be degraded to a point where link performance may be compromised. Appropriate values for limits on output power are determined by the characteristics of the specific hardware and by the expected attenuation of the intended link.

Instead of ensuring limits on maximum emitted power by fixed setting of average laser bias and by imposition of tolerances in the optomechanical connection, conformance to safety limits can be achieved through the use of real-time controls built into the circuitry that controls the laser diode output. For example, assuming that we have duplex modules and some form of communicating circuitry between the transmitter and receiver sections in each module, let us label laser T1 and receiver R1 in the first module, and T2 and R2 in the second module. Then, T1 sends to R2 and T2 sends to R1. We also need some form of sensing and communicating circuitry from each receiver to the laser transmitter next to it within the *same* module. The sensing circuitry must be able to detect a loss of signal and convey this information to the laser transmitter next to it in the same module. Upon receipt of this signal, the laser is wired or programmed to shut off.

To start (or restart) a link, test pulses can be sent out and processed in similar fashion by the receivers. Once the control circuirty has established that an intact link exists (through detection of the test pulses by the receivers and notification of same to the laser sources), a test pattern can be sent and an error rate determined. The signal strength can be increased up to such a level so as to ensure a bit error rate below a certain value. This signal strength can exceed the safety threshold value for the human eye because it will already have been established that the link is intact (in both directions). Furthermore,

if the link is subsequently opened or broken, a loss-of-light signal will be generated by the receiver and will bring about shutdown of the lasers according to the sequence just described.

Such circuits, called *automatic power control* or *open fiber control*, are permissible under IEC rules. ANSI publishes possible designs for circuits that will operate in the way described. Use of such circuits will then allow a link to function at a higher power level than it would otherwise. Without such an automatic power control, each source would have to be built in such a way as to ensure that *each* one never emitted power above a certain level. Furthermore, the coupling efficiency of each source would never be allowed to exceed a certain value for *all* fibers into which it would be plugged. Thus, each source would have to be individually calibrated and trimmed and all connector hardware controlled so as to guarantee a launched power efficiency high enough to ensure a workable link, yet low enough to ensure conformance to safety levels.

This approach diverges from the primary focus of this chapter and further treatment of it will not be taken up here. It is mentioned so that the reader may be aware of the different ways in which the problem of laser safety and optical link power can be resolved.

5.2.4 Modal Noise

Laser sources used together with MM fiber can give rise to a kind of signal noise called *modal noise* that is generated at fiber-to-fiber connections where there is some degree of transverse misalignment of one fiber with respect to the other [9, 10]. (Excessive axial separation of fiber end face pairs can also give rise to this effect, but transverse misalignment is apt to be the main cause in actual links.) The noise is optical in the sense that it is a fluctuation in the optical intensity. Of course, this optical fluctuation is converted into a corresponding fluctuation in voltage at the receiver's detector circuitry and hence becomes electrical noise at the receiver.

Modal noise arises from properties that are intrinsic to the behavior of coherent light in MM fiber, or more generally, in any situation in which single coherent light source is divided among several paths and then recombined. In MM fiber, the different paths are the modes themselves, which, in order of magnitude, number between 100 and 1000 in the usual MM fibers.

Modal noise has its origins in the subtle details of the physics of laser and fiber behavior and of the effect of the first upon the second. Before proceeding further with a description of the phenomenon, it may help to clarify the concept if we say a few words about modes. There are oscillation modes of the laser and propagation modes of the fiber. The two are distinct. Each oscillation mode of a laser results in output of light at a particular wavelength. Light at a given wavelength that radiates from the laser and is launched into the fiber propagates within the fiber along a number of distinct paths, each of

which is a propagation mode. Modal noise arises from interference among the fiber propagation modes, not among the laser oscillation modes. Having said that, it should be understood that the fiber propagation modes obtain their energy from the set of laser oscillation modes. If the energy distribution changes among the different laser oscillation modes, the pattern of energy distribution among the fiber propagation modes will change. The phenomenon is refered to as *mode partition noise.*

The reader may already be familiar with interferometers, such as Michelson or Mach-Zehnder interferometers, or with the steering of radar beams using phased-array antennas. In the interference, a light beam is deliberately split and sent along two different paths and then recombined. In the radar case, the same signal is fed to each antenna element of an array, with possibly a phase shift deliberately induced among certain of them relative to the others. Signal intensity upon recombination is determined by the kind of interference that occurs between or among the wave fronts from the different paths. In the case of MM fiber, the different paths are the fiber modes themselves.

Imagine a plane cutting transversely through a fiber at some point in the link. The time that it takes for the radiation from the source to reach that plane depends on which propagation mode is adopted. (The pulse spreading that occurs because of this dependence is *intermodal dispersion*, mentioned earlier as the dominant factor limiting distance-bandwidth in MM fiber links.) As we stated earlier, in the MM fiber there are more than 100 modes and these are all recombined at our imaginary cut. More precisely, we should say that all the modes interfere. The result is a variegated pattern of intensity across the face of the cut, with bright regions attributed to constructive interference and dark regions attributed to destructive interference. Figure 5.3 is a simplified representation of the (fiber) modal interference pattern in the plane where the exit endface of sending fiber A overlaps that of the entrance endface of receiving fiber B. The amount of transverse misalignment between the two endfaces has been greatly exaggerated for purposes of illustration. The interference pattern of the light from all the modes is represented by three oblong regions of light. An actaul interference pattern would show a mottled arrangement of bright spots, the number of which would be, in order of magnitude, the same as that of the mode count for the fiber. For simplification, we show only three such regions—R_1, R_2, and R_3. Figure 5.3a shows the initial distribution. At time T_1 R_1 and R_3 are both accepted by fiber B, whereas only part of region R_2 is accepted. Figure 5.3b shows a later time, T_2 when the pattern of intensity distribution has shifted somewhat. Only R_1 will be entirely accepted by fiber B. R_2 is still partially accepted, but R_3 is lost entirely. This loss of coupled intensity, which is time-dependent because of the time-varying nature of the interference, is modal noise.

This distribution of intensity at a given cross-section would remain fixed and stationary if optical path lengths for each mode of the propagating set remained fixed; however, the physical environment of the fiber is never

completely static. Mechanical stress and local temperature change the refractive index of the glass, thereby changing the net optical path length for each of the modes. Furthermore, the nominal refractive index of the fiber and that of any mode function of wavelength. Mode partition noise in the laser output also causes a time variation of the distribution of intensities in the various laser mode wavelengths about the laser's nominal centerline frequency. As the distribution of output wavelengths shifts, so does the energy distribution among the various propagation modes of the fiber. The net effect is a fluctuation in the pattern of interference among the fiber modes at a connection interface. Fluctuations on the modal interference pattern due to changes in the fiber (mechanical and thermal strain) will occur on a time scale representative of the strain mechanisms, that is, on the order of milliseconds

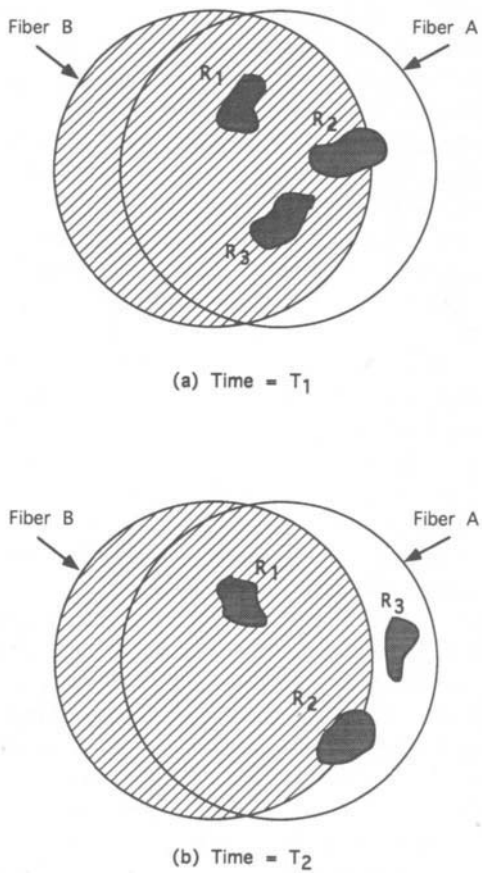


FIGURE 5.3 Simplified physical interpretation of modal noise.

or longer. Fluctuations due to mode partition noise will occur faster, on the order of picoseconds.

Thus it is that the intensity pattern at the fiber cut fluctuates. Such an effect can actually be observed by illuminating a MM fiber with a helium-neon (He-Ne) laser and directing the output from the fiber onto a sheet of white paper or onto a screen in a dimly lit room. The fluctuation can be enhanced by gently contorting the fiber by hand, twisting, bending, or rolling the fiber between the fingers. Actually, one can even see a SM form of modal noise as well. By illuminating fiber made for operation at 1300 nm with the light from a He-Ne laser (output at 633 nm), more than one mode can propagate because of the shorter wavelength of light used. The output will likely not show a Gaussian profile. One may observe a doughnut-shaped or other pattern representative of a higher order mode or an interference pattern of several modes.

It bears restatement that the intensity fluctuations that are modal noise are necessarily a result of the use of coherent light in MM fiber and are unrelated to any variation in signal strength that may be present at the output of the source. Because modal noise is created within the fiber as an effect of intermodal interference, the magnitude of the interference is proportional to the intensity of the light; hence, it does not disappear or diminish with an increase in signal strength at the source. Therefore, modal noise is proportional to the light intensity or signal strength and cannot be reduced by increasing the optical power, as the modal noise will simply increase proportionately. Whenever it is the case that modal noise is the dominant source of noise, an upper limit on attainable signal-to-noise ratio is imposed and therefore a limit as well on attainable BER. The only ways to lessen modal noise are to reduce the coherence of the source or to reduce the mode-selective loss throughout the link, generally by improving the coupling efficiency of the connectors.

Modal noise effects can be demonstrated in the following way. A link is set up with MM fiber and a laser source. Included in the link should be a variable attenuator and a fiber connection allowing adjustable transverse misalignment but initially set up in good alignment. The link is set up within a BER test station where the frequency of errors in transmission can be monitored and tallied according to various link parameters, including attenuation. The variable attenuation is increased, thus decreasing the signal intensity at the receiver until errors begin to occur regularly enough to be counted easily, say, about one error per second. One then replaces some of the variable attenuator's attention with deliberate misalignment of the adjustable connection. Even with a one-for-one replacement of loss through the attenuator with loss through the misaligned connection such that the optical intensity at the receiver remains the same, the frequency of errors is higher than before. One can only reestablish the initial error rate by reducing the variable attenuator setting somewhat more. This necessary additional reduction in attenuation is

called a penalty and represents the added effective loss that is induced in the link as a result of the effects of modal noise generated at the misaligned connection. For this reason, loss at MM connectors due to fiber misalignment is called mode-selective loss. An example of loss in a fiber link that is not mode-selective is the distributed attenuation throughout the entire length of the fiber.

Having said all of this, it should be kept in mind that modal noise need not be of serious practical concern if all fiber connections are of high efficiency and if the detector at the receiver end of the link intercepts all of the fiber output. A point emphasized throughout this chapter is that the great strides that have been made in developing high-quality SM connector hardware have benefited MM even more.

Because modal noise arises from intermodal interference at connectors, modal noise effects where present, will be worse at longer wavelengths. This situation occurs because the coarseness of any pattern of interference, constructive and destructive, is proportional to the number of interfering modes, and the number of allowable propagating modes in a given MM fiber is inversely proportional to the square of the wavelength. A more coarse interference pattern reduces modal noise effects.

Modal noise effects are essentially eliminated by using an incoherent source, that is, an LED. The broad spectral spread of an LED's output means that in effect, the light within the fiber propagates among an infinite number of modes. The mottled interference pattern across the fiber endface in the coherent case washes out into a relatively smooth distribution of intensity under an approximately Gaussian envelope that is symmetric about the fiber axis when an LED is used as the source.

It may be the case that because of the demands of the application environment, one desires the higher modulation rate of a laser together with the relaxed optomechanical tolerances of MM fiber. In such a case, lower coherence of the laser source is beneficial. True sustained laser action produces coherent radiation, so any effort to reduce source coherence is in some sense at cross-purposes with using a laser in the first place. In an effort to preserve the high bandwidth modulation capabilities of a laser and yet make the source more suitable for MM applications, methods can be tried to reduce the coherence of lasers. Two methods are to use "self-pulsating" lasers and to "dither" the modulation signal [6, 13, 29].

Both approaches exploit the fact that a laser's output is of relatively low coherence as it is being turned on and shortly thereafter, as the applied bias is brought up from near threshold to a value corresponding to the bright output level (compared with its output during a sustained and steady DC bias). A *self-pulsating* laser is one that is constructed in such a way that its bright output fluctuates or pulsates because of a high frequency "self-throttling" action imparted by the dynamics characteristic of its cavity structure or by the way current is supplied to it. *Dithering* attempts to

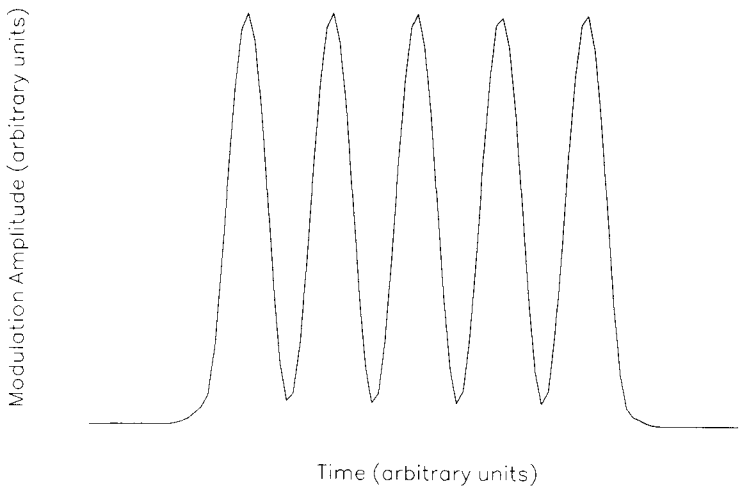


FIGURE 5.4 A dithering signal.

accomplish rather much the same thing but directly, through a high frequency oscillation superimposed on the high levels of the modulation. Figure 5.4 shows a representation of a dithering signal. In either case, the laser output for the bright state does not maintain a relatively constant intensity throughout the duration of the bit period but rather fluctuates rapidly. If there is a sufficient number of oscillations of intensity within the bit period, the oscillations in the electric signal at the output of the receiver can be removed with suitable filtering ahead of the decision circuitry for discrimination of logic level. The difficulties of either approach are fairly obvious. Self-pulsation requires specially tailored construction of the laser diode. Dithering requires electronic circuits that can supply a modulation "overtone" at a frequency higher than that of the data rate.

5.3 The Twelve Elements of Coupled Power Repeatability in Single Mode Systems

It is well known that micron-level tolerances are required to achieve acceptable performance in SM fiber-optic technology. Multimode applications are somewhat more forgiving, but must none the less provide low variability. The mechanical coupling of the TOSAs and ferrules presents a unique challenge when designing a reliable optical interface because the variables associated with this interface determine its performance, reliability, and safety.

In the remainder of this section, only SM systems will be discussed. A similar approach to what follows could be applied to MM applications, for

which optomechanical tolerances are much less stringent. Although laser-to-MM fiber systems may have additional complications, it is the intent of this section to outline the practical considerations for data communication connections; therefore, laser-to-SM fiber will be used as the specific example.

Poor control of the many tolerances involved can cause two basic problems in connector/module mating. The first problem can occur when the same connector is plugged into the same module and extensive variations in the launched power are obtained. This variable is referred to as *plug repeatability* (PR). Plug repeatability is a concern when one is conducting experiments to determine the effect of aging or stress on a single module or connector. Two decibels is typically the most extensive acceptable PR variation. Many tests measure fluctuations smaller than 2 dB, making it difficult to resolve changes if PR is greater than 2 dB.

The second variation is the *cross-plug range* (XPR). It is measured by subtracting the low power reading from the high power reading in a large sample of different connectors mated to the same module. This variable is typically measured in decibels and its importance lies in the fact that any module may be connected to a large number of different connectors. A large PR is of concern because it is a significant source of the variation of the power launched to any given link. It is important to have a number for XPR, as the goal of communications engineering is to minimize *fluctuation* in systems: too large an XPR indicates wide variability and thus the system could and should be modified to reduce this variance. In many cases, improving the peak performance is less important than ensuring a small change in performance when interchanging connections. For example, a system with a coupling efficiency range of 40 to 50% (-4 to -3 dB) may be better than one whose efficiency is 60 to 90% (-2.25 to -0.5 dB), as the latter has more variation.

Plug repeatability and XPR are a concern when mating TOSA to fiber, but not as much for fiber-to-fiber coupling, in which repeatable, low-loss connections are more easily made using techniques borrowed from the telecommunications industry. This better repeatability is a result of the tighter tolerance control in fiber-to-fiber mating than in TOSA-to-fiber mating. It should also be pointed out that the physical alignment of the fiber with the TOSA laser spot is much more demanding than the alignment of the fiber with the ROSA detector. Hence, this discussion will emphasize TOSA-to-fiber alignment.

To reduce XPR between components, we must establish the sources of coupling variations. Once these are defined, it becomes an engineering problem to optimize component performance. One can prioritize on the basis of which parameter causes the most variance, but feasibility and cost must also be considered. In an effort to account for TOSA-to-fiber variations, twelve elements of plug repeatability have been designated and are explained as follows.

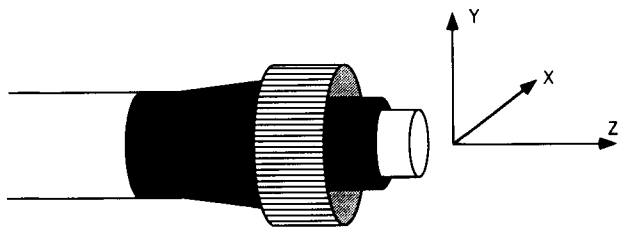


FIGURE 5.5 Coupling coordinate system.

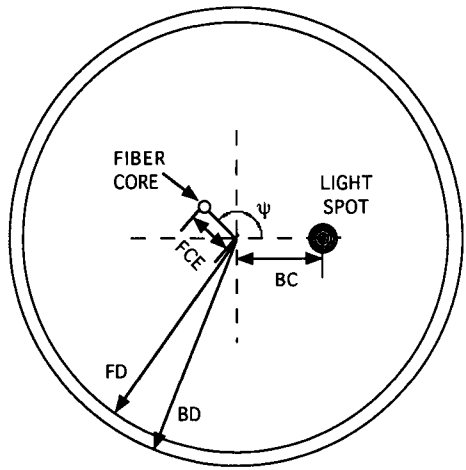
5.3.1 Elements 1–4: The Planar Alignment Elements

Physical misalignments are the most obvious source of coupling variation when mating two optical components such as a ferrule and an OSA bore. We can define the interface plane as the plane normal to the optical axis. Adopting this convention allows one to think of "Cartesian" misalignments between the two devices being coupled. Longitudinal displacements occur along the (optical) z-axis and are not significant [17]. In a Cartesian system, then, we also allow some lateral misalignment in the x and y directions. These variables can be seen in Figure 5.5 and are the directions of misalignments considered here.

There are four elements that contribute to the Cartesian misalignments between the laser spot and the fiber core, all of which are measured in microns. Two are attributable to the module and two to the connector. (Each of the elements in the module has a mate in the connector and vice versa.) Element 1 is *beam centrality* (BC) or the position of the laser beam axis with respect to the mechanical axis of the bore. It is desirable that the two be coaxial (making BC = 0) so that the laser beam center is exactly in the center of the bore. Element 2 is the corresponding centrality misalignment element in the connector, known as the *ferrule/core eccentricity* (FCE). It is nonzero when an imperfect ferruling process results in a fiber core that is non-concentric with the ferrule.

These parameters are shown in Figure 5.6. Note that to ease interpretation, it is assumed that the ferrule fits "perfectly" within the bore so that the ferrule and bore centers coincide. Elements 3 and 4 determine the degree of "coincidence," as will be explained. Therefore, we can define an angle, ψ, between the directions of BC and FCE with respect to some fixed coordinate system (ψ itself does not constitute a new element). For given BC and FCE, the ψ value will be randomly distributed throughout 360 deg. For $\psi = 0$ deg, the misalignments will have the least effect, whereas $\psi = 180$ deg will place the fiber core the farthest from the light spot. The scales of the BC and FCE have been grossly enlarged in the figure for clarity.

The BC alignment is achieved by the manufacturer in a process called *active alignment*. This technique is performed by maximizing the power launched into a reference cable (one with < 0.5 μm FCE) while adjusting the laser's



Note that the ferrule and bore centers coincide

FIGURE 5.6 Beam centrality (BC) and ferrule/core eccentricity (FCE), exaggerated scale.

transverse position (perpendicular to light travel) within the bore. Coupling as a function of purely transverse offset is well documented [17, 21]. When the coupled power indicates an acceptable BC, the laser diode is then welded within the bore. The 0.5-μm FCE error added to the bore/ferrule clearance (Elements 3 and 4) of 2.2 μm average is present before the laser is welded. The welding process adds another 2.5 μm. These three errors added statistically result in an offset of approximately 3.3 μm or less, but unfortunately this parameter is presently difficult to measure to better than ± 2.0 μm.

The FCE is typically controlled to < 1.6 μm. Anecdotal data from a vendor indicates that this element is usually < 0.7 μm; hence, FCE is less critical than BC. To our knowledge, no commercially available tool exists to measure FCE. In the laboratory, however, people have used photodetectors to measure the movement of the light spot coming out of the fiber as it is rotated in a V-groove to measure the FCE to < 0.3 μm precision.

In connectorized fiber-optic components, it has been mentioned that it is standard procedure to ferrule the fibers. These ferrules are then inserted into bores in the OSAs that house the transmitter or receiver. Two styles of bores are available: solid and split-sleeve. A *solid bore* has complete rotational symmetry in the sense that it is a "perfect" cylinder, as shown in Figure 5.7a. It is typically constructed of some rigid ceramic material. The *split-sleeve bore*, in contrast, is separated lengthwise along the cylinder wall, as shown in Figure 5.7b. It is made out of a more flexible material. The split-sleeve bore is designed so that its effective diameter will enlarge slightly to accommodate a larger ferrule.

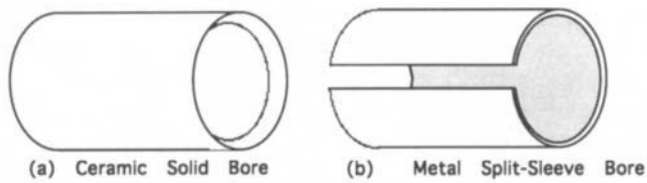


FIGURE 5.7 Solid and split-sleeve bores.

Elements 3 and 4 are the solid *bore inner diameter* (BD) and *ferrule diameter* (FD). The difference in these two diameters creates another offset in solid bore TOSAs that can vary from insertion to insertion in the same module/connector pair. Hence, these two elements are critical in PR for solid bore TOSAs. These elements are shown in Figure 5.8 for an arbitrary direction, ϕ, of displacement. If FD − BD = 0, then ϕ is meaningless and we revert back to Figure 5.6. Depending on the ψ orientation between BC and FCE, ϕ can either improve coupling or degrade it. Because ψ and ϕ are random variables, the best we could hope for is to establish a *range* of possible distances between the fiber core and the light spot. It should be observed, however, that split sleeve TOSAs will not be sensitive to these elements; by design, the split-sleeve walls flex to achieve a snug fit for a wider range of FD.

Typical values for the BD are 2.5010 to 2.5025 mm. For the ferrule dimensions, some specifications require a diameter of 2.4992 to 2.5000 mm, within a 3-standard deviation criterion. These parameters give the largest ferrule-to-bore difference possible of 3.3 μm. The minimum difference is 1 μm, and assuming normal distributions, the average would be 2.15 μm.

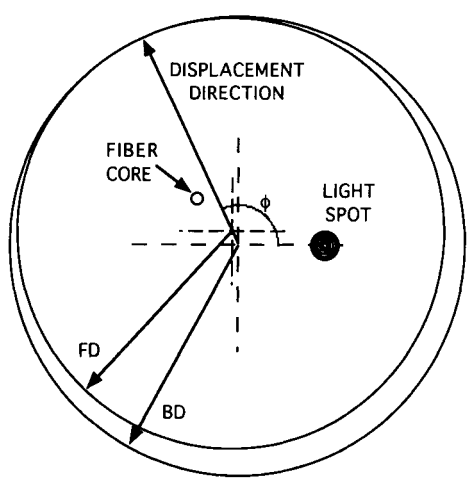


FIGURE 5.8 Bore diameter (BD) and ferrule diameter (FD), exaggerated scale.

Current interconnect products use some method to reduce the variability of longitudinal alignment. Although a given TOSA might not guarantee that the laser beam waist coincides precisely with the fiber endface (indeed, it may be intentionally defocused), the physical pressure exerted on the connector when mated with the TOSA is such that the longitudinal position is consistent between matings. Also, based on the theory of coupling between two quasi-Gaussian beams, it has been shown extensively that a small longitudinal misalignment has negligible effects on the power transfer [18, 22]. For this reason, it is not considered here.

5.3.2 Elements 5 and 6: The Angular Alignment Elements

The ferrule/bore diameter difference can lead to an additional source of misalignment known as the angular offset, or *tilt*. However, because of the small lateral displacement of BD − FD = 3.3 μm, and the long length of the bore of approximately 4 mm, this tilt will be small: the largest tilt resulting from the ferrule/bore difference would be 0.05 deg. This can be shown to contribute insignificantly to the CPR using the theory of Nemoto and Makimoto [22].

There are however, other possibly significant sources of tilt. One potential cause of tilted beams can be found in the TOSA. Even if the beam center spot can be successfully imaged to the bore center (i.e., BC = 0), the laser may possess some tilt, as shown in Figure 5.9. Element 5 is this tilt, known as the *pointing angle* (PA). As explained previously before the laser and bore are welded, the laser is repositioned for maximum coupling. If there is a significant PA to the laser light, the weld will still be optimum for a small BC, as shown, and the engineer will be oblivious to the fact that any tilt is present.

Although the BC can be measured, the tilt of the laser diode beam within the bore is extremely difficult to evaluate. Moreover, the manufacturing process of the TOSA allows little control over PA. Presently, manufacturers

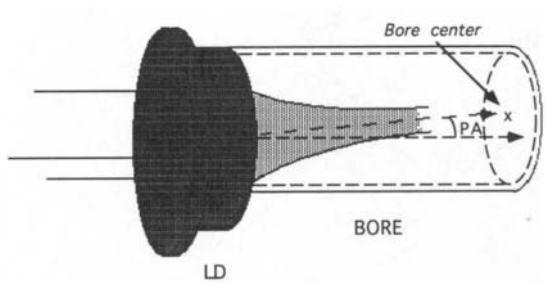


FIGURE 5.9 Laser pointing angle (PA).

use TO packaging for the TOSA because it is cheap and well tested. (The *TO* in TOSA has nothing to do with its being based on TO can designs.) This packaging may suffice for CD-players, but the impact of poorly controlled TO technology on datacommunication is generally underappreciated. The PA can be as large as 3 deg, and must be taken as an unknown. It is hoped that as datacommunications become more common, the need for well-controlled tolerances will drive manufacturers to produce modules with much smaller laser tilts. A necessary precursor to this optimization is the development of an accurate measurement technique. Presently, this technique has not yet been developed. Therefore, considerable fluctuation in XPR between TOSAs can be attributed to this tilt.

Element 6 is the fiber *endface angle* (EFA) which is a result of the polishing process of the ferruled fiber. Fibers are typically arc-polished so that the hemispherical tip deforms slightly when two ferruled fibers are butt-coupled. This process, known as the PC polish ("physical contact") assumes that the polished apex coincides with the fiber core center. If the polishing tool is not precisely aligned, the endface normal of the fiber will not be parallel to the fiber/ferrule axis (Figure 5.10, not drawn to scale). An nonzero fiber FCE further complicates this situation.

Thus, refraction of the entering/emerging light will cause the effective light axis to be nonparallel, or tilted, with respect to the alignment axis.

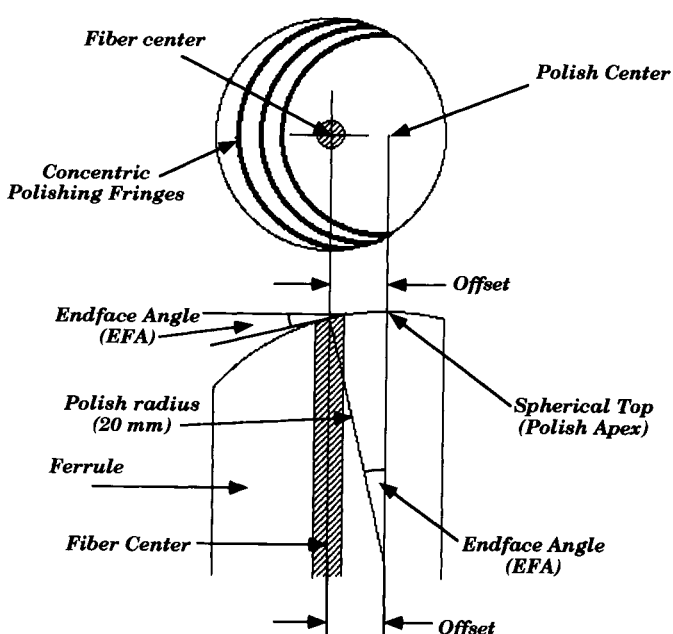


FIGURE 5.10 Fiber polishing geometry.

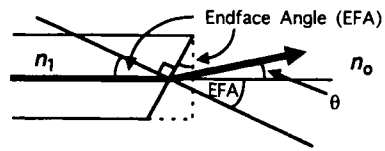


FIGURE 5.11 Fiber endface angle (EFA).

Depending on how stringent the manufacturer's constraints on longitudinal offsets are, this refraction can produce a significant reduction in coupling efficiency. As seen in Figure 5.11, a beam emerging from an obliquely polished fiber is refracted according to Snell's law:

$$n_1 \sin(\text{EFA}) = n_0 \sin(\text{EFA} + \theta) \tag{5.1}$$

It can be easily seen that when a fiber endface is polished at an angle EFA, the beam emerges at an angle θ with respect to the fiber axis (which is an angle EFA + θ with respect to the endface normal). Using the small angle approximation, θ can be found:

$$n_1(\text{EFA}) \approx n_0(\text{EFA} + \theta)$$

or

$$\theta \approx \left(\frac{n_1}{n_0} - 1\right)\text{EFA} \tag{5.2}$$

Using 1.47 for the fiber's index of refraction and a value of 1 for air, the relation between θ and EFA is $\theta \approx (0.47)$ EFA. Note that for analytical purposes, θ can be seen as the fiber analogue to the TOSA PA.

Random sampling of nonrejected spherically polished fibers shows up to 0.3 deg of effective tilt, θ. The measurement of the EFA can typically be made using interferometric techniques so that samples with, say, < 0.1 deg EFA can be rejected; however, this will produce large numbers of unused cables and increase manufacturing costs.

5.3.3 Elements 7–10: Connector-to-Module Mating Elements

Elements 7 and 8 are the *subassembly misalignment* (SAM) and connector *ferrule float* (FF), respectively, both measured in millimeters. Subassembly misalignment is the axial misalignment of the entire OSA/bore assembly from its desired true position within the transceiver module. Remember that we are discussing duplex interconnections, so that both the transmitting and receiving bore/ferrule matings must be aligned (see Fig. 5.12). If one of the couples cannot be mated, then by design, the other pair would be prevented from achieving successful plugging.

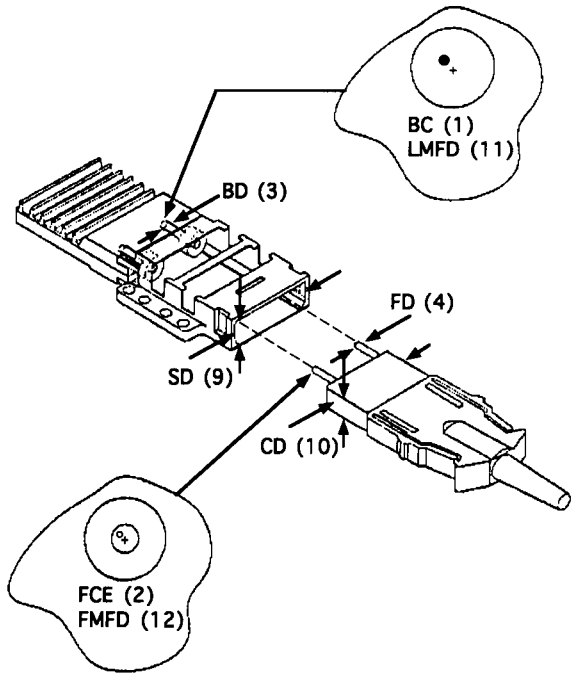


FIGURE 5.12 Connector/module mating elements.

Ferrule float is the amount of lateral movement or float possessed by the connector ferrules. Once a fiber has been ferruled (see Ref. 21 for ferruling procedures), it must be fully connectorized using a mechanical housing to ease pluggability. Too rigid a setting would impede ease of use and/or would make component damage more likely. Thus, spring forces built into the design allow a little "give." The most obvious "give" is that of longitudinal compression, but there is some leeway for radial flexing as well. This can be seen (simplified) in Figure 5.13.

If there is enough FF, a slight SAM could be compensated for. Of course, too large a FF could also prevent proper mating because of the "slopiness" of the ferrule position. These elements are important to ensure that the connector will physically plug into the module. If the SAM is too great to be accomodated by the FF, plugging may be impossible in solid bore OSAs. In split-sleeve OSAs, excessive SAM may cause a large off-axis force on the split sleeves that can cause PR problems or damage the split sleeve. It is important to note, however, that these two elements do not strongly affect the PR or XPR for solid bore systems, as the bore stiffness overwhelms them. A detailed account of the forces involving FF and SAM is given in Reference 25.

Elements 9 and 10 are the module *shroud dimension* (SD) and the *connector body dimension* (CD), respectively (also measured in millimeters). In a sense,

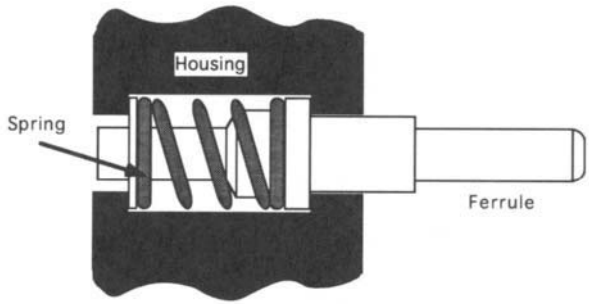


FIGURE 5.13 Ferrule spring forces.

the clearance between the two is the variable that affects PR and XPR. Excessive clearance can create several mechanically stable plug modes that can affect PR in the same module/connector pair. Data indicate that this clearance has a much stronger effect on split-sleeve OSAs than on solid bores. This observation is logical, as the clearance only follows the ferrule to "rattle" in the solid bore, whereas it may cause actual off-axis displacement in a split sleeve. These dimensions can be measured with calipers or a micrometer.

5.3.4 Elements 11 and 12: The Optical Elements

The last two elements are optical in nature. Element 11 is the *laser mode field diameter* (LMFD) and element 12 is the *fiber mode field diameter* (FMFD). Here, MFD is defined as twice the radial distance at which the field *amplitude* falls to $1/e$ of its center value. The laser MFD can be determined by deconvolving the power versus displacement curve obtained by scanning a fiber of known MFD across the spot, in the optical reference frame, while measuring the power launched into the fiber. Several techniques exist to measure MFD, the most common of which is the transverse-offset technique [2]. Even when all mechanical alignments are perfect, a difference in MFD between components will reduce the coupling efficiency. Although the MFD of fibers is well measured and controlled, the LMFD is usually poorly known or poorly controlled.

5.3.5 Examples

5.3.5.1 Example 1: A Split-sleeve TOSA Module

As an example of mating ferruled fibers to split-sleeve modules, we examined various tests with TOSAs possessing LMFD = 14 μm. The LMFD, however, is a difficult measurement and it can only be determined to within ± 2 μm. Figure 5.14a shows an example of the PR data for seven duplex cables and eight split-sleeve TOSAs in which the test was repeated for statistical confidence. Eighteen percent of the module/cable matings exhibited PR above 2 dB, with a disturbingly long tail above 5 dB. Figure 5.14b

shows the individual PR data for one module with seven cables. Each point is a measure of the power launched into a cable in one mating. Ten plugs were performed for each cable. Plug repeatability is measured by taking the maximum and subtracting the minimum for each cable. Thus, each data point in Figure 5.14a is a result of one cable "cluster" corresponding to Figure 5.14b. The XPR is calculated by subtracting the minimum from the maximum for all of the cables. In this example, the maximum PR is 3.5 dB for one cable, and the XPR is 5.5 dB. This graph indicates that PR is poor. (The data of Fig 5.14 was interpreted from data collected by IBM's Optoelectronic Enterprise in Endicott, New York.)

This work was done before the development of the beam center ratio test (section 5.4); therefore, the effect of BC was not assumed. However, if the BC was poor (large), we would have expected some especially low readings when the FCE was opposite the BC. Instead, the power launched consistently met specifications. The FCE was measured indirectly with an insertion loss technique that ensures a FCE of < 1.6 μm, and as stated before, there are data to suggest that it is typically 0.7 μm.

Split sleeves have the advantage that the bore/ferrule clearance variable is negated by design: they dilate to accommodate a range of ferrule diameters. Hence, elements 3 and 4 are of little consequence. However, elements 7 through 10 are of considerable concern as they can add variable off-axis forces to the split-sleeve, thus moving the ferrule in differing amounts from mating to mating. Careful observation during each plugging may show that the clearance between the connector and the shroud is excessive (elements 9 and 10). This was indeed the case here. This situation allows several mechanically stable modes of plugging (see Fig. 5.14c). These different modes of plugging produce variable off-axis forces that move the ferrules in the split-sleeve. These forces can only be created, however, if the FF of the connector (element 8) is exceeded by this clearance problem.

The use of shims can reduce the clearance between the connector and the shroud and this problem can be resolved. By inserting small wedges or shims in the spaces created by the SD and CD, the connection can be guided into a consistent position among the different matings of Figure 5.14c.

Repeating the tests using shims to confine the connector within the module shroud provides striking improvement as noted in Figure 5.14d. These are well within specifications. This suggests that the LMPD of 14 ± 2 μm promotes acceptable PR and XPR. However, even with shims, pulling on the cable can result in significant changes in coupled power.

Note that no consideration was made specifically to include PA and EFA in this discussion. The small range of EFA (< 0.2 deg) should not affect the XPR significantly. However, the PA could be expected to be a significant source of XPR because of the large variability in PA between modules. If a vendor were able to reduce the possible range of laser PAs, it is expected that the XPR would decrease accordingly.

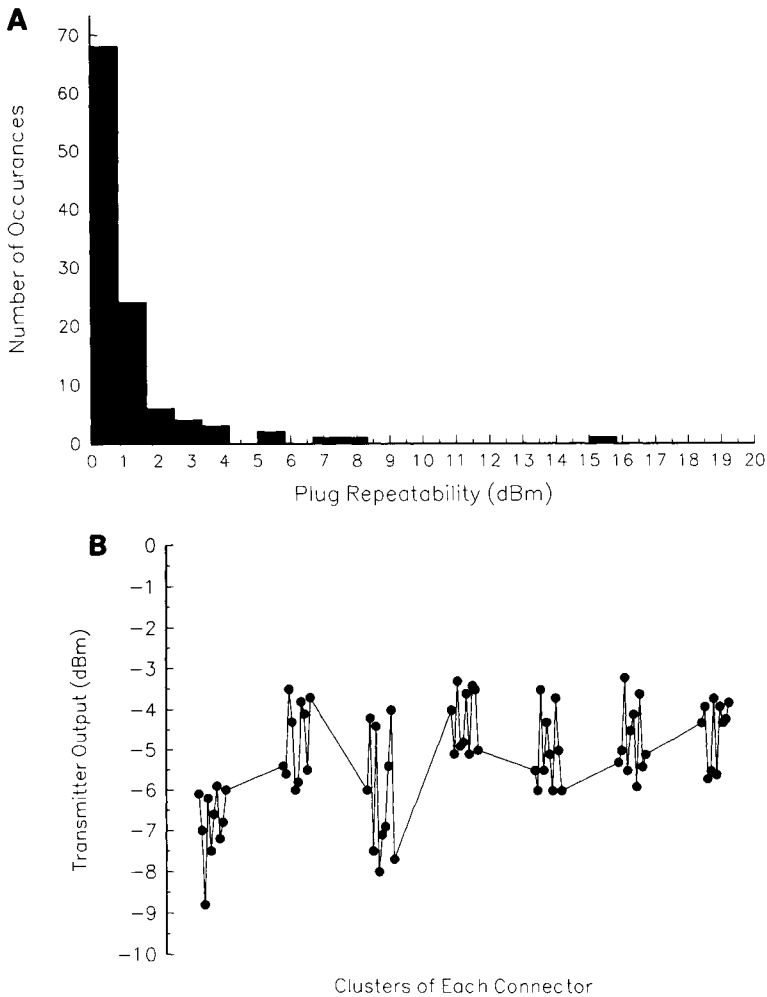


FIGURE 5.14 Split-sleeve bore coupling example.

As a result of the preceding considerations, split-sleeve OSAs are not recommended for SM, duplex connector/module systems.

5.3.5.2 A Solid Bore TOSA System

When considering solid bore OSAs, elements 1 through 4 are clearly very important, whereas because of the nearly infinite stiffness of the solid bore, elements 7 through 10 are only of secondary importance. As in example 1, the FMFD (element 12) is well controlled in a sample of duplex cables, and the LMFD (element 11) is specified on the solid bore TOSAs at approximately 14 μm ± 2 μm. Figure 5.15a is the PR data for seven modules and eight cables

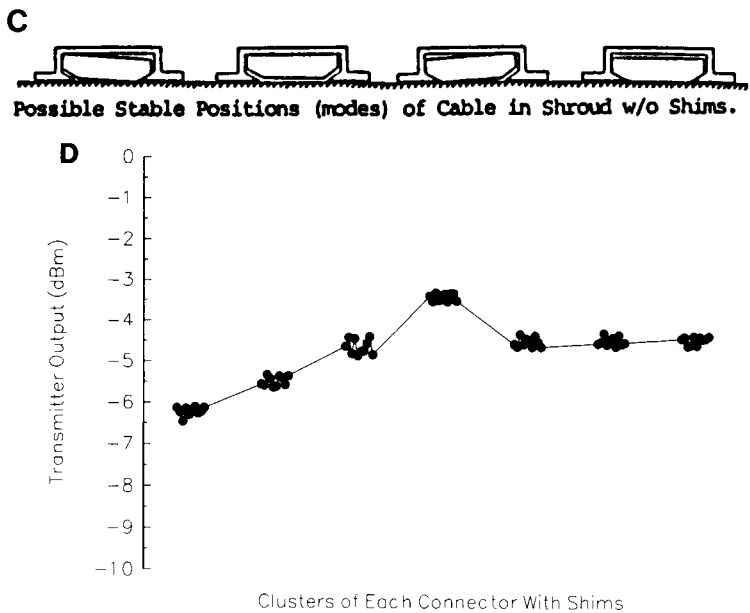


FIGURE 5.14 *Continued.*

using these TOSAs. In this experiment, five pluggings were performed with each connector. These data indicate that although 3 out of 56 module/cable combinations failed the 2-dB criteria, the lack of a tail in the data makes solid bores appear more promising than those with split-sleeve TOSAs. Figure 5.15b shows individual power-launched readings similar to Figure 5.14b. As in the last example, we also show data for when shims were used to minimze shroud/connector clearance. Figure 5.15c gives the PR using shims. With shims, the PR data do not improve as dramatically as in the split-sleeve case and therefore there is less incentive to use them: they would simply drive up module cost.

Beam centrality, as stated before, was not measured on these TOSAs. However, the results in our seven-module, eight-cable test resulted in a relatively consistent power launch and reasonably good PR and XPR. Pointing angles may have been largely responsible for much of the XPR for this experiment, as in the discussion for split-sleeve bores.

5.4 Metrology: Measuring the Coupled Power Range

Once the elements that effect the CPR have been identified, they must be measured to within desired tolerances. As explained in section 5.3, it would

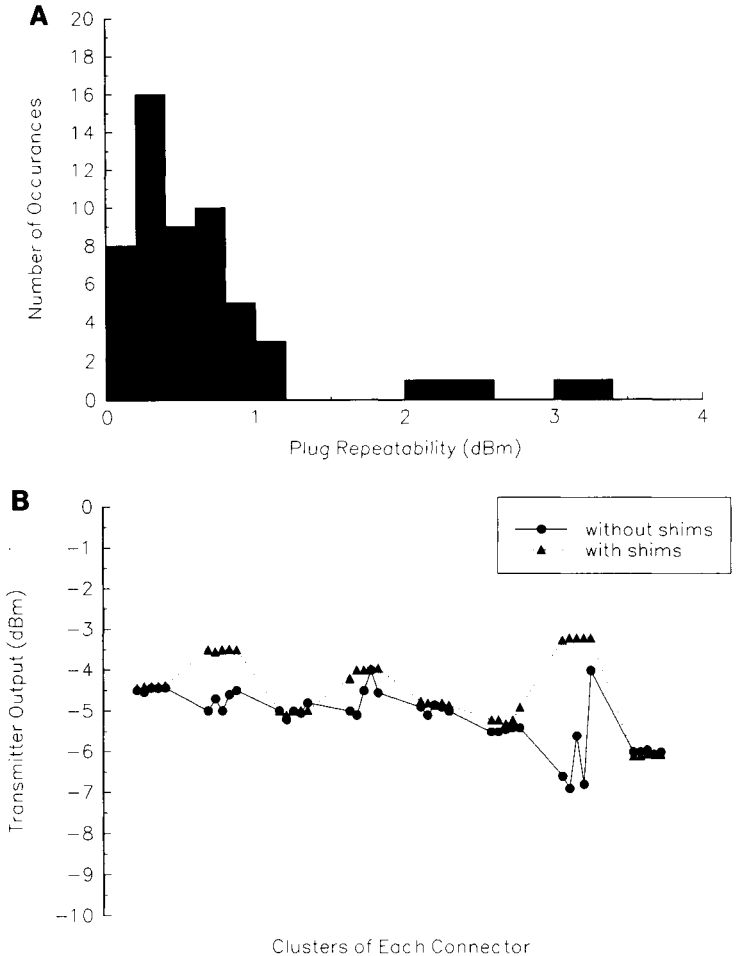


FIGURE 5.15 Solid bore coupling example.

be inefficient to measure every manufactured part, and so parameters are evaluated within some confidence interval. For reference, Table 5.1 shows typical values for the twelve elements. The ✓-marks are determined by the influence of a parameter's *variability*. If a parameter either has little variation, or has little effect on the coupling efficiency to begin with, it will be ranked with fewer ✓. Given these measurements, it is desirable to predict how much variation to expect from a given product, which could include any combination of values.

A typical CPR might be from −3 dBm to −9 dBm—a 6 dBm range, barring safety considerations. All of the variables that affect the CPR are

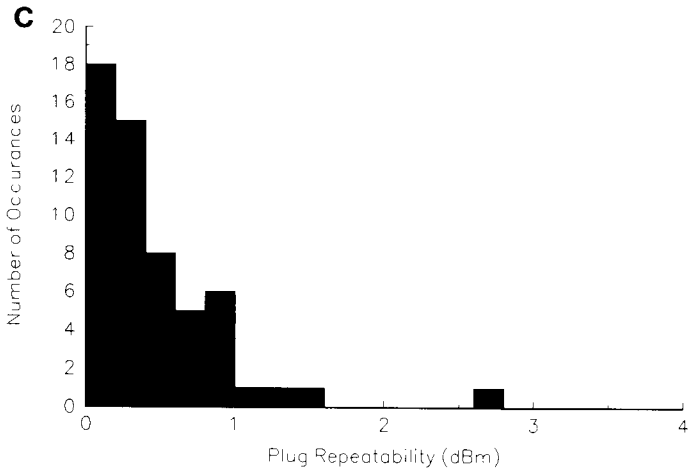


FIGURE 5.15 *Continued.*

usually analyzed with a Monte Carlo technique to ensure that the CPR can be achieved (section 5.5 will explore Monte Carlo simulations). The CPR can be viewed as the sum of variations resulting from aging effects on the electrical components and from changes in coupling from mechanical/optical variations in the connectors and TOSAs. The latter coupling variations are, of course, referred to as the XPR, which is defined as the range of coupled power from a large number of mating of many different connectors into one transceiver module, assuming no electrical variations.

TABLE 5.1 Typical values of the twelve elements

Element	Typical Value	CPR Influence	
		Solid Bore	Split-Sleeve
BC	(3.3 ± 2.0) μm	✓ ✓	✓ ✓
FCE	< 1.6 μm (0.7 μm typ.)	✓	✓
BD	(2.5010–2.5025) mm	✓	
FD	(2.4992–2.5000) mm	✓	
PA	< 3 deg	✓ ✓	✓ ✓
EFA	0.1–0.2 deg (barring rejection)	✓	✓
SAM	Spec. ± 0.34 mm		✓
FF	Spec. ± 0.34 mm		✓
SD	Spec. ± 0.125 mm		✓
CD	Spec. ± 0.125 mm		✓
LMFD	Variable	✓ ✓	✓ ✓
FMFD	(9.5 ± 0.5) μm typ. for SM	✓	✓

In general, all twelve elements affect XPR. However, if we limit ourselves to solid bore TOSAs, only eight variables are operable. Four variables are related to the connector and four to the module. These variables are

Module	Connector
BD	FD
BC	FCE
PA	EFA
LMFD	FMFD

Each of these factors will produce variablility in both PR (same connector/module pair) and XPR (multiple connectors/one module).

The actual distance between the fiber core and the laser spot is essentially determined by BC, FCE, BD, and FD, as shown in Figure 5.16 (cf. Figs. 5.6 and 5.8). Knowing these quantities, along with the directions of displacement ψ and ϕ, we can develop a relation for the distance R:

$$R^2 = [\mathrm{BC} - \mathrm{FCE} \times \cos\psi - \tfrac{1}{2}(\mathrm{BD} - \mathrm{FD}) \times \cos\phi]^2 + [\mathrm{FCE} \times \sin\psi + \tfrac{1}{2}(\mathrm{BD} - \mathrm{FD}) \times \sin\phi]^2 \tag{5.3}$$

It is clear that the maximum value for R occurs when ψ and ϕ are 180 deg and the minimum occurs for both directions equal to zero. Thus,

$$R_{\max} = \mathrm{BC} + \mathrm{FCE} + \tfrac{1}{2}(\mathrm{BD} - \mathrm{FD})$$

and

$$R_{\min} = \mathrm{BC} - \mathrm{FCE} - \tfrac{1}{2}(\mathrm{BD} - \mathrm{FD}) \tag{5.4}$$

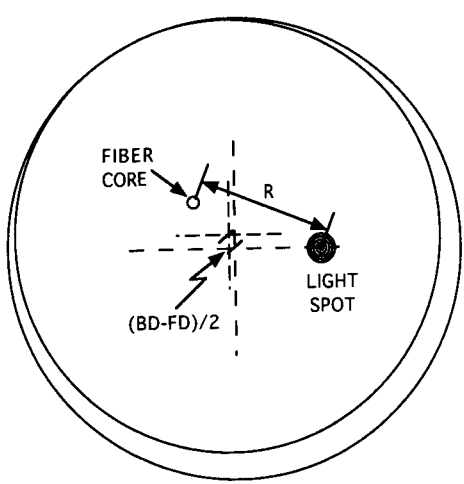


FIGURE 5.16 Fiber core and laser spot separation.

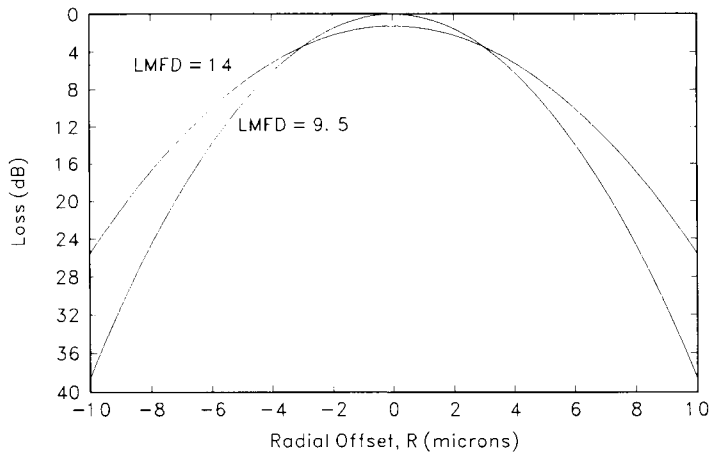


FIGURE 5.17 Coupling loss as a function of radial offset.

Clearly, BC, FCE, and BD − FD can give rise to an R_{min} of zero. Figure 5.17 shows the theoretical results for coupling from LDs with LMFD = 9.5 μm and 14 μm into a standard fiber with FMFD = 9.5 μm, for purely radial offset R. Unfortunately, actual coupling losses are about 3 dB greater than those shown in Figure 5.17, but the shapes of the experimental curves are surprisingly close to Figure 5.17. It is believed that the difference is caused by lens aberrations in the TOSA [32].

As mentioned, longitudinal offset plays a much smaller role in determining variable coupling loss. This is demonstrated in Figure 5.18. The FMFD is

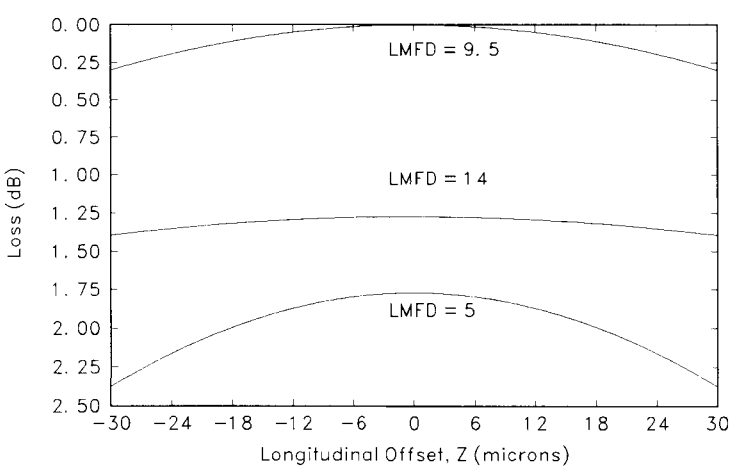


FIGURE 5.18 Coupling loss as a function of longitudinal offset.

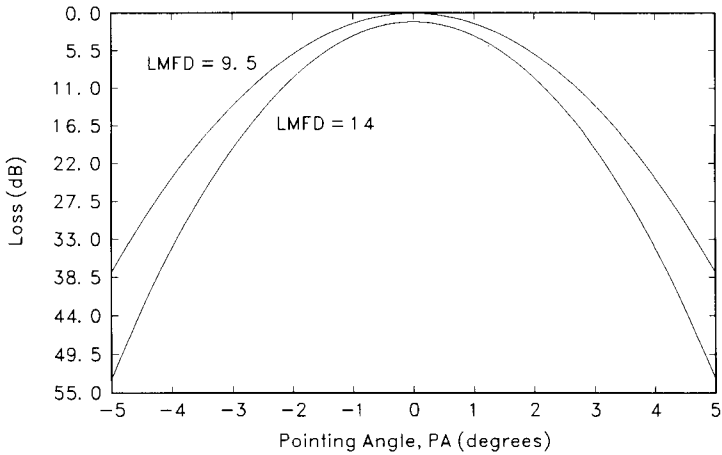


FIGURE 5.19 Coupling loss as a function of laser pointing angle.

taken as 9.5 μm. The loss is shown to change very little for several values of LMFD; thus it will not be included in further analysis. The references at the end of this chapter will give the reader any additional information desired [7, 8, 18, 21, 22].

Different modules will necessarily have different PAs, even if BC is well controlled (see description of element 5). Figure 5.19 shows the effect on coupling loss these various PAs will create. Generally speaking, this variation can be responsible for up to 10-dB differences in loss between matings of different modules with a given connector. However, PR will not likely be affected, as the PA for a given module can be assumed to be constant.

The fiber EFA has a much less severe influence on coupling efficiency, when it is assumed that other misalignments are small. In fact, as it is common that EFA < 0.2 deg, one can use Figure 5.19 to estimate its effect on overall coupling loss, provided $PA = \theta = 0.47$ (EFA) from Eq. (5.2). The variability comes from different insertions of one connector/module pair. With 360 deg of possible rotation within the bore, a fiber with an EFA = 0.2 deg will provide minimal coupling variation. Figure 5.20 shows this effect for small $R = 1$ μm, a realistic longitudal offset of 10 μm, LMFD = FMFD = 9.5 μm, and a "worst case" PA of 3 deg. The fluctuation is barely evident. The EFA has much more influence for larger disparities between MFDs [7]. Nonetheless, it will not be included in further discussion of CPR.

It should be noted, however, that many applications call for *intentionally* angled end faces to reduce back-reflection into the source. The industry standard for the EFA in this case is 8 deg. Here, for the same parameter values as in the preceding paragraph, different insertion orientations have dramatic effects, as shown in Figure 5.21 [8]. We believe that these situations are

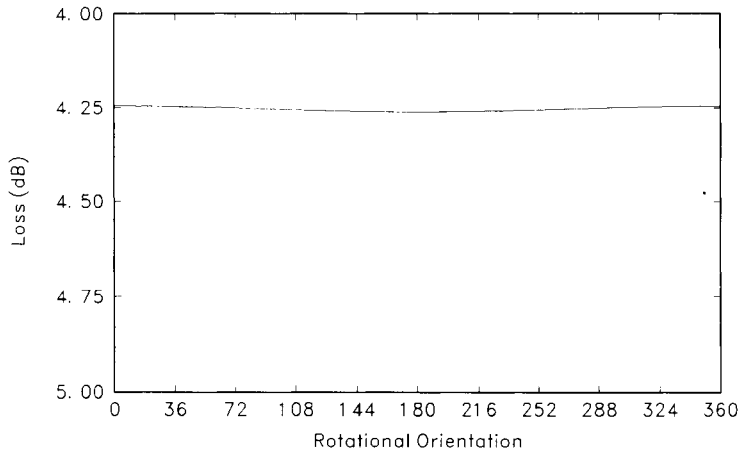


FIGURE 5.20 Coupling loss as a function of fiber end-face angle for closely matched mode field diameters.

unique and we only wish to alert the reader to the potential difficulties such fibers may present.

It is well worth noting that in our discussions, we have primarily considered *isolated* misalignments: radial, axial, and angular. When *all* of these are present, they can each become more significant to coupling. When the radial offset is significant, the PA has a more dramatic effect and vice versa. Nemoto and Makimoto give analytic solutions for when the three misalignment

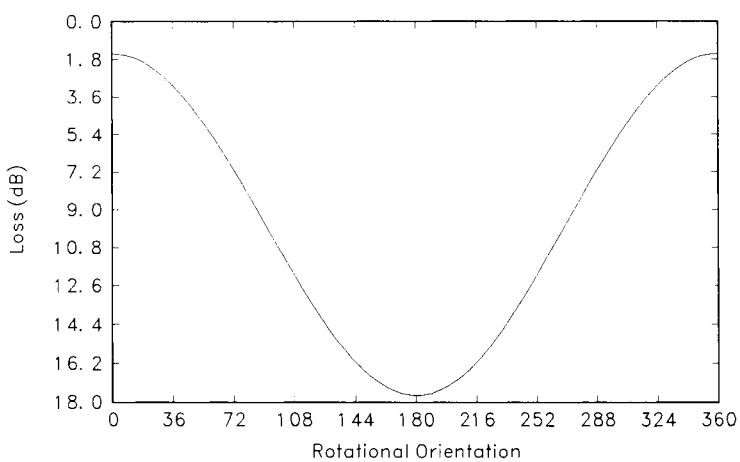


FIGURE 5.21 Change in coupling loss during fiber rotation for a large end-face angle.

"categories" can be assumed to occur in the same plane [22]. Out-of-plane misalignments are explored in References 7 and 8.

5.4.1 Theory of the BCR Test

Measurement of the BC has proven to be a formidable task. Although work could and should progress in this endeavor, it is possible instead to develop a product test such that the known parameters are used to ensure a desired CPR without direct measurement of the BC. Using worst-case scenarios, a test called the *beam center ratio* (BCR) test has been implemented. This functional test was developed to ensure the CPR in a cost-effective manner. Other approaches, such as measuring the elements and the BC, would be prohibitively expensive and would result in a *calculation* of the coupled power and not in an actual measurement.

The random mating of cables and modules will produce combinations at the extremes of the tolerances of the twelve elements. For this analysis, we will only be concerned with solid bore modules. For this reason and for those reasons just stated, the elements to be considered are BC, FCE, BD, FD, LMFD, and FMFD. (The laser PA only reduces the *maximum* coupling efficiency and will not change during the BCR test because the same module is used throughout.) A test was desired that would predict how each module would perform with any and all cables (that met the cable specifications). The BCR test was developed as a result of this need and is crucial to ensure CPR in SM systems.

We begin with a physical description of the process. A selected cable with a large FCE = 1.5 μm, FD = 2.500 μm, and FMFD = 10 μm is inserted into the TOSA. A small off-axis force is applied to the ferrule in the opposite direction of the BC offset to force the fiber core to be maximum distance from the center of the bore. While maintaining this force, the fiber is rotated through 360 deg. This procedure is shown in Figure 5.22. Because the beam will seldom be at BC = 0, the coupled power will vary as the fiber is rotated. A Monte Carlo analysis can show that the maximum to minimum beam-to-fiber core distances in this test should far exceed the 3-sigma limit of random matings in the field. The maximum and minimum coupled power readings are recorded and pooled with the maximum coupled power in a separate coupling efficiency test conducted with no external forces. This test is performed with an undersized ferrule. Perfect alignment is achieved with optical power feedback to a robot that aligns the ferrule with a laser spot. This test is performed to determine the "absolute" maximum coupled power. (The "control" test consisted of recording the highest coupling efficiency for a "normal" plugging, as would typically occur in a real product.) The BCR is defined as the maximum minus the minimum power from these three readings. The current limit is 4.5 dB for BCR with the preceding specified cable.

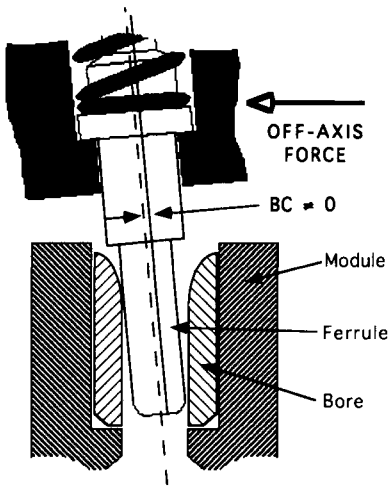


FIGURE 5.22 The BCR tester.

The Nemoto/Makimoto solution for light coupling can be used to determine the BCR as a function of the six elements [22]. For a source at wavelength 1.3 μm, the resulting equation for BCR is

$$\text{BCR (dB)} = \frac{8 \cdot 20 \cdot \log(e)}{\text{LMFD}^2 + \text{FMFD}^2} [R_{\max}^2 - R_{\min}^2] \tag{5.5}$$

Using Eq. (5.4), $R_{\max}^2 - R_{\min}^2$ can be reduced to give

$$\text{BCR (dB)} = \frac{8 \cdot 20 \cdot \log(e) \cdot 2\text{BC} \cdot [\text{FCE} + \frac{1}{2}(\text{BD} - \text{FD})]}{\text{LMFD}^2 + \text{FMFD}^2} \tag{5.6}$$

One must bear in mind the differences between the "spot size" of Reference 21 and the MFDs used here: MFD = $(2\sqrt{2})$ *spot size*. Hence the factor of 8 in Eq. (5.6).

5.4.2 Test Results

By using Eq. (5.6), one can define a TOSA/cable combination that gives a 4.5 BCR. The cable parameters are FD = 2.5000 mm, FCE = 1.5 μm, and FMFD = 10 μm, while in this example the TOSA has a BD = 2.5025 mm, BC = 2.5 μm, and a LMFD = 11.5 μm. The intent of the BCR test, to ensure statistically a 3-dB XPR with a 4.5-dB BCR, can be verified with a Monte Carlo analysis (see section 5.5). In such a simulation, this worst-caseTOSA is "mated" mathematically with a large number of cables that are randomly

selected, each having its FD, FCE, and FMFD within the specification range as given in Table 5.1. Monte Carlo analysis indicates that a BCR of 4.5 dB results in an XPR of slightly >3 dB at a 3-sigma limit. Additional Monte Carlo simulations indicate that to have consistently an XPR <3.0 dB, the BCR must be <4.0 dB. Unfortunately, it has been difficult to obtain reasonable quantities of TOSAs with BCRs below 4.0 dB because of difficulties in manufacturing and in the BCR test correlation and repeatability. This lack of correlation is the result of small errors in measuring the FD, FCE, and FMFD of the BCR test cable. These errors can easily lead to results that differ by more than 0.5 dB between two testers. In addition to these measurement errors, it is generally not possible to obtain a BCR test cable as described here in practice; hence, one must use a correction factor. As an example, let us assume that our test cable has the following characteristics: FD = 2.499 mm, FCE = 1.4 μm, and FMFD = 10.2 μm. Using Eq. 5.6 and the imaginary TOSA (with a BCR = 4.5 dB), as previously described, one obtains a BCR of 5.14 dB instead of 4.5 dB. Thus the cable has a correction factor of 4.5/5.14 = 0.875. Therefore any readings with this cable should be multiplied by 0.875.

As an example of BCR testing and XPR correlation, 16 modules and cables were tested in a cross-plug experiment. In this experiment, all cables are mated with all modules; hence, there are 256 combinations (16 × 16). The BCRs of the TOSAs in the modules are graphed in Figure 5.23. The BCR data indicate that 2 of the 16 TOSAs have a BCR of 4 dB. This percentage is quite high for a small population and suggests that the 3-sigma, 3-dB XPR criteria may not be met by the sample. Figure 5.24 is a Weibull probability plot of the XPR for each module with the cables and it is apparant that the tail of the 3-sigma limit (99.73%) is >3 dB.

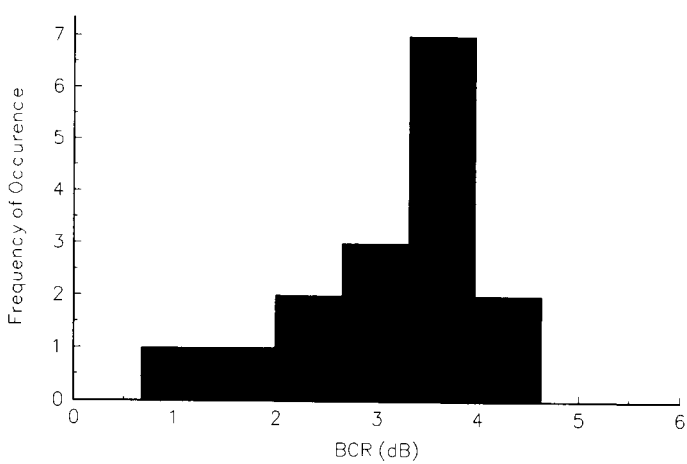


FIGURE 5.23 The BCR test results.

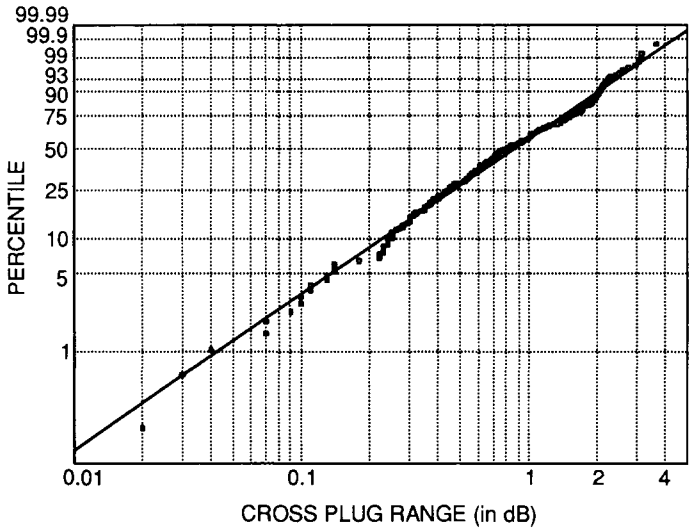


FIGURE 5.24 Weibull probability plot of XPR results.

In an ideal world, one would prefer to have all TOSAs with a BCR < 4.0 dB to achieve the acceptable XPR. Unfortunately, with today's technology, this goal is not economically reasonable. One must then decide what is an acceptable distribution of BCRs in population of TOSAs. Qualitatively, it is easy to imagine that an occasional TOSA with BCR > 4.0 dB can be tolerated and still result in the average XPR of < 3.0 if many of the TOSAs have low BCRs, say, < 2.0 dB. To determine the acceptable BCR distribution, a Monte Carlo analysis was performed with TOSAs defined from Eq. 5.6 with BCRs of 2.5, 3.0, 3.5, 4.0, and 4.5 dB. The total number of matings (out of 10,000) with an XPR that exceeded 3.0 dB was recorded at each BCR value. these matings were with random cables within the specification limit. The 3-sigma limit distribution criteria for 3-dB XPR must then be

$$\frac{\Sigma_i N_1 W_1}{\Sigma_i N_1} < 27 \tag{5.7}$$

where N_i is the number of TOSAs with a BCR of BCR_i and W_i is the number of readings above 3.0 dB XPR at a given BCR_i. The pertinent values are listed in Table 5.2 The Monte Carlo analysis of those data indicates that 46.9/10 000 (0.469%) of the mating should exceed 3 dB. Figure 5.24 shows that in the actual product testing, approximately 2.0% of the cross plugs exceeded 3.0 dB. When one considers the difficulty in accurately measuring FD, FCE, and FMFD on the BCR test cable, this level of agreement between data and theory is reasonable. However, to assure an acceptable XPR one may want to scrap some of the higher level BCR TOSAs from a population such

TABLE 5.2 Pertinent values of cross-plug test results

BCR_i	W_i
>4.5	334
4.0–4.5	100
3.5–4.0	35
3.0–3.5	12
2.5–3.0	5
<2.5	1

as this one. These sources of BCR error will be discussed in the next section.

5.4.3 Sources of Error in BCR Testing

It would appear that our concern should be the error in BCR between the "true" value and the measured value. However, in a practical sense the tolerances are too small and the testers too variable to obtain the true value. Thus, we are actually concerned with the difference between sets of data from different testers. If the differences between two data sets are small, we assume that both tests are close to reality. One would like the BCR test to be consistent; a check of this consistency is to have different testers check a number of TOSAs. So our concern and comparison for sources of error will be between two or more possible testers. It must also be remembered that all the TOSAs tested should be the same.

Since the specification for BCR is 4.5 dB, sources of error in BCR testing will be most critical near this value. One does not want to reject a 4.0 dB TOSA nor accept one that is 5.0 dB. We now refer back to Figure 5.17. The position of maximum coupled power, P_{max}, is $R = 0$. When the lateral offset is a maximum, this will give the least amount of coupled power, P_{min}. The maximum R can be found using the maximum offsets: BC = 3.0 μm, FCE = 1.5 μm, and 0.6 (BD − FD) = 1.5 μm. The coefficient of 0.6 appears (rather than 0.5) as a result of the experimental conditions, particularly the tilt angle. Geometric arguments can show that this factor is just under 0.6, and it is rounded up. Even if each individual element was at its maximum, it is unlikely that they would all align themselves to add up completely constructively. These offsets can be added to get an approximate extreme of the BCR core-to-laser spot offset of R_{max} = 6.0 μm.

From Figure 5.17, it can be seen that the P_{min} value is much more sensitive to errors in R than is P_{max}. A 1-μm error in the R position for measuring P_{min} can result in a greater than 1.5 dB difference. The first obvious source of error

is in the metrology of the BCR test cable. We compared measurements for FD and FCE for eight BCR cables. Data for FD were taken by two vendors, *vendor 1* and *vendor 2*, whereas FCE was measured by *vendor 1* and *vendor 3*. Using a *paired t*-test, the average difference between vendor methods for measuring FD is 0.220 ± 0.046 μm, and for FCE is 0.176 ± 0.114 μm. The *paired t*-test statistic for the FD is 13.9, implying virtually zero probability for rejection of the hypothesis that the two FD measurement methods are equivalent. Similarly, the *t*-test statistic for FCE measurement is 4.4, giving a confidence level of about 99.9%. The combination of the two average deviations in the same cable could cause a total test error of 0.4 μm × 1.5 dB/μm = 0.6 dB in a TOSA that was near 4.5 dB in BCR.

The measurements of FD and BD are averages. Although the technology involved in making the ferrules and bores is impressive, the processes result in roundness and straightness errors in both FD and BD. These imperfections create additional sources of error. Two examples are illustrated in Figure 5.25. Example (a) illustrates a case in which the ovality of the bore and ferrule match to give an additional offset. Example (b) is the opposite case in which they cancel and this results in a reduced offset. Production data indicate that the roundness error ($r_{max} - r_{min}$) is typically 0.4 μm for BD and 0.3 μm for FD. In example (a), this situation creates an additional offset of 0.35 μm (i.e., [0.4 + 0.3]/2). In example (b), a reduced offset of the same magnitude results.

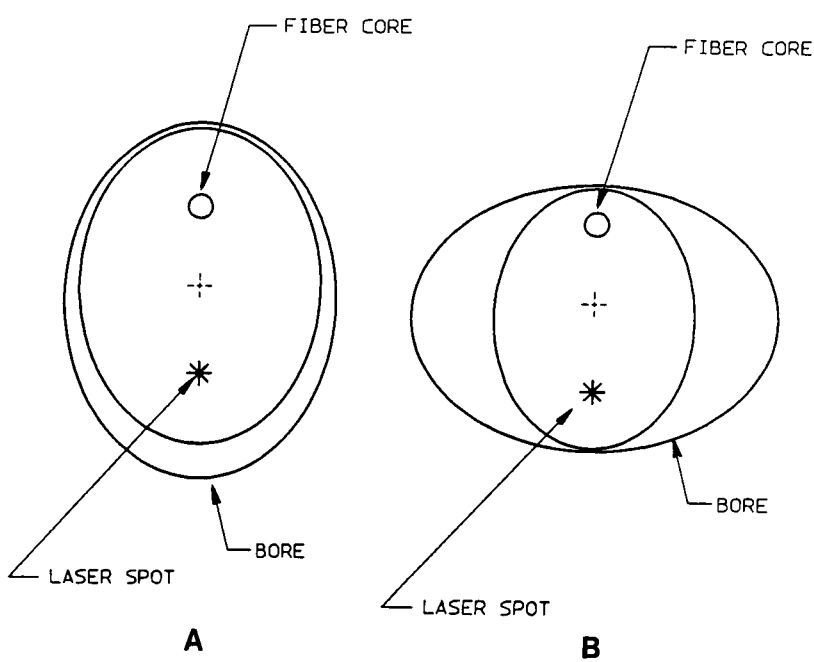


FIGURE 5.25 Ferrule and bore roundness errors.

In a near-4.5 dB TOSA, either of the two situations results in a 0.35 μm × 1.5 dB/μm = 0.525 dB difference from the mean. This error phenomenon is not likely to produce such an extreme deviation, as it requires the roundness errors to be aligned. A Monte Carlo analysis has shown that these errors in roundness produce an average of 0.2 dB in error in the readings.

Measurement errors in the FMFD can also affect BCR readings. Preliminary results suggest that at least a 0.3 μm difference can be expected between different testers. By using Eq. 5.6, this difference results in a 0.15-dB BCR measurement difference.

There are at least two other significant sources of error, both relating to the angular alignments. The PA is constant because we are using the same TOSA, but the angle of the ferrule surface (EFA) will vary among BCR test probes. Theory predicts that the coupling loss between a probe with no surface angle and one with a surface angle of only 0.25 deg (a typical maximum value) can be an additional 0.5 dB if the PA is 3 deg [8].

Another lossy mechanism is called the *etalon effect*. It is the result of light reflecting off the ferrule back into the laser, creating an extended laser cavity that results in a power distribution versus z, as seen in Figure 5.26. Based on resonant cavity analysis, the peak-to-null separation is $\lambda/2$. Because only a 1.3 μm/2 = 0.65 μm z-difference is needed to measure max-to-min in Figure 5.26, a variation of about 0.5 dB will be common.

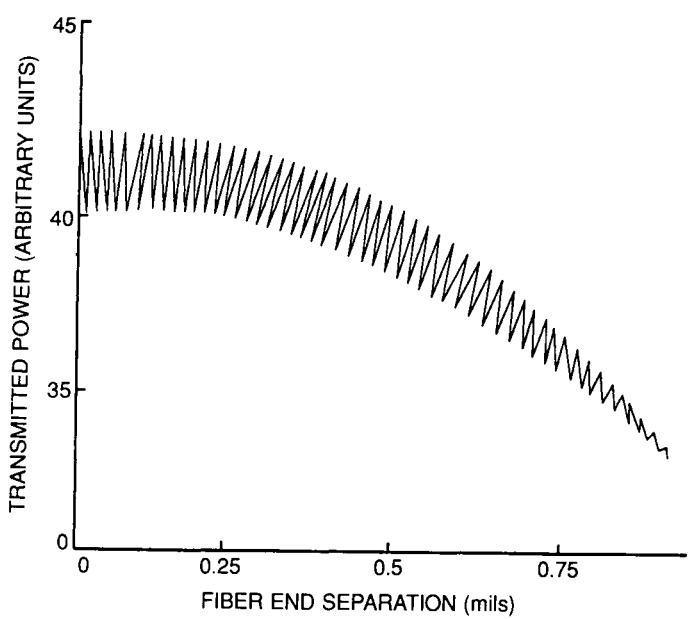


FIGURE 5.26 The etalon effect.

In summary, there are many sources of error, including many that are not considered here. Other fluctuations from temperature, humidity, electrical noise, and so on, are either small or will average out to produce no change between tests. However, the errors outlined in the previous paragraphs cannot be glossed over. The state of affairs is not quite as depressing as it may seem, as the errors seldom will exist together and will often cancel. Adding these errors in quadrature may give a reasonable estimate of the differences one might expect in comparing test data on the same TOSAs from different testers. We will assume that the errors are 3-sigma limits and can thus be added in quadrature. This approach results in $(0.6^2 + 0.525^2 + 0.2^2 + 0.15^2 + 0.5^2 + 0.5^2)^{1/2} = 1.09$ dB or essentially 1 dB. Thus, with today's metrology and manufacturing processes, one should expect about a 3-sigma limit of 1 dB variation between different test houses.

5.4.4 Experimental Verification

As just summarized, when all uncertainties associated with metrology are considered in creating the BCR standard, the net variance will be approximately 1 dB. An additional penalty of 0.5 dB should also be added for repeatability and drift. This brings the BCR limits to 4.5 ± 1.5 dB, which requires the TOSAs to be screened to 3 dB. To reduce this variability, a BCR standard connector has been established and is screened through cross plugging to maintain a ± 1.0-dB error.

To establish a consistency for BCR measurement, it is important to develop criteria for these test cables. We will refer to cables meeting these demands as the Golden Standard. To create the BCR Golden Standard cables, all cables are initially screened for spurious response and rejected. The remaining cables are tested in the cross-plugging fashion and the BCR value is measured for each cable. This criterion is used to handpick the most accurate and repeatable cables (it is assumed that the metrology readings are uniformly distributed). The BCR test cables are composed of either tungsten carbide or zirconia ferrules. The tungsten carbide BCR cables achieve their BC offset ferrules (those with BC $\neq 0$) by moving the ferrule insert (which holds the cable) to one side and soldering this insert to the ferrule. Several problems arise with these cables, such as the cable may be recessed within the ferrule, a "soft" area may often occur on the ferrule, and the FMFD may be distorted (noncircular). Their advantage is that they are durable and their manufacturing is well controlled. The zirconia ferruled cables, on the other hand, typically have more uniform FMFD and are more similar to in-the-field applications. However, these ferrules have poor roundness and wear quickly.

In a comparison of a pool of Golden Standard test cables, Figure 5.27 shows the BCR cross-plug results of eight BCR cables and eight TOSAs. In these tests, the standard deviation of the corrected BCR for each TOSA varies from

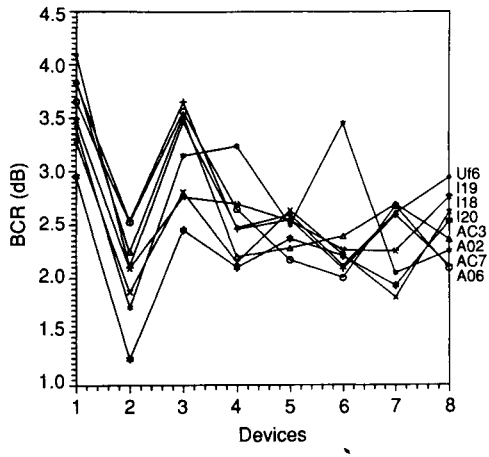


FIGURE 5.27 BCR cross-plug results for eight TOSA and eight cables.

0.164 to 0.463 dB. (See *Test Results* section for information on the "correction" factor.) As the repeatability for this measurement has been shown to be 0.3 dB, no cable should be rejected for a 0.164-dB deviation from the mean, but instead a 0.4-dB deviation should cause rejection. Using the criteria that a deviation from the mean must be less than 0.4 dB for a test cable to be acceptable, only cables AC3, I18, and I20 could be accepted in this evaluation. (Note that these data [Figs. 5.27–5.29] have been reproduced from experiments conducted at IBM's Optoelectronics Enterprise in Endicott, New York.)

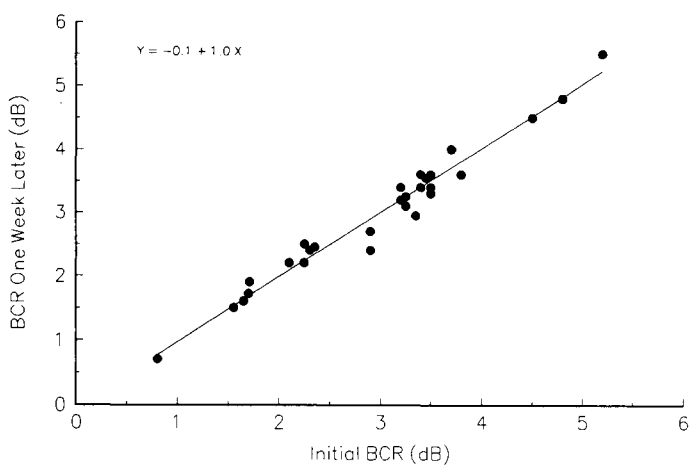


FIGURE 5.28 Long-term repeatability of 30 TOSAS.

Another source of error is tester short-term repeatability arising from the random perturbations associated with fiber rotation. This error is also affected by fiber wear, contaminants, and uniformity of the off-axis forces. Data show that the short-term repeatability gives a BCR of approximately 0.10 ± 0.08 dB.

The final source of error to be mentioned here is that of long-term repeatability and drifts. This has been an area of concern because it results in chaotic measurements. Contamination, fiber wear, pressure exerted from the TOSA bores/mounts, postproduction TOSA movement, ambient temperature, and humidity may all conribute to drifts. Cleaning procedures have been instituted, such as using the spectroscopic grade of 2-propanol along with physical inspection of the bore and lenses. Gradual fiber wear can be accounted for with controls and metrology. Excessive pressure on the TOSA mount is resolved by remounting with the minimal force. TOSA movements resulting from relaxation, excessive force from the mount, or from fiber binding to the bore is of concern. The long-term repeatability of 30 TOSAs is shown in Figure 5.28 with a correlation coefficient of 0.985. The average difference after 1 week is 0.2 dB. An example in Figure 5.29 shows vendor-to-vendor correlation for 30 parts with a correlation factor of 0.854.

The best BCR absolute accuracy today is ±1 dB: 0.5 dB for metrology, 0.3 dB for repeatability, and 0.2 dB for drifts. Nevertheless, it is the best means of determining the extreme cases of coupled power. Ceramic ferrules with better roundness and straightness could reduce some errors, as will producing testers with reduced applied force in their floating mechanisms. Finally, the BCR tester could be placed under a clean room hood. Given all of these variables, it will be difficult to improve the BCR test to maintain ±0.5 dB accuracy.

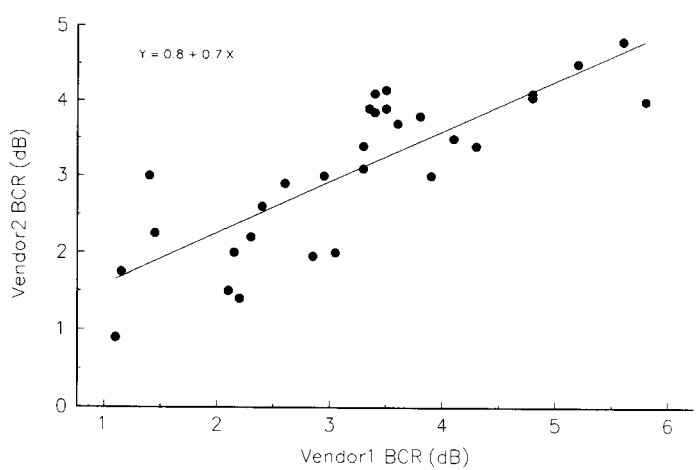


FIGURE 5.29 Vendor-to-vendor correlation.

5.5 Computer Simulations to Guarantee the Coupled Power Range

To guarantee that any manufacturing process performs to specifications, it is necessary to show statistically that the percentage of failures is acceptable. Although it is impossible to produce anything with absolutely zero errors, it is financially beneficial in both the short term and the long term to reduce the incidence of part failures. Knowing the percentage of total production that must be discarded/repaired provides both the manufacturer and the purchaser with the information needed to improve their respective products.

There are many ways to establish a reliability estimate, but all fall into two basic groups: random sampling of the finished product or computer simulations of the product, as generated using manufacturing tolerances. Although both are necessary in establishing client confidence, random sampling can become very costly, especially at the early stages of production before the kinks have been worked out of the manufacturing process. Therefore, computer simulations are typically performed first.

One method of simulating statistical variations in a process uses the Monte Carlo technique. Monte Carlo simulation randomly selects a set of independent parameters to estimate the probabilistic distribution (hence the reliability) of a dependent parameter. We will not go into detail regarding the theory of the Monte Carlo numerical methods (see Ref. 26 for details). For the case considered here, coupling from a TOSA into an optical fiber, the critical parameter is the coupled power ratio. If a certain CPR is exceeded, the expected data transmission cannot be achieved.

The CPR is dependent on the variability in cable parameters and on the BCR, as measured by the methods of section 5.4:

$$\text{CPR} = f(\text{field cable parameters, BCR})$$

The cable parameters are assumed to be controlled to within specifications: FCE $< 1.6\ \mu$m, FMFD $= 9.5\ \mu\text{m} \pm 0.5\ \mu$m, and FD $= 2.5010 - 2.5025$ mm. When the cable parameters satisfy these criteria, CPR can be determined using a Monte Carlo analysis that will only depend on the BCR. This BCR can be varied accordingly.

In this analysis, each selected TOSA is combined with 10 000 randomly selected "simulated" cables. Thus, if one knows the BCR for a given TOSA, the Monte Carlo simulation will generate 10 000 cables with random magnitudes/orientations of parameters and mate them with the selected TOSA, giving the CPR for each link.

When generating the CPR data, it is important to account accurately for all random variations that could occur. We will consider solid bore modules only and therefore only six elements of CPR as in the solid bore example in section 5.3.

As shown in section 5.4, the radial offset is a function of several parameters:

$$R^2 = [\mathrm{BC} - \mathrm{FCE} \times \cos\theta - 0.5(\mathrm{BD} - \mathrm{FD})\cos\phi]^2 + [\mathrm{FCE} \times \sin\theta + 0.5(\mathrm{BD} - \mathrm{FD}) \times \sin\phi]^2$$

From this equation, the maximum coupled power will occur when $R_{min} = 0$, whereas the minimum coupled power corresponds to $R_{max} = \mathrm{BC} + \mathrm{FCE} + (\mathrm{BD} - \mathrm{FD})/2$.

Assuming controlled cable elements of FCE, FD, and FMFD, the remaining three of the six elements, BC, BD, and LMFD correspond to the TOSA. As outlined in section 5.4, these are not separately measured but rather are combined in the BCR test. Thus, the minimum power coupled into the link can be specified statistically according to the parameter variation. However, as mentioned earlier, laser safety requires < -4.5 dBm of coupled light. Because -9 dBm is required to satisfy the link budget, this leaves a 4.5 CPR that must be guaranteed by hardware and Monte Carlo simulations.

5.6 Conclusions

This chapter has presented the reader with some new techniques with which to analyze losses from coupling light from laser diode modules into optical cables. The section on the twelve elements has introduced standardized parameters that can be used to evaluate the quality of a coupled system. These elements were utilized in section 5.4 where experimental testing was illustrated for the SM fiber case. Practical considerations were used to develop these methods.

In addition to the new metrology, this chapter has also introduced new terminology used in data communication. Much of this might be unfamiliar to those acquainted with only telecommunication. It is expected, however, that the concepts introduced here can be applied to any information-transmitting system that makes use of the distinct advantages of fiber-optic technology.

References

1. American National Standard for the Safe Use of Optical Fiber Communication Systems Utilizing Laser Diode and LED Sources (1988). American National Standards Institute, Z136.2.
2. W. T. Anderson and D. L. Philen (March 1983). Spot size measurements for single-mode cables—a comparison of four techniques. *IEEE Journal of Lightwave Technology* **LT-1**(1), 20–26.
3. N. R. Aulet, D. W. Boerstler, G. DeMario, F. D. Ferraiolo, C. E. Hayward, C. D. Heath, A. L. Huffman, W. R. Kelly, G. W. Peterson, and D. J. Stigliani, Jr (July 1992). IBM Enterprise Systems multimode fiber optic technology. *IBM J. Res. Develop.* **36**(4), 553–576.

4. D. B. Keck (1981). Optical fiber waveguides. In *Fundamentals of Optical Fiber Communications* (Ch. 1). M. K. Barnowski, ed. New York: Academic Press.
5. M. K. Barnowski (1981). Coupling components for optical fiber waveguides. In *Fundamentals of Optical Fiber Communications* (Ch. 3). M. K. Barnowski, ed. New York: Academic Press.
6. F. Bosch, G. L. Dybwad, and C. B. Swan (1980). Laser fiber-optic digital system performance improvements with superimposed microwave modulation. Paper TuDD7, *Conference on Lasers and Electro-optics.* San Diego, CA, February 26, 1980.
7. D. P. Clement, U. Österberg, and R. C. Lasky (1993). Analytical coupling model for obliquely polished single-mode cables. *IEEE Photon. Technol. Lett.* **5**, 1442–1444.
8. D. P. Clement and U. Österberg (1995). Laser diode to single-mode coupling using an "out-of-plane misalignment" model. *Opt. Engineering* **34**(1), 63–74.
9. S. Das, C. G. Englefield, and P. A. Goud (1984). Modal noise and distortion caused by a longitudinal gap between two multimode cables. *Appl. Opt.* **23**(7), 1110–1115.
10. R. E. Epworth (1978). The phenomenon of modal noise in analog and digital fiber systems (pp. 492–501). *Proceedings of the 14th European Conference on Optical Communications*, Genoa, Italy.
11. P. E. Green (1983). *Fiber Optic Networks.* Englewood Cliffs, NJ: Prentice-Hall.
12. M. C. Hudson (1974). Calculation of the maximum optical coupling efficiency in multimode optical waveguides. *Appl. Opt.* **13**(5), 1029–1033.
13. R. Huegli, and R. J. S. Bates (1991). Reduction of modal noise in high-speed short-distance computer data links by an electrical signal. *IEEE Journal of Lightwave Technology* **9**, 1788–1793.
14. International Electrotechnical Commission (IEC). International Standards 825-1 and 825-2, safety of laser products. Central Bureau of the IEC, Geneva, Switzerland.
15. K. Kawano, and O. Mitomi (1986). Coupling characteristics of laser diode to multimode fiber using separate lens methods. *Appl. Opt.* **15**(1), 136–141.
16. G. Keiser (1983). *Optical Fiber Communications* (Ch. 5). New York: McGraw-Hill.
17. H. Kogelnik (1964). Coupling and conversion coefficients for optical modes. *Proceedings of the Symposium on Quasi-Optics*, June 8–10, 1964. In Vol. XIV of *Microwave Research Institute Symposia Series.* New York: Polytechnic Press of the Polytechnic Institute of Brooklyn.
18. D. Marcuse (1977). Loss analysis of single-mode fiber splices. *Bell System Technical Journal* **56**(5), 703–718.
19. D. Marcuse (1966). Physical limitation on ray oscillation suppressors. *Bell System Technical Journal* **45**, 743–751.
20. D. Marcuse (1971). Compression of a bundle of light rays. *Appl. Opt.* **10**(3), 494–497.
21. C. M. Miller (1986). *Optical Fiber Splices and Connectors.* New York: Marcel Dekker.
22. S. Nemoto and T. Makimoto (1979). Analysis of splice loss in single-mode fibers using a Gaussian field approximation. *Opt. Quantum Electron.* **11**, 447–457.
23. E. G. Neumann (1988). *Single-Mode Cables.* Heidelberg: Springer-Verlag.
24. C. M. Olsen (1993). Superposition of a strong microwave signal on the laser injection current: Spectral characteristics and impact on modal noise. *IEEE Photon. Technol. Lett.* **5**, 13–16.
25. O. Paz, H. B. Schwartz, D. E. Smith, E. B. Flint (1992). Measurements and modeling of fiber optic connector plugability. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology* **15**(6), 983–991.
26. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C.* Cambridge University Press.
27. B. E. A. Saleh, and M. C. Teich (1991). *Fundamentals of Photonics.* New York: John Wiley & Sons.
28. A. W. Snyder and J. D. Love (1983). *Optical Waveguide Theory* (Part 1). London: Chapman & Hall.

29. J. Vanderwall and J. Blackburn (1979). Suppression of some artifacts of modal noise in fiber-optic systems. *Opt. Lett.* **4**, 295–296.
30. U.S. Department of Health and Human Services (1988). Regulations for the administration and enforcement of the Radiation and Control for Health and Safety Act of 1968. Publication FDA 88-8035, April 1988, U.S. Dept. of H&HS, Public Health Service, Food & Drug Administration Center for Devices & Radiological Health, Rockville, MD.
31. International Non-ionizing Radiation Committee of the International Radiation Protection Association (1985). Guidelines on limits of exposure to laser radiation of wavelengths between 180 nm and 1 mm. *Health Physics* **49**(2), pp. 341–359.
32. H. Karstensen, and K. Drögemuller (1990). Loss analysis of single-mode fiber couplers with glass spheres or silicon plano-convex lenses. *IEEE Journal of Lightwave Technology* **8**(5), 739–747.

This Page Intentionally Left Blank

Chapter 6

Data Processing Systems and Optoelectronics

Casimer DeCusatis
IBM Corporation
Poughkeepsie, New York

6.1 Data Processing System Overview

The role of fiber optics and optoelectronics in the data processing environment is closely related to the needs of the various computer system architectures available. Traditionally, technological advances have appeared first in the high-end, commercial mainframe markets and are later applied to smaller, personal computer (PC) systems. In this respect, fiber-optic data communications is unique in that high bandwidth optical data links have found applications in networks of personal computers and midrange processors even before their introduction into mainframe enterprises. With the increasing emphasis on connectivity, optical data links are being used in a wide range of applications across all of the major computer platforms. We begin this chapter with an overview of the major architectures in use today and a discussion of the role played by fiber optics in these systems. With this background, we

Optoelectronics for Data Communication

can proceed with a discussion of some of the basic problems common to most data communication systems, including link design, sources of noise and signal degradation, and link performance modeling. We then discuss the most commonly used fiber-optic data communication protocols and apply some of the basic principles to the design of various link physical layers. The chapter concludes with a description of emerging technologies and applications that are expected to become important for the next generation of digital data communications.

It is convenient to organize the various data processing environments into a few common platforms on the basis of processing power and functionality. At the low end, personal computers have been steadily increasing in power and versatility since their introduction, to the point where the distinction between a high-end PC and a low-end workstation is often unclear. By using modems to transfer data across public carrier telephone lines, which often contained a fiber-optic backbone, these machines were probably the first to indirectly use optical fiber for data transfer. This had minimal impact on data communication, however, because the high bandwidth telecommunication fiber was bottlenecked by much lower speed copper coax or twisted pair wiring at either end. As the commercial use of PCs and workstations grew, dedicated networks to share data became increasingly important. Most business communication takes place within a single building or in a relatively small area; the networks that were developed to manage this data became known as local area networks, or LANs. As the distance covered by these networks and their functionality increased, they expended to form the so-called wide area networks (WANs) and even campus-sized or metropolitan area networks (MANs). The distinction between LAN, WAN, and MAN systems is somewhat arbitrary; typically, a LAN handles communications within a building, a WAN extends to several buildings, and a MAN may cover a university campus-sized area or even larger area.

There are four basic network topologies that have emerged, the bus, star, ring, and star-wired ring as shown in Figure 1. Bus networks are essentially just an extension of the internal machine data bus to a serial link; all devices are attached to a main cable that is used as a common transmission medium for sharing data. The main cable is bidirectional and devices are attached with a special tap that does not require interruption of the cable to install a new device. Each device is assigned a unique address to ensure that it receives the intended data. Although the bus topology requires the least amount of cable, its length is limited by the use of a single continuous cable; even with amplifiers, most bus networks are less than 500 m long. More common is the star topology, in which each device is connected to a common central point where a hub is installed to provide interconnection paths between the devices. Each device has an independent communication pathway to and from the hub, so the star network can accommodate mixed transmission speeds and media (either

optical fiber, copper twisted pair, shielded twisted pair, or coaxial cable.) The hub can be either a crossbar switch or an intelligent processor; either way, cable congestion is a problem for large star networks. The second most common topology is a ring in which each device is connected only to adjacent devices; because each device must repeat the information it receives, much longer distances are possible. As before, each device or node on the ring is assigned a unique address and the media may be of mixed type and speed. Of course, the ring must be temporarily opened to facilitate addition of new devices. Finally, the star-wired ring combines features of both previous designs; the physical layout is a star configuration, but devices are wired in a ring topology. Instead of a central hub, devices are connected to a wiring concentrator by two unidirectional paths. The concentrator contains relays that connect the path of the upstream device to the next device, as in a ring topology. By using this

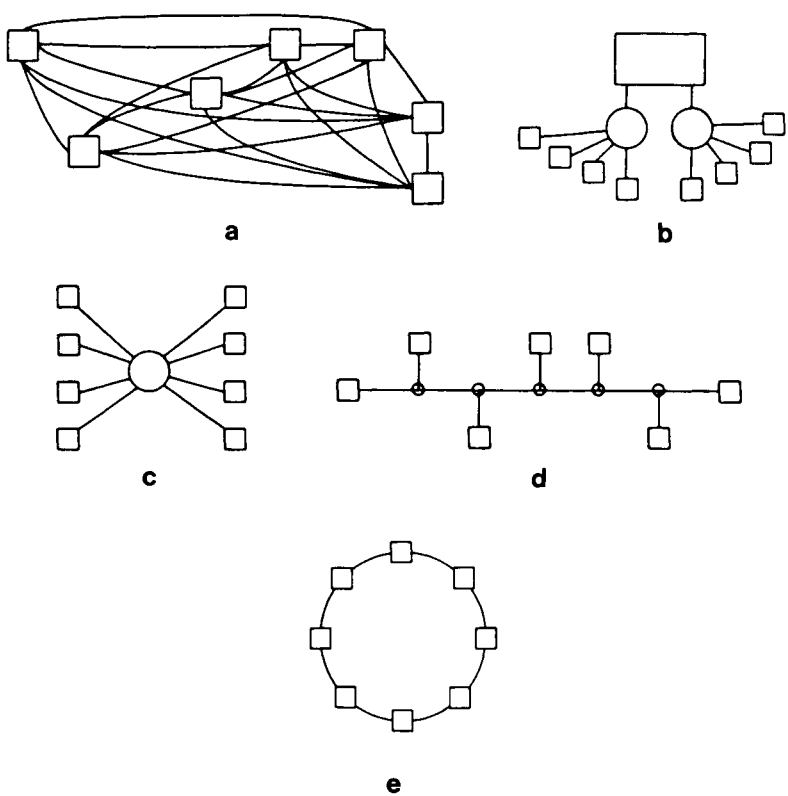


FIGURE 6.1 Different network topologies: (a) mesh, (b) tree, (c) star, (d) bus, (e) ring. Squares indicate nodes (workstations, terminals, etc.); circles indicate connection points (either active or passive) that may involve data routing functions.

design, it is possible to wire a building once for all possible future cable drops; concentrators can be connected together to provide for future network growth.

Similar types of LANs can be interconnected by switching devices called bridges or by more complex devices known as routers, which can create paths between individual nodes on different networks. Dissimilar networks can be connected through gateways that account for the different network protocols in use. As can be imagined, most networks in use today have evolved into a combination of different types of equipment and topologies interconnected on a variety of media. A complete description of LAN, WAN, and MAN interconnection requirements and strategies is well beyond the scope of this chapter, we will confine our attention to network protocols designed specifically for use with an optical fiber transmission media. However, there are some basic features common to many data communication systems that we should discuss. Throughout this chapter, we will refer to fiber-optic networks, channels, and fabrics in different contexts. Generally speaking, a data communication network serves many users and is software intensive, requiring much higher overhead to transfer information. In contrast, a data channel has much lower overhead and is very hardware intensive; for example, simple error correction and retransmission may occur in the link hardware without consulting the operating system. A loose interconnection between networks or channels with central control provided by a switching point forms a data communication fabric. Later in this chapter, we will discuss examples of local area networks (FDDI), fabrics (FCS), and channels (ESCON[1]); the distinction is growing somewhat arbitrary, as high-end data communication channels are merging with existing networks. The flow of data in a network is not evenly distributed; random bursts of data that have variable length are sent at irregular intervals. There are two main ways to manage data flow in a network, circuit switching and packet switching. In a *circuit switching* network, a central device or switch makes a direct, dedicated connection between two devices that lasts as long as necessary for the devices to exchange data. This approach is most useful when there are fewer devices in contention for the transmission media or when data integrity is critical; a circuit switch topology is used in the ESCON interconnection architecture for mainframe computers. An alternative is *packet switching*, in which data is packaged into electronic packets that contain the address of the sending and receiving devices, plus additional information for data routing and functions, such as error checking. Useful for bursty data and large networks of nodes, packet switching is the basis for the emerging asynchronous transfer mode (ATM) technology.

Most of the early LANs developed as peer-to-peer networks; each user on the network contributes some resources to the LAN and controls what

[1] ESCON is a trademark of the International Business Machines Corporation.

resources are made available. This type of decentralized control works best in smaller networks; initially, a single PC or workstation can be dedicated to act as a source for data shared by many users. For larger networks, a more rigid control structure known as client-server is required; here, resources are controlled by a network administrator rather than by the users. The administrator is not a LAN user, but instead controls security, centralized data backup, allocation of disk space to users from a central dedicated hard disk server, and performs other LAN functions. As networks have grown, mainframes have begun to play an increasing role in network administration; because of its large central memory, large number of I/O (input/output) connections, processing power, data integrity, and high data security, the mainframe is in many respects the ultimate server. Many users are beginning to recognize this, and high-end processors are being used as "super servers" on many networks. This reflects the changing role of the mainframe in today's environment. Previously, mainframes were used as large centralized processors, independent of outside networks and with a largely proprietary operating system. Because all of the processing hardware was contained in a central computer room (commonly known as a "glass house"), mainframes were slow to adopt the networking requirements of other lower-end systems.

The role of the mainframe has changed drastically in recent years, however, and in 1990, IBM became one of the first in the industry to adopt fiber-optic technology for I/O communications on the System/390[2] family of mainframe computers. This represented a vast improvement in data rate and distance over conventional copper channels; ESCON is capable of 17 Mbytes/s up to 20 km, whereas conventional channels were limited to about 6 Mbytes/s and under 200 m. This technology allowed for significant increases in performance and extended the mainframe into a distributed computing environment. We will describe the ESCON channel later in more detail. While ESCON became an ad hoc industry standard for high speed fiber data communication, the fundamental structure of the mainframe itself was changing. Rather than using a single large processor, which was expensive to develop and required elaborate maintenance such as water-cooled processors, modern mainframes use a distributed processing approach. Several smaller, less expensive, air-cooled processors based on CMOS technology can be interconnected to operate in parallel on large problems; such machines form a new class of highly coupled facilities to replace conventional mainframe designs. In order for the coupling facility concept to work, high bandwidth data transfer is required to couple closely individual processors; because of speed and distance requirements, fiber-optic data buses are a natural choice for these new distributed architectures. We will discuss the Intersystem Channel used on IBM coupling facilities such as the Parallel Query Server (PQS) and Parallel

[2] System/390 is a trademark of the International Business Machines Corporation.

Transaction Server (PTS), both announced in 1994. The parallel coupled systems use optical links compatible with the FC-0 layer of the emerging ANSI Fiber Channel Standard (FCS). These closely coupled systems still use ESCON technology for I/O communication to the System/390 machines. They also feature adapters that connect mainframes to outside networks such as fiber distributed data interface (FDDI); these adapters allow the mainframe to function in an open environment and bring its established processing power to the network. As the role of the mainframe continues to evolve and networks become increasingly complex, fiber-optic interconnects are expected to play an increasingly critical role in the fabric of data communications.

6.2 Optical Data Link Design

6.2.1 Requirements

There are many factors to be considered in the design of a fiber-optic data link, such as the intended data rate, distance, and fidelity of the data transmission. In this section, we discuss some of the requirements for data link design and then discuss how these requirements can be met with available hardware. We present accepted design methods for modeling link performance and noise sources on the link, as well as common design practices to ensure the quality of a fiber-optic channel.

Any fiber-optic channel consists of an optical source or transmitter, the cable plant, and an optical receiver. The transmitter is capable of launching a limited amount of optical power into the fiber; the receiver has a lower limit on sensitivity or the ability to detect a weak optical signal. A fundamental consideration is the optical link budget, or the difference between the allowable transmitter output and receiver input signal strength requirement. The link budget determines the allowable loss of the cable plant resulting from fiber attenuation, couplers, and splices, as well as from the acceptable level of noise and interference on the link. The optical power level (usually measured in dB) is related to link performance as measured by the bit error rate (BER). Link BER is measured as the number of errors per second, and is related to the signal-to-noise ratio, Q, by the well-known Gaussian integral

$$\mathrm{BER} = \frac{1}{\sqrt{2\pi}} \int_{Q}^{\infty} \mathrm{e}^{-Q^2/2}\, dQ \cong \frac{1}{Q\sqrt{2\pi}} \mathrm{e}^{-Q^2/2} \tag{6.1}$$

Thus if we plot BER versus received optical power on a semilog scale, the resulting curve is a straight line as shown in Figure 6.2. Because the received signal strength depends on the amount and type of noise on the link, the slope of the BER curve will vary for different link and receiver designs. For typical

Gaussian noise sources (to be discussed later), the slope is about 1.8 dB/decade; variations in slope or deviation from a straight line can indicate the presence of nonlinear noise sources. Note that when the curve is linear, it is possible to improve the BER by increasing the received optical power. There are certain types of noise sources, however, that are independent of transmitted signal strength; these can produce BER floors, as shown by curve B in Figure 6.2. These sources must be given careful attention, as they can seriously limit the link performance.

There are two basic approaches to modeling the link performance and the link budget. Both recognize that the transmitted and received power levels may vary by several decibels as a result of changes in the operating temperature, power supply fluctuations, electrical noise on either end of the link, manufacturing process variations, or degradation of components near the end of their lifetimes. Variations in the type of data pattern can also affect receiver

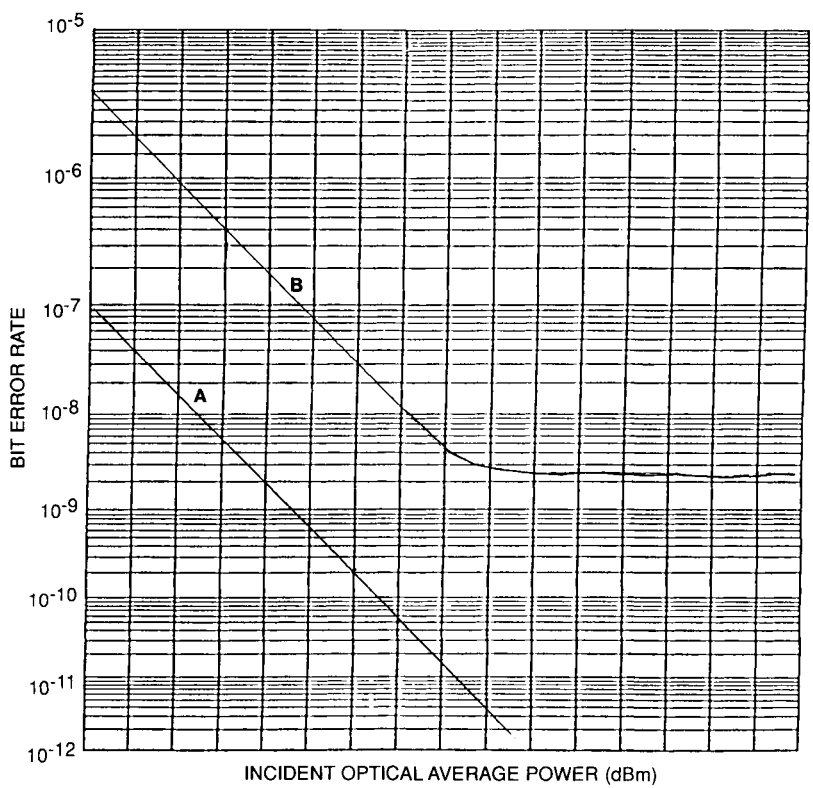


FIGURE 6.2 Bit error rate as a function of received power. Curve A shows a typical performance; curve B shows a BER floor.

sensitivity. For these and other reasons, transmitter output and receiver sensitivity are specified as a range of acceptable values. Additionally, there are some sources of attenuation that are inherently variable, such as the loss caused by a connector or splice in the fiber. One approach is to calculate the worst-case performance of all link components and to design the link budget to operate at the desired BER under these conditions. This approach is desirable when very high performance and consistency is required in the data, or when it is difficult to repair the link during its installed lifetime. However, a worst-case analysis tends to be conservative; as it is unlikely that all link parameters will assume their worst-case performance in a single link, a typical link will be overdesigned and have a large safety margin that could have been utilized to improve performance. An alternative approach is a statistical link model. In this method, the actual distributions of transmitter power, receiver sensitivity, connector variability, and other parameters are measured and statistically combined to determine the overall expected link performance. Non-Gaussian contributions may be treated as part of a Monte Carlo or other simulation; once a cumulative distribution function has been obtained, link performance can be specified at the desired confidence level. The industry standard is a plus or minus 3-sigma design, which considers link performance to be acceptable if it falls within three times the standard deviation of the mean when operating at the BER limit. The statistical approach allows the designer more flexibility in choosing components, and can result in significantly improved distance for the same BER compared with a worst-case analysis. The disadvantage is that all statistical distributions of the link components must be known in addition to the variation of these distributions with temperature, end of life, and so on. Accurate data can be difficult or impossible to obtain, especially as many component vendors consider the distribution of their components to be proprietary because it reflects on their manufacturing process. A useful compromise is to combine the two methods by using worst-case values for those parameters that are not well known, such as transmitter and receiver performance, and statistical analysis for those that are, such as the variability of certain types of optical connectors. This approach exploits the strengths of both methods and will be demonstrated later with a link design example.

Clearly, the acceptable BER is a key factor in link design because of its impact on the link budget. The BER is not only determined by the link configuration, however; data are almost always encoded before transmission and decoded at the destination, and the type of code used can affect the BER. We have already mentioned that the receiver design may be sensitive to certain types of data patterns which degrade the BER performance when they are present. Certain data codes can also be used to check for bit errors in the link, and even to correct some kinds of bit errors. If an error detection or correction code is used, the link may pass this information on to the operating system, which can implement different kinds of error recovery strategies. For

example, some applications may choose to use only single-error detection and request retransmission of all data packets that contain one or more errors. This helps to ensure that error-free data will be received, and might be appropriate for a data center remote backup application. Other systems may find retransmission impractical, such as a real-time video display; in this case, the system may choose to discard any erroneous data so that transmission of a video image is not interrupted. There are many error recovery strategies for different applications, and the link design must consider this when establishing the BER requirement. A system with excellent error recovery, for instance, may be able to tolerate a higher BER on the link.

Other link design requirements may come from conformance to various industry standards, such as those discussed in section 6.3, which place limits on the link physical layer to ensure interoperability among equipment from different vendors. Optical cable installed in modern buildings must also conform to a host of regulations concerning safety, building code conformance, flammability, and so on. The availability of cable that meets installation requirements may be a limiting factor in the choice of optical fiber. Fiber-optic links must also conform to optical safety standards, which limit the maximum optical power in the link. Optical safety standards apply whenever the link can be viewed by someone not trained and certified in the safe handling of high power optical sources; infrared sources of less than a milliwatt can still cause serious damage to the unprotected eye. The exact power levels and measurement procedures vary somewhat from country to country, but always apply to both intentional and accidental viewing of the source and require the source classification to be clearly labeled. In the United States, optical safety is defined by the Department of Health and Human Services of the Occupational Safety and Health Administration. A source that is "inherently safe" for unaided viewing is limited to an observable power level less than -2.0 dBm. International safety requirements as defined by the IEC are more stringent, setting the maximum level for class 1 operation at -6.0 dBm. The power measurement procedures also vary and are generally more stringent for international class 1 certification. Although early standards treated only laser safety, recent revisions may apply to all types of optical sources, including light emitting diodes (LEDs). The type of standards required depends on the link environment and application. For example, in a typical mainframe system, fiber optics may be used to interconnect a central processor with tape drives and other equipment through a "patch panel," or switchboard. It is often desirable to reconfigure the system by changing the patch panel connections; this implies that the connectors should be easy to handle, maintain low loss after multiple insertions, and ensure that class 1 operation is maintained whenever the link is opened. The ability to reconfigure a computer system easily and quickly by changing the interconnections is one of the main advantages of fiber optics, because of their light weight and flexibility.

6.2.2 Optical Link Modeling

We will now discuss important factors in the modeling and design of fiber-optic data links, with emphasis on the link budget requirements [1, 2]. It is convenient to break down the link budget into two areas: installation loss and available power. Installation or DC loss refers to optical losses associated with the fiber cable plant, such as fiber attenuation, connector loss, and splice loss. Available optical power is the difference between the transmitter output and receiver input powers, minus additional losses resulting from optical noise sources on the link (also known as AC losses). With this approach, for instance, one may choose to treat the installation loss budget as statistical and the available power budget as worst case.

First, we consider the installation loss budget. Some of the important factors to be considered include the following:

- Fiber cable loss
- Connector loss
- Splice loss
- Link measurement uncertainty

A typical point-to-point fiber-optic link connecting two devices is shown in Figure 6.3. There are two types of fiber cable commercially available: trunk and jumper cables. Trunk cable is used for long installations (several hundred meters to several kilometers, usually between buildings). A single trunk cable may contain any number of individual fibers. Jumper cables are used for short

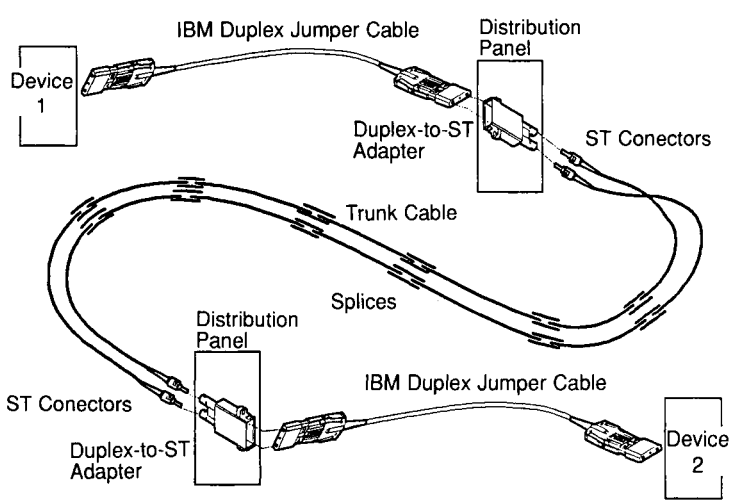


FIGURE 6.3 Typical point-to-point data communication link, with trunk and jumper cables, for an ESCON environment [60].

connections between devices that are close together or between a device and a distribution panel; they typically contain only one or two fibers. Jumper cables are typically several tens of meters in length and may have a slightly higher attenuation loss than trunk cable. Typical trunk cable will have a loss of around 0.5 dB/km, near 1300 or 1550 nm; the maximum loss of jumper cables can be as high as 0.8 dB/km. Fiber attenuation, in general, is usually well controlled, and fibers are available with loss as low as 0.2 dB/km for some applications.

Transmission loss is perhaps the most important property of an optical fiber or cable; it affects the link budget and maximum unrepeated distance. The number and separation between optical repeaters and regenerators is largely determined by this loss. The mechanisms responsible for this loss (material absorption, linear and nonlinear scattering) have been discussed in previous chapters. It was also noted that fiber loss varied with the operating wavelength, providing transmission windows for optical communication. In most optical transmitters, the central wavelength is subject to variation resulting from the manufacturing process, as well as from environmental factors such as temperature. Even for laser sources with a spectral width of 6 nm or less, the center wavelength may vary 80 nm or more. This causes additional loss which must be accounted for in the link budget. Although wavelength-dependent attenuation data is available from most fiber manufacturers, it is often not practical to perform measurements of the fiber at each wavelength of interest, especially during a repair activity. An accurate model for fiber loss as a function of wavelength has been developed by Walker [3]. This model accounts for the effects of linear scattering, macrobending, and material absorption due to ultraviolet and infrared band edges, hydroxide (OH) absorption, and absorption from common impurities such as phosphorous. Using this model, it is possible to calculate the fiber loss as a function of wavelength for different impurity levels; an example plot is shown in Figure 6.4. By using this method, the fiber properties can be specified along with the acceptable wavelength limits of the source to limit the fiber loss over the entire operating wavelength range; design trade-offs are possible between center wavelength and fiber composition to achieve the desired result. Typical loss due to wavelength-dependent attenuation for laser sources on single-mode fiber can be held below 0.1 dB/km.

In previous chapters, optical fiber loss resulting from microbends and macrobends was considered; clearly, these are to be avoided during fiber installation. Another important environmental factor is exposure of the fiber to ionizing radiation damage, which is important in some environments such as in research labs, nuclear power plants, medical facilities, military applications, and space satellites. Although optical fiber is by definition immune to electromagnetic interference, there is a large body of literature concerning the effects of ionizing radiation on fiber links [4–8]. There are many factors that can affect the radiation susceptibility of optical fiber, including the type

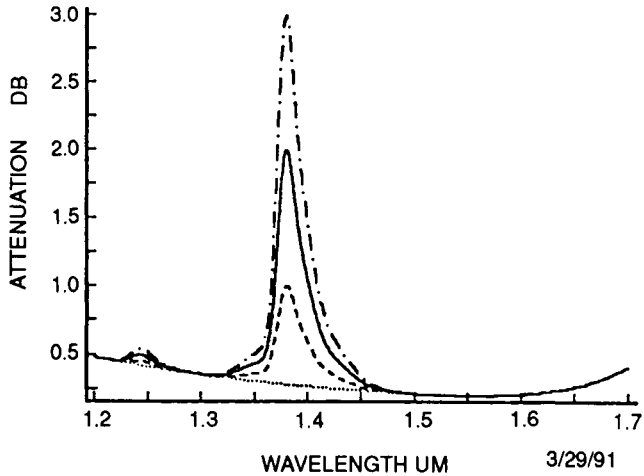


FIGURE 6.4 Spectral loss characteristics of single-mode fiber for different impurity levels (from Ref. 3).

of fiber, type of radiation (gamma radiation is usually assumed to be representative), total dose, dose rate (important only for higher exposure levels), irradiation history of the fiber, temperature, wavelength, and data rate. Because of the many factors involved, there does not exist a comprehensive theory with which to model radiation damage in optical fibers. Models for the fiber's response to high radiation levels have been proposed [5–8]. There have also been studies of the long-term reliability and lifetime effects of background radiation levels, such as those encountered in buried or undersea fibers [9]. Recently, several models have been advanced [10, 11] for the performance of fiber under moderate radiation levels; the effect on BER is a power law model of the form

$$\text{BER} = \text{BER}_0 + \text{A}\,(\text{dose})^b \tag{6.2}$$

where BER_0 is the link BER before irradiation, the dose is given in rads, and the constants A and b are empirically fitted. We will not discuss in detail the modeling work being done in this area, information on which is available in several references [10, 11]. In most applications, the link can be sufficiently shielded from exposure to mitigate these effects, or specially fabricated radiation-hardened fiber can be used to reduce the effects. The loss from normal background radiation exposure over a typical link lifetime can be held below about 0.5 dB.

There are also installation losses associated with fiber-optic connectors and splices; both of these are inherently statistical in nature. There are many different kinds of standardized optical connectors, some of which are shown

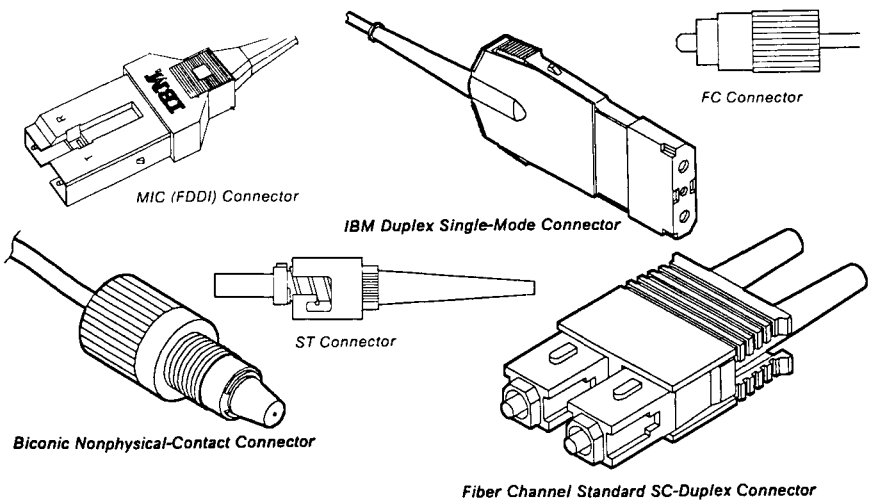


FIGURE 6.5 Typical fiber-optic connector embodiments: MIC (FDDI), ST, FC, ESCON (duplex), fiber channel standard SC (duplex), and biconic (nonphysical contact) [60].

in Figure 6.5. Typical IBM ESCON cable plant optical losses are shown in Table 6.1. An important property of optical connectors (besides loss) is repeatability or consistent loss when the same pair of connectors are mated several times. The loss between randomly chosen pairs of connectors is generally not the same due to optical and mechanical differences. By "making and breaking" random pairs of connnectors many times, a connector loss distribution can be measured; for large number of connections, the distribution is modeled as Gaussian and can be characterized by a mean and sigma value as shown. Typically, published manufacturer's connection loss values assume similar fiber parameters on both sides of the connection. This may not always be the case; for example, because of different tolerances on jumper and trunk cables, there may be a mode field diameter mismatch between a jumper-to-trunk connection at a distribution panel. Usually, the variation attributable to mechanical tolerances on the connectors will dominate any excess connector loss resulting from mode field diameter mismatch; for some applications, this represents an additional statistical penalty which must be accounted for in the installation loss budget. Usually, the net effect is limited to a few tenths of a decibel. There are many different models that have been published for estimating connection loss resulting from fiber misalignment [12–14]; most of these treat losses resulting from misalignment of fiber cores, offset of fibers on either side of the connector, and angular misalignment of fibers. The loss due to all of these effects is then combined into an overall estimate of the connector performance. Although some of these models have

experimental verification, there is still no general model available to treat all types of connectors, which would then allow the modeling information to be used to improve the connector design. The development of optical connectors remains something of an art; even so-called "standardized" connectors exhibit widely different properties depending on the manufacturer. The standard merely ensures that connectors will be mechanically compatible and interoperate; they do not guarantee system performance or other desirable properties such as ease of plugging and durability.

Optical splices are required for longer links, as fiber is usually available in spools of 1 to 5 km, or to repair broken fibers. There are two basic types, mechanical splices (which involve placing the two fiber ends in a receptacle that holds them close together, usually with epoxy) and fusion splices (in which the fibers are aligned, then heated sufficiently to fuse the two ends together). Fusion splices have become much more common because of their low loss and the advent of new portable tools for efficiently performing fusion splices in the field. Typical splice loss values are also given in the Table 6.1.

TABLE 6.1 Typical ESCON cable plant optical losses [60]

Component	Description	Size (μm)	Mean Loss	Variance (dB^2)
Connector[a]	Physical contact	62.5–62.5	0.40 dB	0.02
		50.0–50.0	0.40 dB	0.02
		9.0–9.0[b]	0.35 dB	0.06
		62.5–50.0	2.10 dB	0.12
		50.0–62.5	0.00 dB	0.01
Connector[a]	Nonphysical contact (multimode only)	62.5–62.5	0.70 dB	0.04
		50.0–50.0	0.70 dB	0.04
		62.5–50.0	2.40 dB	0.12
		50.0–62.5	0.30 dB	0.01
Splice	Mechanical	62.5–62.5	0.15 dB	0.01
		50.0–50.0	0.15 dB	0.01
		9.0–9.0[b]	0.15 dB	0.01
Splice	Fusion	62.5–62.5	0.40 dB	0.01
		50.0–50.0	0.40 dB	0.01
		9.0–9.0[b]	0.40 dB	0.01
Cable	IBM multimode jumper	62.5	1.75 dB/km	NA
	IBM multimode jumper	50.0	3.00 dB/km at 850 nm	NA
	IBM single-mode jumper	9.0	0.8 dB/km	NA
	Trunk	62.5	1.00 dB/km	NA
	Trunk	50.0	0.90 dB/km	NA
	Trunk	9.0	0.50 dB/km	NA

[a] The connector loss value is typical when attaching identical connectors. The loss can vary significantly if attaching different connector types.

[b] Single-mode connectors and splices must meet a minimum return loss specification of 28 dB.

An additional effect of lossy connectors and splices is modal noise. Because high-capacity optical links tend to use highly coherent laser transmitters, different modes propagating in the fiber may interfere with one another. Random coupling between fiber modes causes fluctuations in the optical power coupled through splices and connectors; this phenomena is known as modal noise [15]. As one might expect, modal noise is primarily a concern when using laser sources in conjunction with multimode fiber; recent industry standards [16] have allowed the use of shortwave lasers on 50 μm fiber because of potential cost savings on this type of link hardware. Shortwave lasers are currently readily available because of their wide use in compact disk systems for optical storage, and the relaxed tolerances of multimode (MM) fiber makes connectorization simpler and less expensive. The main disadvantage to such systems is model noise, along with the limitation of using only 50 μm fiber (the bandwidth of 62.5-μm fiber seriously limits link distance).

Modal noise is usually considered to be nonexistent in single-mode (SM) systems. However, modal noise in SM fibers can arise when higher order modes are generated at imperfect connections or splices. If the lossy mode is not completely attenuated before it reaches the next connection, interference with the dominant mode may occur. This is especially important in LANs where large numbers of connectors and/or splices are used over relatively short distances. The effects of modal noise have been modeled previously [15], assuming that the only significant interaction occurs between the LP_{01} and LP_{11} modes for a sufficiently coherent laser. For N sections of fiber, each of length L in a SM link, the worst-case sigma, σ_m, for model noise can be given by

$$\sigma_m = \sqrt{2}N\eta(1 - \eta)e^{-\alpha L} \tag{6.3}$$

where a is the attenuation coefficient of the LP_{11} mode and η is the splice transmission efficiency given by

$$\eta = 10^{-(\eta_o/10)} \tag{6.4}$$

where η_o is the mean splice loss (typically, splice transmission efficiency will exceed 90%). The corresponding optical power penalty P (dB) resulting from modal noise is given by

$$P = -5 \log (1 - Q^2\sigma_m^2) \tag{6.5}$$

where Q is the signal-to-noise ratio corresponding to the desired BER, as in Eq. (6.1).

One final factor that deserves attention as part of the installation loss budget is the uncertainty in performing link loss measurements. There are many different standardized techniques for measuring link loss [18]. One approach involves using a known source and detector to measure the loss of a sample cable, then replacing the cable with the link under consideration and

measuring the link loss. By taking the difference between these measurements, random factors such as the connection loss can be reduced. However, these effects cannot be totally eliminated, and the replacement loss technique will itself produce a distribution of measured loss values centered about the mean or average link loss. If this variability is small, as for the large tolerances associated with MM fiber, then the statistics of the replacement loss technique can be measured and incorporated into the installation loss budget. However, for some applications (particularly for SM fiber), this approach is not accurate enough. Instead of using typical jumper cables, special reference jumper cables must be used that are manufactured to very tight tolerances. Although reference jumpers are often expensive, they are required to provide the level of accuracy needed in most SM measurements. The requirements of reference jumpers must be determined on an individual basis for different applications; even a good reference jumper results in some link loss variability which must be accounted for in the installation loss budget. Note that because of the nature of the measurement process, it is possible to measure link attenuation that is below the true value; enough link budget margin must be provided to allow for such an error. The design of reference jumper cables and dependence of the assembly loss on the type of cable have been the subject of much research [19, 20] and ongoing work continues in this area.

6.2.3 Optical Power Penalties

Next, we consider the available power budget, which is the difference between the transmitter output and receiver input powers, allowing for optical power penalties due to noise sources in the link. The most important of these effects, and the most important fiber characteristic after transmission loss, is dispersion, which refers to the broadening of optical pulses as they propagate along the fiber. As pulses broaden, they tend to interfere with adjacent pulses, thus limiting the data rate and bandwidth of the optical link. In MM fibers, there are two dominant kinds of dispersion, modal and chromatic. *Modal dispersion* refers to the fact that light can propagate along many different paths, or modes, in a MM fiber; because not all of these paths are the same length, different modes will travel at different velocities and cause pulse broadening. The fiber manufacturer will measure and specify modal bandwidth of the fiber, in units of MHz-km. This has been characterized [23] according to

$$BW_{\text{modal}} = BW_1 / L^{\gamma} \tag{6.6}$$

where BW_{modal} (MHz) is the modal bandwidth for a length L of fiber, BW_1 (MHz-km) is the manufacturer-specified modal bandwidth of a 1-km section of fiber, and γ is a constant known as the modal bandwidth concatenation length scaling factor. The term γ usually assumes a value between 0.5 and 1,

depending on details of the fiber manufacturing and design, as well as the operating wavelength; it is conservative to take $\gamma = 1.0$. Modal bandwidth can be increased by mode mixing, which promotes the interchange of energy between modes to average out the effects of modal dispersion. Fiber splices tend to increase the modal bandwidth, although it is conservative to discard this effect when designing a link. There have been many attempts to fabricate fibers with enhanced modal bandwidth, most of which have not produced commercially viable products [21, 22].

The other major dispersion contribution is *chromatic dispersion*, which occurs because different wavelengths of light propagate at different velocities in the fiber. Chromatic dispersion, BW_{chrom} (MHz), for MM fiber is given by an empirical model of the form

$$BW_{\text{chrom}} = \frac{L^{\gamma_c}}{\sqrt{\lambda_w}(a_0 + a_1 |\lambda_c - \lambda_{\text{eff}}|)} \tag{6.7}$$

where L is the fiber length in kilometers; λ_c is the center wavelength of the source in nanometers; λ_w is the source FWHM spectral width in nanometers; γ_c is the chromatic bandwidth length scaling coefficient, a constant; λ_{eff} is the effective wavelength, which combines the effects of the fiber zero dispersion wavelength and spectral loss signature; and the constants a_1 and a_0 are determined by a regression fit of measured data. From Reference 23, the chromatic bandwidth for 62.5/125-μm fiber is empirically given by

$$BW_{\text{chrom}} = \frac{10^4 L^{-0.69}}{\sqrt{\lambda_w}(1.1 + 0.0189 |\lambda_c - 1370|)} \tag{6.8}$$

For this expression, the center wavelength was 1335 nm and λ_{eff} was chosen midway between λ_c and the water absorption peak at 1390 nm; although λ_{eff} was estimated in this case, the expression still provides a good fit to the data. For 50/125-μm fiber, the expression becomes

$$BW_{\text{chrom}} = \frac{10^4 L^{-0.65}}{\sqrt{\lambda_w}(1.01 + 0.0177 |\lambda_c - 1330|)} \tag{6.9}$$

For this case, λ_c was 1313 nm and the chromatic bandwidth peaked at $\lambda_{\text{eff}} = 1330$ nm. Recall that this is only one possible model for fiber bandwidth (refer to Refs. 61 and 62 for further details). The total bandwidth, BW_t, capacity of MM fiber is obtained by combining the modal and chromatic dispersion contributions, according to

$$\frac{1}{BW_t^2} = \frac{1}{BW_{\text{chrom}}^2} + \frac{1}{BW_{\text{modal}}^2} \tag{6.10}$$

Once the total bandwidth is known, the dispersion penalty can be calculated for a given baud rate. This is also an empirical process; the type of receiver

used can also compensate to some extent for the fiber's bandwidth. For example, the dispersion penalty P_d (dB) is given by

$$P_d = 1.22\left[\frac{\text{Bit Rate (Mb/s)}}{BW_t(\text{MHz})}\right]^2 \tag{6.11}$$

By design, SM fiber does not suffer modal dispersion. Chromatic dispersion is an important effect, however, even given the relatively narrow spectral width of most laser diodes. The dispersion of SM fiber corresponds to the first derivative of group delay, τ_g, with respect to wavelength, and is given by

$$D = \frac{d\tau_g}{d\lambda} = \frac{S_o}{4}\left(\lambda_c - \frac{\lambda_o^4}{\lambda_c^3}\right) \tag{6.12}$$

where D is the dispersion in ps/(km-nm), and λ_c is the laser center wavelength. The fiber is characterized by its zero dispersion wavelength, λ_o, and zero dispersion slope, S_o. Usually, both center wavelength and zero dispersion wavelength are specified over a range of values; it is necessary to consider both upper and lower bounds in order to determine the worst-case dispersion penalty. This can be seen from Figure 6.6 in which fiber dispersion is a function of wavelength for some typical values of λ_o and λ_c; the largest absolute value of dispersion occurs at the extremes of this region. Once the dispersion is determined, the intersymbol interference penalty as a function of link length L can be determined to a good approximation from a model proposed by Agrawal *et al.* [24]:

$$P_d = 5\log(1 + 2\pi(BD\Delta\lambda)^2L^2) \tag{6.13}$$

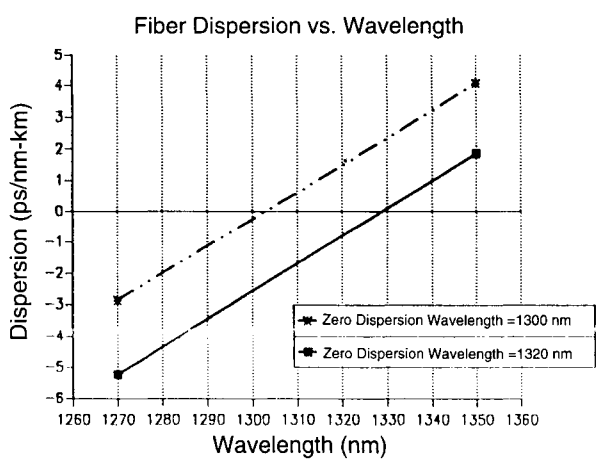


FIGURE 6.6 Graph of fiber dispersion as a function of wavelength.

where B is the bit rate (Mb/s) and $\Delta\lambda$ is the RMS spectral width of the source. More detailed models of dispersion penalty have been proposed [25] that take into account the design details of the receiver and equalizer circuits. By maintaining a close match between the operating and zero dispersion wavelengths, this penalty can be kept to a tolerable 0.5 to 1.0 dB in most cases.

Group velocity dispersion contributes to another optical penalty that remains the subject of continuing research, mode partition noise and mode hopping. This penalty is related to the properties of a Fabry-Perot-type laser diode cavity; although the total optical power output from the laser may remain constant, the optical power distribution among the laser's longitudinal modes will fluctuate. This is illustrated by the model depicted in Figure 6.7. When a laser diode is directly modulated with injection current, the total output power stays constant from pulse to pulse; however, the power distribution among several longitudinal modes will vary between pulses. We must be careful to distinguish this behavior of the instantaneous laser spectrum, which varies with time, from the time-averaged spectrum, which is normally observed experimentally. The light propagates through a fiber with wavelength-dependent dispersion or attenuation, which deforms the pulse shape. Each mode is delayed by a different amount because of the group velocity dispersion in the fiber; this leads to additional signal degradation at the receiver, in addition to the intersymbol interference caused by chromatic dispersion alone, discussed earlier. This is known as mode partition noise; it is capable of generating BER floors, such that additional optical power into the receiver will not improve the link BER. This is because mode partition noise is a function of the laser spectral fluctuations and wavelength-dependent dispersion of the fiber, so the signal-to-noise ratio resulting from this effect is independent of the signal power. For this reason, mode partition noise has been closely studied by many authors; we do not attempt a complete treatment here, but rather summarize some of the more recent models of this effect. The power penalty, P_{mp}, due to mode partition noise was first calculated by Ogawa [26] as

$$P_{mp} = 5 \log (1 - Q^2 \sigma_{mp}^2) \tag{6.14}$$

where

$$\sigma_{mp}^2 = \tfrac{1}{2} k^2 (\pi B)^4 [A_1^4 \Delta\lambda^4 + 42 A_1^2 A_2^2 \Delta\lambda^6 + 48 A_2^4 \Delta\lambda^8] \tag{6.15a}$$

$$A_1 = DL \tag{6.15b}$$

and

$$A_2 = \frac{A_1}{2(\lambda_c - \lambda_o)} \tag{6.15c}$$

The mode partition coefficient, k, is a number between 0 and 1 that describes how much of the optical power is randomly shared between modes; it summarizes the statistical nature of mode partition noise. According to

Ogawa, k depends on the number of interacting modes and root-mean-square spectral width of the source, the exact dependence being complex. However, subsequent work has shown [27] that Ogawa's model tends to underestimate the power penalty due to mode partition noise because it does not consider the variation of longitudinal mode power between successive baud periods, and because it assumes a linear model of chromatic dispersion rather than the nonlinear model given in Eq. (6.14). A more detailed model has been proposed by Campbell [28], which is general enough to include effects of the

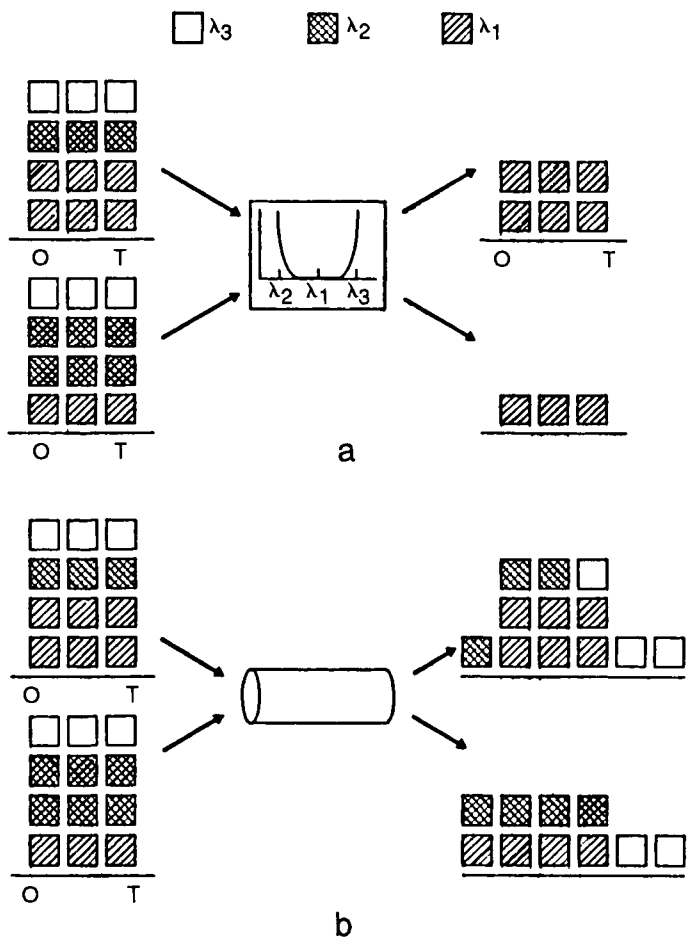


FIGURE 6.7 Conceptual model of mode partition noise. Each block represents a group of photons with the same wavelength. A laser diode emits a combination of wavelengths (different longitudinal modes); top figure (a) shows wavelength-dependent loss, lower figure (b) shows chromatic dispersion effects [26].

laser diode spectrum, pulse shaping, transmitter extinction ratio, and statistics of the data stream. Whereas Ogawa's model assumed an equiprobable distribution of zeros and ones in the data stream, Campbell showed that mode partition noise is data-dependent as well. Recent work based on this model [29] has rederived the signal variance:

$$\sigma_{mp}^2 = E_{av}(\sigma_0^2 + \sigma_{+1}^2 + \sigma_{-1}^2) \tag{6.16}$$

where the mode partition noise contributed by adjacent baud periods is defined by

$$\sigma_{+1}^2 + \sigma_{-1}^2 = \tfrac{1}{2}k^2(\pi B)^4 \times [1.25A_1^4\Delta\lambda^4 + 40.95A_1^2A_2^2\Delta\lambda^6 + 50.25A_2^4\Delta\lambda^8] \tag{6.17}$$

and the time-average extinction ratio $E_{av} = 10\log(P_1/P_0)$, where P_1, P_0 represent the optical power by a 1 and 0, respectively. If the operating wavelength is far from the zero dispersion wavelength, the noise variance simplifies to

$$\sigma_{mp}^2 = 2.25\frac{k^2}{2}E_{av}(1 - e^{-\beta L^2})^2 \tag{6.18}$$

which is valid provided that

$$\beta = (\pi BD\Delta\lambda)^2 \ll 1 \tag{6.19}$$

Mode partition effects deserve careful consideration, as they can often limit the performance of a link or generate BER floors. However, many diode lasers have been observed to exhibit mode hopping or mode splitting, in which the spectrum appears to split optical power between two or three modes for brief periods of time. The exact mechanism is not fully understood, but stable Gaussian spectra are generally only observed for continuous wave operation and temperature-stabilized lasers [30]. During these mode hops, the preceding theory does not apply because the spectrum is non-Gaussian and the model will overpredict the power penalty; hence, it is not possible to model mode hops as mode partitioning with $k = 1$. There is no currently published model that describes a treatment of mode hopping noise, although several recent papers [31, 32] suggest approximate calculations based on the statistical properties of the laser cavity. In a practical link, some amount of mode hopping is probably unavoidable as a contributor to burst noise; empirical testing of link hardware remains the only way to reduce this effect.

In the previous discussion, we mentioned the effect of the receiver extinction ratio on mode partition noise. There is a much more direct effect of this parameter as well. The receiver BER is a function of the modulated AC signal power; if the laser transmitter has a small extinction ratio, the DC component of total optical power is significant. Gain or loss can be introduced in the link budget if the extinction ratio at which the receiver sensitivity is measured

differs from the worst-case transmitter extinction ratio. If the linear extinction ratio E_t at the transmitter is defined as the ratio of optical power when a 1 is transmitted versus when a 0 is transmitted,

$$E_t = \frac{\text{Power (1)}}{\text{Power (0)}} \tag{6.20}$$

then the modulation index at the transmitter M_t can be defined as

$$M_t = \frac{E_t - 1}{E_t + 1} \tag{6.21}$$

Similarly, if the linear extinction ratio is measured at the optical receiver input, then a modulation index M_r at the receiver can be defined. The extinction ratio penalty is given by

$$P_{er} = -10 \log \left(\frac{M_t}{M_r} \right) \tag{6.22}$$

where M_t and M_r refer to modulation specifications for the transmiter and receiver, respectively.

Another important property of the optical link is the amount of reflected light from the fiber endfaces that returns up the link back into the transmitter. Whenever there is a connection or splice in the link, some fraction of the light is reflected back; each connection is thus a potential noise generator, as the reflected fields can interfere with one another to create noise in the detected optical signal. The phenomenon is analogous to the noise caused by multiple atmospheric reflections of radio waves, and is known as multipath interference noise. To limit this noise, connectors and splices are specified with a minimum return loss. If there are a total of N reflection points in a link and the geometric mean of the connector reflections is α, then based on the model of Duff *et al.* [33], the power penalty due to multipath interference (adjusted for BER and bandwidth) is closely approximated by

$$P_{mpi} = 10 \log (1 - 0.7\, N\alpha) \tag{6.23}$$

Multipath noise can usually be reduced well below 0.5 dB with available connectors, whose return loss is often better than 25 dB. A far more serious effect occurs when stray light is reflected back into a Fabry-Perot-type laser diode, giving rise to intensity fluctuations in the laser output. This is a complicated phenomena, strongly dependent on the type of laser, and is called either reflection-induced intensity noise or relative intensity noise (RIN). This effect is very important, as it can also generate BER floors. The power penalty from RIN is the subject of ongoing research; because the reflected light is measured at a specified signal level, RIN is data-dependent, although it is independent of link length. As many laser diodes are packaged in windowed containers, it is difficult to correlate the RIN measurements on an unpackaged laser with those of a commercial product. There have been

several detailed attempts to characterize RIN [34–36]; typically, the RIN noise is assumed Gaussian in amplitude and uniform in frequency over the receiver bandwidth of interest. The RIN value is specified for a given laser by measuring changes in the optical power when a controlled amount of light is fed back into the laser; it is signal dependent and is also influenced by temperature, bias voltage, laser structure, and other factors which typically influence laser output power [36]. If the effect of RIN is to produce an equivalent noise current at the receiver, then the RIN receiver noise, σ_r, may be modeled as

$$\sigma_r = \gamma^2 S^{2g} B \tag{6.24}$$

where S is the signal level during a bit period, B is the bit rate, and g is a noise exponent that defines the amount of signal-dependent noise. If $g = 0$, noise power is independent of the signal, whereas for $g = 1$, noise power is proportional to the square of the signal strength. The coefficient γ is determined by

$$\gamma^2 = S^{2(1-g)}\, 10^{(RIN_i/10)} \tag{6.25}$$

where RIN_i is the measured RIN value at the average signal level S_i, measured for a given application, including worst-case back-reflection conditions and operating temperatures. Clearly, the RIN model contains many parameters that are application-dependent; if these can be determined, the Gaussian BER probability due to the additional RIN noise current is given by

$$P_{error} = \frac{1}{2}\left[P_e^1\left(\frac{S_1 - S_0}{2\sigma_1}\right) + P_e^0\left(\frac{S_1 - S_0}{2\sigma_0}\right)\right] \tag{6.26}$$

where σ_1, σ_0 represent the total noise current during transmission of a digital 1 and 0, respectively, and P_e^1, P_e^0 are the probabilities of error during transmission of a 1 and 0, respectively. The power penalty from RIN may then be calculated by determining the additional signal power required to achieve the same BER with RIN noise present as without the RIN contribution. One approximation from Hakki, Bosch, and Lumish [36] for the RIN power penalty is given by

$$P_{rin} = -5 \log\left[1 - Q^2(BW)(1 + M_r)^{2g}(10^{RIN/10})\left(\frac{1}{M_r}\right)^2\right] \tag{6.27}$$

where the RIN value is specified in dB/Hz, BW is the receiver bandwidth, M_r is the receiver modulation index, and the exponent g is a constant varying between 0 and 1 which relates the magnitude of RIN noise to the optical power level. Because g may vary depending on factors such as the bias point and extinction ratio, this expression provides only an approximation of the true penalty.

Another important area in link design deals with the effects of timing jitter on the optical signal. In a typical optical link, a clock is extracted from the incoming data signal which is used to retime and reshape the received digital

pulse; the received pulse is then compared with a threshold to determine whether a digital 1 or 0 was transmitted. So far, we have discussed BER testing with the implicit assumption that the measurement was made in the center of the received data bit; to achieve this, a clock transition at the center of the bit is required. When the clock is generated from a receiver timing recovery circuit, it will have some variation in time and the exact location of the clock edge will be uncertain. Even if the clock is positioned at the center of the bit, its position may drift over time. There will be a region of the bit interval, or eye, in the time domain where the BER is acceptable; this region is defined as the *eyewidth* [37]. Eyewidth measurements are an important parameter for evaluation of fiber-optic links; they are intimately related to the BER, as well as to the acceptable clock drift, pulse-width distortion, and optical power. At low optical power levels, the receiver signal-to-noise ratio is reduced; increased noise causes amplitude variations in the received signal. These amplitude variations are translated into time domain variations in the receiver decision circuitry, which narrows the eyewidth. At the other extreme, an optical receiver may become saturated at high optical power, reducing the eyewidth and making the system more sensitive to timing jitter. This behavior results in the typical "bathtub" curve shown in Figure 6.8; for this

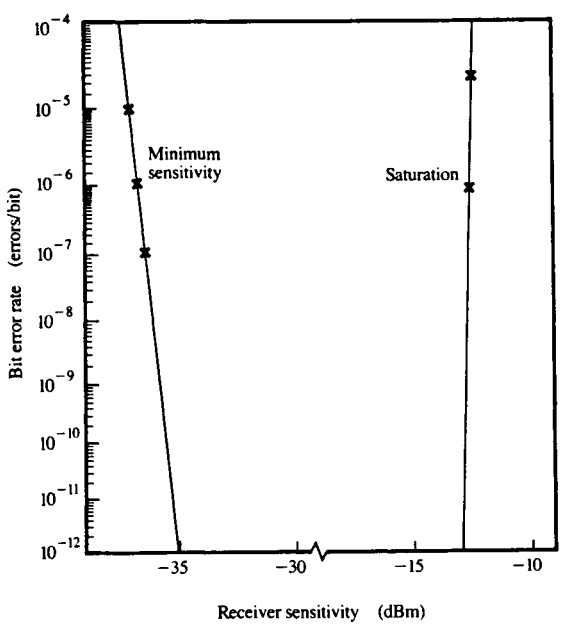


FIGURE 6.8 Bit error rate as a function of received optical power, range of operation is from minimum sensitivity to saturation. From Ref. 59, copyright 1992 by International Business Machines Corporation, reprinted with permission.

measurement, the clock is delayed from one end of the bit cell to the other, with the BER calculated at each position. Near the ends of the cell, a large number of errors occur; toward the center of the cell, the BER decreases to its true value. The eye opening may be defined as the portion of the eye for which the BER remains constant; pulsewidth distortion occurs near the edges of the eye, which denotes the limits of the valid clock timing. Uncertainty in the data pulse arrival time causes errors to occur by closing the eye window and causing the eye pattern to be sampled away from the center. This is one of the fundamental problems of optical and digital signal processing, and a large body of work has been done in this area [38–40]. For the purposes of this chapter, we provide only a brief outline of the problem; this remains an active area of research and the reader is encouraged to follow the recent literature on this subject.

The CCITT had adopted a standard definition of *jitter* [41] as short-term variations of the significant instants (rising or falling edges) of a digital signal from their ideal position in time. Longer-term variations are described as *wander*. In terms of frequency, the distinction between jitter and wander is somewhat unclear. The predominant sources of jitter include the following:

- phase noise in receiver clock recovery circuits, particularly crystal-controlled oscillator circuits, and noise in digital logic resulting from restricted rise and fall times;
- imperfect timing recovery in digital regenerative repeaters, which is usually dependent on the data pattern:
- low frequency jitter resulting from instabilities in clock sources and modulation of transmitters; and
- very low frequency jitter caused by variations in the propagation delay of fibers, connectors, and so on, typically resulting from small temperature variations (this can make it especially difficult to perform long-term jitter measurements); if the jitter is slow enough to be tracked by the receiver clock recovery circuit, it is defined as wander.

In general, jitter from each of these sources will be uncorrelated; jitter related to modulation components of the digital signal may be coherent, and cumulative jitter from a series of repeaters or regenerators may also contain some well-correlated components.

There are several parameters of interest in characterizing jitter performance. Jitter may be classified as either random or deterministic, depending on whether it is associated with pattern-dependent effects; these are distinct from the duty cycle distortion which often accompanies imperfect signal timing. Each component of the optical link (data source, serializer, transmitter, encoder, fiber, receiver, retiming/clock recovery/deserialization, decision circuit) will contribute some fraction of the total system jitter. If we consider the link to be a "black box" (but not necessarily a linear system), then we can

measure the level of output jitter in the absence of input jitter, this is known as the "intrinsic jitter" of the link. The relative importance of jitter from different sources may be evaluated by measuring the spectral density of the jitter. Another approach is the maximum tolerable input jitter (MTIJ) for the link. Finally, because jitter is essentially a stochastic process, we may attempt to characterize the jitter transfer function (JTF) of the link or estimate the probability density function of the jitter. All of these approaches have their advantages and drawbacks; but first, we consider some simple attempts to model jitter in a real link.

One of the first attempts to model the optical power penalty due to jitter [42] considered the general case of a receiver whose input was a raised cosine signal of the form

$$S_t = \tfrac{1}{2}(1 + \cos(\pi Bt)) \tag{6.28}$$

where B is the bit rate. A decision circuit samples this signal at some interval $t = n/B$. In the presence of random timing jitter, the sampling point fluctuates and the jitter-induced noise depends on the probability density function PDF_j of the random timing fluctuations. Determination of the actual PDF_j is quite difficult. If we can approximate $Bt < 1$, then it has been derived [43] that for a uniform PDF_j, the jitter-induced noise σ_j is given by

$$\sigma_j^2 = \tfrac{4}{5}(\pi Bt/4)^4 \tag{6.29}$$

Whereas for the less conservative case of a Gaussian PDF_j in the same limit,

$$\sigma_j^2 = 2(\pi Bt/4)^4 \tag{6.30}$$

The optical power penalty P_j (dB) due to jitter noise is then given by

$$P_j = -5\log(1 - 4Q^2\sigma_j^2) \tag{6.31}$$

Based on this expression, the penalty for a Gaussian system is much larger than for a uniform PDF_j. When $Bt = 0.35$, a BER floor appears at 10^{-9}. However, it was subsequently shown [44] that the approximation on Bt is very restrictive, and the actual PDF_j is far from Gaussian; indeed, it is given by

$$P_{\text{error}} = \int_{-1/B}^{1/B} (PDF_j)\, Q\left(\frac{A\cos(\pi Bt)}{2\sigma_j}\right) dt \tag{6.32}$$

where the probability density function of the jitter is included under the integral (PDF_j) and A is a constant. Numerical integration of Eq. (6.32) shows that the preceding approximate results tend to underestimate the effects of Gaussian jitter and overestimate the effects of uniform jitter by over 2 dB. This points out one of the basic problems in attempting to model a power penalty due to jitter effects, namely, the limitations of the assumptions which must be made to reduce the expression to closed form versus the real-world jitter behavior. Using similar approximations, jitter power penalties have been derived for both PIN and avalanche photodiode (APD) receivers [45],

although the limiting assumptions make it difficult to generalize these models. An alternative modeling approach has been to drive a "worst-case" distribution for the PDF, which will provide an upper bound on the performance of the optical link and which can be more easily evaluated for analytical purposes [46].

The problem of jitter accumulation in a chain of repeaters becomes increasingly complex; however, we can state some general rules of thumb. It has been shown [47] that jitter can generally be divided into two components, one as a result of repetitive patterns and one as a result of random data. In receivers with phase-lock loop timing recovery circuits, repetitive data patterns will tend to cause jitter accumulation, especially for long run lengths. This effect is commonly modeled as a second-order receiver transfer function [48]. The jitter accumulates according to the relationship

$$Jitter \propto N + \left(\frac{N}{\xi} L\right)^2 \tag{6.33}$$

where N is the number of identical repeaters and ξ is the loop-damping factor specific for a given receiver circuit. For large ξ, jitter accumulates almost linearly with the number of repeaters, whereas for small ξ, the accumulation is much more rapid. Jitter will also accumulate when the link is transferring random data. Jitter due to random data is of two types, systematic and random. The classic model for systematic jitter accumulation in cascaded repeaters was published by Byrne [49]. The Byrne model assumes cascaded identical timing recovery circuits; the general expression for jitter has been generalized to networks consisting of different components [50] and to nonidentical repeaters [51]. Despite these considerations, for well designed practical networks, the basic results of the Byrne model remain valid for N nominally identical repeaters transmitting random data:

- Systematic jitter accumulates in proportion to $N^{1/2}$
- Random jitter accumulates in proportion to $N^{1/4}$

Details of jitter accumulation modeling are provided in References 41 through 51.

Except for those penalties that produce BER floors, such as mode partitioning, most penalties can be reduced by increasing the transmitted or received optical power. This brute force approach is subject to limitations such as maintaining class 1 laser safety; even if it were possible to increase optical power significantly (as we will discuss later using optical amplifiers), a new class of nonlinear optical penalties becomes important at high power levels. We will briefly discuss some of these here (Ref. 52 gives a more complete treatment of nonlinear phenomena). Class 1 laser systems typically do not experience these nonlinear effects, although they may be important for optical fiber amplifiers or for systems using open fiber control (OFC), for which there may be significantly higher power levels present in the fiber.

At high optical power levels, nonlinear scattering may limit the link performance. This is usually only significant in telecommunication systems. The dominant effects are stimulated Raman and Brillouin scattering. When incident optical power exceeds a threshold value, a significant amount of light scatters from small imperfections in the fiber core. This is known as stimulated Brillouin scattering; under these conditions, the output light intensity becomes nonlinear. When the scattered light experiences frequency shifts due to modulation by impurities or molecular vibrations in the fiber core, the effect is known as Raman scattering. Both effects do not occur below optical power levels of approximately 5 mW [52, 53].

The final nonlinear effect we will consider is frequency chirping of the optical signal. Chirping refers to a change in frequency with time, and takes its name from the sound of an acoustic signal whose frequency increases or decreases linearly with time. There are three ways in which chirping can affect a fiber-optic link. First, the laser transmitter can be chirped as a result of physical processes within the laser [54, 55]; this effect has its origin in carrier-induced refractive index changes, making it an inevitable consequence of high power direct modulation of semiconductor lasers. Second, a sufficiently intense light pulse will be chirped by the nonlinear process of self-phase modulation in an optical fiber [56]. This effect arises from the interaction of the light and intensity-dependent portion of the fiber's refractive index. Finally, there is a power penalty arising from the propagation of a chirped optical pulse in a dispersive fiber, because the new frequency components propagate at different group velocities. This may be treated as simply a much worse case of the conventional dispersion penalty [57], provided that one of the first two effects exists to chirp the optical signal.

6.2.4 Link Budget Example

To illustrate some of these principles, the following example discusses the design of a data communication link using SM optical fiber operating at 1300 nm. The application is remote backup of data from a hard disk to a tape drive located in a building 20 km away; because of system-wide fault tolerance, the link has a target BER of 10^{-12}. The link adheres to an industry standard data rate of 200 Mbit/s so that the equipment will interoperate. The transmitter is guaranteed to have a worst-case output of -9.0 dBm, accounting for temperature variation and end-of-life; the worst-case receiver sensitivity is -27.0 dBm. Table 6.2 contains the transmitter, receiver, and single-mode optical fiber specification used in this example. Note that the link center wavelength is sufficiently close to the zero dispersion wavelength of the fiber so that the dispersion penalty from Eq. (6.11) is kept to 0.2 dB worst-case. The transmitter is class 1, so we can neglect nonlinear effects in the link; connector return loss is held low enough so that multipath effects are not a concern either. Using the manufacturer's specification for RIN and

TABLE 6.2 Transmitter, receiver, and optical fiber specifications used in the illustrated example

Transmitter		
Power output	−3 to −9	dBm
Extinction ratio	6	dB
RIN	−112	dB/Hz
Center wavelength	1270 to 1340	nm
Spectral width	6	nm
Receiver		
Sensitivity	−27	dBm
Saturation	−3	dBm
Fiber		
Zero dispersion wavelength	1310 nom. 1295–1322 (range)	nm
Zero dispersion slope	0.095	ps/(nm-km)
Cutoff wavelength	1280	nm
Attenuation	0.5 @ 1310 nm	dB/km
Max. attenuation delta (1270 to 1340)	0.06	dB/km

assuming $g = 0.5$, we can estimate the RIN power penalty of 0.7 dB, which should be tolerable. The wavelength-dependent attenuation of the fiber was selected so that the penalty would be limited to 1.2 dB at the required 20 km distance. For this distance, we have chosen a SM link so that modal noise is not a concern provided that we use sufficiently low loss splices and connectors to minimize modal noise. Although mode partition noise should be verified experimentally, we estimate that the power penalty is less than 1.5 dB (note that $\beta < 1$, so the approximate model is valid). Rounding down to be conservative, the preceding yields a total available power of

$$P_{avail} = Tx(\text{output}) - Rx(\text{sens}) - \text{Penalties} = 14\ \text{dB} \tag{6.37}$$

Next, we compute the installation loss. The link contains one distribution panel near the source end and uses previously installed connectors as shown in Table 6.1. We will consider a link with four connectors (mean loss = 0.35 dB, variance = 0.06 dB), two fusion splices (mean loss = 0.15 dB, variance = 0.01 dB), and a 20-km link with 240 m of jumper cable (loss = 0.8 dB/km for the jumpers and 0.5 dB/km for the trunk). By performing a statistical calculation on the installed budget for connectors, splices, and cable transmission loss, the mean installed loss is 11.9 dB with a variance of 0.54 dB. By using a 3-sigma design philosophy, the calculated installation loss is 13.5 dB. This allows us to operate the link with 0.5 dB to spare, the difference between the available power and installation loss budgets. A conservative design should always allow some link safety margin, as we have done, to allow for the many assumptions used in deriving the various link performance models we have used. Of course, there are other possible ways to compute the link budget; if we had obtained the statistics for all of the link

parameters, we could perform a Monte Carlo simulation for the entire link that might yield a different link budget margin. It should be apparent that link design requires not only the latest valid performance models but also a good amount of engineering judgment in how best to apply these tools. Also, real world problems are often not as clear cut as the example we have just provided; a link designer must be able to manage the trade-offs involved in frequently working with incomplete information. Still, this example should illustrate some of the typical steps required to perform link planning and installation.

6.3 Industry Standards for Optical Fiber Data Communication

6.3.1 Fiber Distributed Data Interface (FDDI)

The FDDI standard [58] was designed primarily to be an all-fiber LAN attachment protocol and is compatible with the IEEE 802.2 interface for WANs. Although FDDI currently represents only about 10% of installed LAN protocols (the most popular is Ethernet), it is still a significant LAN for commercial network installations. The FDDI network is a 125-Mbit/s token ring design that uses dual counter-rotating rings to provide redundancy. One ring is used for data transmission, and the other provides a backup in case of a fault in the primary ring. Because it uses a 5/4 bit coding scheme, the effective bit rate for FDDI is actually about 100 Mbit/s; it can, however, be used as a high speed backbone for lower speed 4-, 10-, and 16-Mbit/s LANs.

FDDI uses the token ring protocol to control data flow, which allows devices having different data requirements to be served appropriately. Data is transferred in packets; the token is a special packet whose only function is to give a node permission to transmit data. The token passes around the ring, and a device can only use the ring to transmit data when it has the token. Devices may be given the token in equal turns, or they can be given priority by receiving it more frequently or by holding it longer after they receive it. The token always travels in the same direction around a ring; this protocol can also be implemented on the star or bus topologies.

FDDI is an LED-based, 62.5/125 μm MM link with duplex connectors. The standard also provides for 50-, 85-, and 100-μm fiber, in addition to SM and copper cable options. The 11 dB link budget supports up to 2 km between successive stations; up to 500 stations may be connected on a 100-km duplex cable. Each station on the loop regenerates and repeats the data; because of limitations in the clocking architecture, the variable length frame size is limited to 4500 bytes. If there is a break in the ring, the station adjacent to the interruption wraps the secondary and primary rings together to form a

new ring. If there are multiple failures, this process can result in multiple disjoint rings. The standard also allows for bypass capability in which an optical switch is used to bypass any station on the ring; this allows a station to be taken off line for repairs without disrupting the rest of the ring. In practice, a concentrator is used to attach several physical devices to a single station on the ring. This is more economical than placing each device separately on the ring.

An FDDI station can transmit a data frame only after capturing a token; the station may then continue transmission until it has no more information to transmit or its token-hold timer expires. The station then issues a new token. As more and more data frames are added to the ring by successive stations, the token appears to a downstream station to be successively bumped to the rear, such that it takes longer and longer for the token to complete one circuit of the ring. Thus, the time between the appearances of successive tokens at a given station, called the token rotation time, is directly proportional to the data load on the ring. Accordingly, each station measures the time between successive appearances of a token; as this time (and the data load) increases, only higher priority traffic is allowed on the ring. This enforces fairness for all stations on the ring. Most of the transmission between stations is asynchronous, although the standard does provide for an optional synchronous service protocol as well. Future enhancements, such as FDDI II, use time-division multiplexing to create up to 16 separate isochronous channels in increments of 6.144 Mbits/s. These can be in turn reallocated into 8-kbit/s subchannels to match voice data on North American and European systems; several channels may also be combined to form a larger channel for video and other applications.

The structure of the FDDI standard does not guarantee that a link which meets the prescribed loss budget and distance constraints will be operational; links may also be dispersion limited. Even if link length, fiber type and specifications, source spectral width, and source center wavelengths are all within FDDI specifications, it is still possible to produce link errors via dispersion. For most applications, this does not occur because the dispersion penalty is included in the receiver sensitivity measurement. However, insertion of an FDDI bypass switch increases the link length and may cause dispersion errors even if the loss budget is within specifications. This makes problem determination somewhat difficult, especially as dispersion testing is an expensive and difficult procedure. Often, for a 2-km link with fiber bandwidth greater than 500 MHz-km at 1300 nm, attenuation is the most probable cause of failures, a substitution technique is required to isolate the defective link segment. Furthermore, just as it is possible for a weak FDDI transmitter and receiver pair individually to meet specifications but still fail to function in a link, it is also possible for individual link components to be out of specification and yet still function with a sensitive receiver and a "hot" transmitter. Because of these difficulties, it is recommended that a link budget

be calculated for each FDDI installation and careful records be kept of the link performance for future reference in the event of a component failure. Key features of FDDI include the specification of a distinctive connector type, the media interface connector or MIC connector, which features field-installable, color-coded keys that allow the connector to be configured into one of four possible configurations. This prevents installation errors and is intended to assist in cable management.

6.3.2 Enterprise System Connection (ESCON)

Since the introduction of System 360 in 1964, only incremental upgrades have been made to the mainframe computer I/O interface; the basic interconnection technology remained unchanged until the advent of ESCON. The ESCON architecture [59] was introduced by IBM on the System/390 mainframes in 1990 as an alternative high-speed I/O channel attachment, complementing parallel copper cables. It has become a de facto industry standard for mainframe data communication because of the dominant role IBM plays in the global mainframe marketplace (the ESCON specifications have been published for use by original equipment manufacturers [60]). ESCON is used as a high-speed backbone for data communication, and includes not only the fiber-optic channels but also the supporting architecture, protocols, and control units which make up a mainframe computing enterprise.

ESCON uses a switched point-to-point topology that uses a central nonblocking dynamic switch. Mainframe data channels, control units, and various types of direct access storage devices (DASD) attach to the switch by means of point-to-point bidirectional fiber-optic data links. The result is similar to a star-wired ring, which provides both efficient bandwidth utilization and reduced cabling requirements. It is also convenient to add and reconfigure equipment on such a ring while maintaining the ability to isolate and partition subdivisions of the network. With a properly designed switching element, the switched point-to-point topology allows the host to communicate with many different types of I/O devices. This function is provided by the ESCON Director, an electronic switch developed for the ESCON architecture. Although ESCON uses 8 bit/10 data encoding and bundles data into frames, it is not a packet-switching network. Instead, the leading frame of a data transmission includes a request for connection; the ESCON Director then establishes and maintains this connection in the manner of a circuit switch for the duration of the data communication. Because of limitations in the architecture, a maximum of two Directors is supported in a single ESCON link. Recently, in an effort both to increase supported distance and provide a buffer between the mainframe and a public network interface, ESCON repeaters were introduced in both laser–laser repeater and laser–LED (or LED–laser) converters. Performance constraints on the

channel and timing jitter at the repeaters results in a limit of three repeaters per link, which is more than adequate for most applications. Unlike FDDI, the ESCON specification guarantees operation of the link if all devices are within specification; because the specification tends to be conservative, customized ESCON links have been installed that can function 10 to 20% or more beyond their distance limitations.

There are two basic ESCON channels: multimode, which supports both 50 and 62.5 μm trunk cables, and single-mode (also known as the Extended Distance Feature, or ESCON XDF). All channels operate at a wavelength of 1300 nm and a data rate of 200 Mbit/s, use 8 bit/10 data encoding, and provide international class 1 laser safety. Multimode ESCON using LED transmitters supports a maximum distance of 2 km on 50-μm fiber and 3 km on 62.5-μm fiber, both with an 8 dB loss budget. ESCON XDF, using laser transmitters and SM fiber, supports up to 20 km with a 14 dB loss budget. Because of their use in mainframe applications, data integrity is critical for ESCON channels; although their published BER specification is 10^{-12}, these channels commonly run at much lower rates, typically about 10^{-15} (as compared with about 10^{-9} for typical voice communications). All ESCON channels use a specially developed duplex connector, illustrated in Figure 6.5. The ESCON connector is designed to prevent accidental insertion of a multimode cable into a single-mode receptacle for laser safety purposes. Additionally, the connector features a unique spring-loaded retractable dust cover, which protects the ferrules when the connector is not in use and retracts when it is plugged into an ESCON receptacle. Because the ESCON connector was intended for use by mainframe customers who may frequently reconfigure their computer systems, a great deal of attention was given to designing a connector that combined ruggedness with ease of use. Figure 6.9 shows the statistical connection loss for randomly paired ESCON connectors; the mean loss is less than 0.3 dB.

6.3.3 Fiber Channel Standard (FCS)

In an attempt to address the need for high bandwidth data communication links, a large number of different standards have emerged over the years. These include the Small Computer Systems Interface (SCSI) and High Performance Parallel Interface (HIPPI), to name a few. In an effort to meet the needs of equipment designers while removing some of the confusion proliferating the data communication industry, the American National Standards Institute (ANSI) has developed a new standard that encompasses many of the existing standard documents while providing a path for growth and development into higher speed links. The ANSI Fiber Channel Physical and Signaling Interface (FC-PH) standard began in 1988 under the auspices of X3T9.3 Working Group of the Accredited Standards Committee; the most recent draft is currently completing final public review and is expected to be

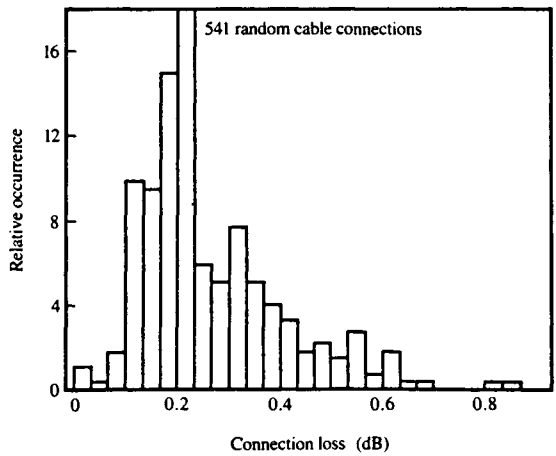


FIGURE 6.9 Connection loss of randomly selected ESCON duplex jumpers [59]. Copyright 1992 by International Business Machines Corporation, reprinted with permission.

adopted by 1995 [16]. Many leading data communication equipment manufacturers have already stated their approval, however, and early FCS-compatible hardware is now available from some vendors. Although FCS will continue to coexist with installed fiber and cable plants, it is also intended to supersede many of these standards, which currently serve niche markets (see Fig. 6.10).

Fiber Channel is logically a bidirectional point-to-point serial data link. Physically, it may consist of either a dedicated point-to-point topology or an interconnected switching network or fabric which ties together multiple communication stations with dedicated single links (a third option, the arbitrated loop, permits connection of three or more devices without use of a fabric). A switching fabric, as defined by FCS, refers to a network in which all station management functions are controlled by a switching point, rather than by the users at each station. An analogy is the telephone company's network, in which users specify an address (phone number) for their communication needs and the network provides them with an interconnection path. This removes the need for complex switching algorithms at each station; FCS allows for more than 16 million different addresses. Although we are concerned with the use of optical fiber transmission media, the standard does provide for attachment of twisted pair and copper coaxial cable data links. Fiber Channel uses the standard 8 bit/10 bit DC-balanced code for data transmision, maintaining a minimal BER of 10^{-12}. Three classes of service are defined in the standard. Class 1 establishes dedicated connections between two devices, which is then guaranteed by the fabric and makes optimal use of the link bandwidth. Classes 2 and 3 provide multiplexing services at data

frame boundaries, with class 2 providing acknowledgement of frame transmission while class 3 does not (if no fabric is present, classes 2 and 3 become special cases of point-to-point interconnection).

The standard provides for long-wave (1300 nm) SM data links at 265.625 Mbit/s, 531.25 Mbit/s, and 1062.5 Mbit/s over distances from 2 to 10 km using laser sources. Multimode links are specified as well, with an additional data rate of 132.8125 Mbit/s. The two slowest MM speeds use long-wave LED sources, whereas the higher speeds on MM fiber use short wavelength (780–850 nm) lasers. For each data rate/distance combination, the standard specifies the physical layer characteristics of the transmitter and receiver (see Fig. 6.11). These specifications are derived from a link budget model proposed in an appendix to the standard; they guarantee operation of links at the required distances and data rates if all components fall within specification. FCS uses the SC duplex connector for optical fiber, keyed to prevent insertion of a multimode fiber into a SM receptacle (see Fig. 6.5). All FCS links must meet international class 1 laser safety levels; this can be

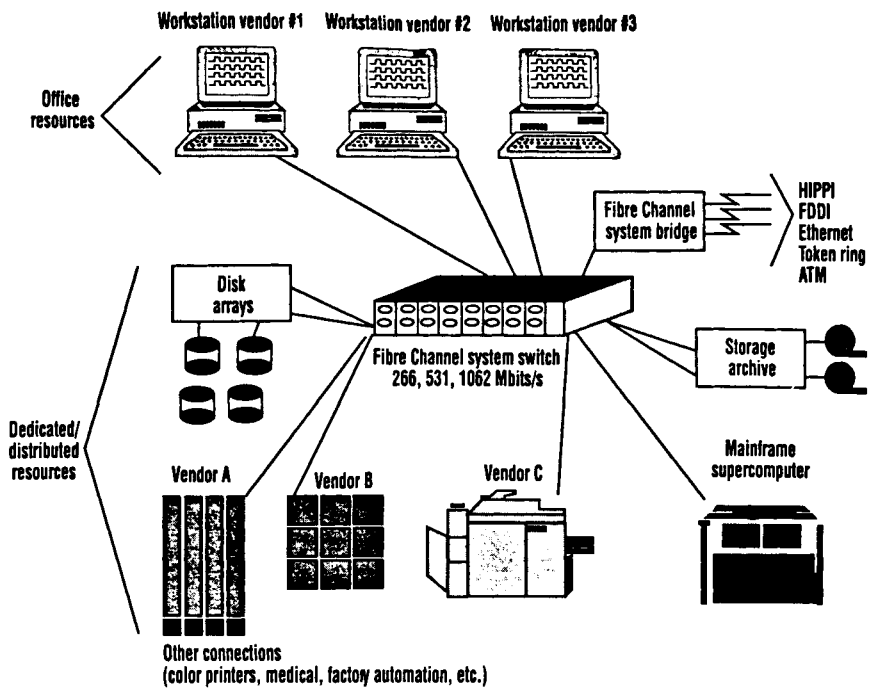


FIGURE 6.10 Example of a Fiber Channel Fabric, interconnecting workstations, mainframes, multiple vendors, and storage through a circuit or packet switched network [61]. A bridge provides connections to other protocols, such as ATM, FDDI, and HIPPI. Reprinted with permission from *Electronic Design*, May 1993. Copyright 1993, Penton Publishing Co.

FIBRE CHANNEL MEDIA PERFORMANCE

Media type	Data-transmission rate (Mbytes/s)	Maximum distance	Signaling rate (Mbaud)	Light source/ logic levels
Single-mode fiber	100	10 km	1062.5	Longwave laser
	50	10 km	531.25	Longwave laser
	25	10 km	265.6	Longwave laser
50-μm multimode fiber	25	2 km	265.6	Shortwave laser
	50	1 km	531.25	Shortwave laser
	100	0.5 km	1002.5	Shortwave laser
62.5-μm multimode fiber	25	1 km	265.6	Longwave LED
	12.5	500 m	132.8	Longwave LED
Video coax	100	25 m	1062.5	ECL
	50	50 m	531.25	ECL
	25	75 m	265.6	ECL
	12.5	100 m	132.8	ECL
Miniature coax	100	10 m	1062.5	ECL
	50	20 m	531.25	ECL
	25	30 m	265.6	ECL
	12.5	40 m	132.8	ECL
Twisted pair	25	50 m	265.6	ECL
	12.5	100 m	132.8	ECL

FIGURE 6.11 Fiber Channel Standard performance for various link technologies [61]. Reprinted with permission from *Electronic Design*, May 1993. Copyright 1993, Penton Publishing Co.

achieved either by using conventional safety shutters and limited transmitter power designs or by a new method called open fiber control (OFC) on MM links. Using OFC, the link transceivers perform a handshake to initialize the channel; if the full duplex link is interrupted (because of a broken fiber or failed transmitter, for example), then the link automatically detects this condition and shuts down transmitters at both ends of the link. This occurs quickly enough so that there is no danger of human exposure to unsafe optical power levels; the OFC timings are specified as part of the standard. When the link enters this state, both transmitters emit a very low duty cycle pulse train, waiting for a closed link to be reestablished via a response from the other end. When the link is closed again, the OFC pulses perform a handshake and return the transmitters to full operating power levels. This system removes the need for mechanical safety shutters on both transmitters and allows the use of higher power levels to achieve longer distances in an operating link. Both transceivers contain identifying information to guarantee that a module with OFC capability will only interoperate with a similar module type.

Fiber Channel is designed as a high speed, point-to-point data bus, and is therefore expected to complement current efforts in ATM development of high speed packet switching. One of the most significant applications so far has been the use of FCS links in the highly parallel coupling facilities

announced by IBM in 1994 [62]. These intersystem channels offer two options: a shortwave, 531-Mbit/s link on MM fiber with a maximum distance of 1 km, and a long-wave 1062.5-Mbit/s link on SM fiber with a maximum distance of 3 km (1 km longer than the maximum distance in the standard). Both links use OFC safety control, although the standard does not specify timing for the gigabit links; this implementation uses timings for the 266-Mbit/s links on the gigabit OFC. FCS is also being adopted by a number of leading workstation vendors; in particular, the Common Optics Group represents a consortium of HP, SUN, and IBM whose goal is to develop a common profile for workstation communications. Today, many workstations operate at a benchmark of about 150 MHz; with performance increasing about 30% per year in recent years, this level is expected to reach 1 GHz by the year 2000 [61]. Fiber Channel may be expected to play an important role in emerging workstation connectivity requirements, as well as in the mainframe and high-end processor environments.

6.4 Intraconnection Applications

So far, we have discussed the design of fiber-optic links and their application to industry standard data channels. The interconnection links discussed in this chapter thus far deal with interconnection applications, that is, links between computing devices. In this section, we consider another important area, namely, intraconnection links or optical channels within a computer device. Typical applications include optical data buses and optical clock distribution. Such applications do not require the kilometer long distances typical of interconnection applications; however, they can still benefit from the use of fiber optics. As the density of high-speed integrated circuits continues to increase, the bandwidth of clock and data bus channels must keep pace; the inability of coaxial cable to meet this requirement has led to the so-called "I/O bottleneck," which is a well-known limitation on computational speed [63, 64]. In addition to providing high bandwidth, fiber optics has the potential to provide a large number of connections in a relatively small space; channels can be placed close together without crosstalk or electromagnetic interference, and optical fibers can also carry bidirectional signals.

Other optical technologies have been considered for the relatively short distances required in clock distribution systems. One example is holographic optical elements [65]. Holograms are attractive because of their ability to redirect and focus light to many points with minimal loss, replacing conventional optical splitters when the application permits line-of-sight operation. This technology remains immature, and although the potential is great, we cannot expect holographic elements to appear in commercial products anytime soon. Another significant problem is the integration of optical

systems with conventional electronic packaging technology; several authors have investigated the use of thin-film optical waveguides for this purpose [66, 67]. Infrared waveguides can be fabricated out of polymeric or organic materials which can be deposited and patterned using conventional photolithography techniques; optical fibers, transmitters, and receivers can all be closely interfaced with such waveguides. Most waveguides have very low dispersion (a few picoseconds per centimeter); misalignment between guides can lead to mode selective loss, but this can be controlled with careful processing. Single-mode waveguides with submicron feature control have been demonstrated [67]. Their drawbacks include high loss (0.5 to more than 1 dB/cm) and poor thermal/mechanical compatibility with other semiconductor materials and processing methods. Many new materials are under investigation in the laboratory and some are commercially available for use today; implementation of this technology has been demonstrated, but probably remains years away from volume production.

The advantages of optical clocks and buses have been known for some time, and there has been a great deal of research in this area [68, 69]. There are still no products available with optical clock or bus designs, mainly because ongoing efforts to squeeze the remaining performance out of electronic distribution systems have pushed back the requirements for higher bandwidth within the machine. There is also some reluctance to abandon the familiar coaxial distribution technologies in favor of optical fiber, given the large investment most companies have in existing facilities. Higher levels of integration have also reduced the length requirements between processing elements. The potential of optical intracommunication has not been ignored, however, and practically all major computer industry companies have published some amount of research in this area, often sponsored by government initiatives such as ARPA programs [70]. There is at least one initiative under way in Europe that has demonstrated optical clock and bus technologies outside the laboratory in a distributed processing environment [71, 72], with the stated goal of commercializing advanced computer products by the end of the decade. This is the ESPRIT program, a consortium of government, university and private industry sponsors that have been working on this problem since 1989. Given the large potential for this area, we will discuss some of the unique requirements and potential solutions for such a system.

Because the optical clock and data bus have similar requirements, we will consider the clock design in some detail; results can be generalized to optical data bus systems. Traditionally, the performance of both large mainframes and smaller workstations have been improved by increasing the clock frequency, increasing the number of microprocessors that make up a single machine image, and reducing the number of cycles per instruction. Whereas the third element depends on factors such as the machine architecture, the first two are directly related to the clocking system. Potential advantages of optical clocking include improved fidelity of the clock waveform at high

frequencies, sharp rising edges of the clock pulse (reducing timing skew and jitter), and immunity to electromagnetic interference. Although current mainframes offer typical performance cycle times at about 10 ns, optical systems have been demonstrated that are capable of reducing this to a few nanoseconds or less. In fact, it is ironic that one obstacle to the adoption of optical clocking are the limitations of the remaining hardware in the computer; system architects are unaccustomed to designing a traditional system around such short cycle times. The full benefits of this technology will probably only be realized when the rest of the system is designed to take advantage of a 500-MHz or faster clock.

Many optical clock designs use either gain-switched or mode-locked lasers, which are capable of drastically reducing the timing skew and jitter. This is important because the clock speed of digital systems is often limited by electrical intraconnect and packaging technology (we shall discuss several alternatives to the packaging problems shortly). In the related area of optical data buses, design problems emerge when the distance between synchronously clocked circuit elements is on the same order as the product of the clock pulse rise time and the signal propagation velocity. For example, the averaging spacing between integrated circuit chips is about 6 cm, whereas a mainframe system with a 7.5-ns cycle time might utilize a clock rise time of about 400 ps with a typical velocity in coax cable of 1.5×10^{10} cm/s. Additional difficulties arise when the typical integrated circuit gate delay (75 ps/cm) is a significant fraction of one-quarter of the clock period for many data bus designs.

For clocking systems, skew is defined as the differences in arrival times of the clock signal to any two points in the circuitry. It is a relative measurement; in an ideal system, the clock would reach all points simultaneously. In reality, differences in the optical path length (especially for connectorized optical fibers) and variations in receiver circuit response introduce offset in the clock signal arrival times. Distribution of an optical signal to many points involves an optical splitter or star coupler. Commercially available products can produce skew as low as 10 ps between branches for 25 or more branches. Allowing for skew resulting from the electronics in the system, an overall performance of less than 20 ps skew is not unreasonable. The variation in arrival times of successive clock pulses to a given point in the computer is called jitter. Usually, passive components such as the coupler contribute negligible jitter to the signal; the dominant source is the optical transmitter. Both mode-locked and gain-switched lasers are capable of producing pulse trains with less than 10 ps jitter; this is far superior to existing electrical designs, which often measure a system-level performance of hundreds of picoseconds.

We have noted that optical clock and bus systems are often only a few hundred meters in length; the link power budget is still important, however, as it is not a point-to-point connection. Rather, the use of optical splitters

introduces a significant loss in the link; modal noise can also be a concern even on short links, and, of course, length-independent properties such as RIN are still present and can lead to BER floors if not handled carefully. Dispersion-related noise is usually not a problem because of the short distances involved. We can see that the link budget concerns for a clock or bus system are quite different from a long haul I/O channel; still, the same principles discussed earlier can be applied to modeling performance of the fiber-optic channel. Because of the critical nature of the clock and bus signals, the desired BER is usually well below 10^{-15}; from a system design point of view, there can be essentially no observed errors in the clock signal without potentially interrupting the entire computer.

There is another area that should not be neglected in the design of future optical clock and data bus systems, namely, the interface between the optical fiber and the electronic package. Often, chips are bundled together in multichip modules (MCMs), collections of chips on a common silicon or glass-ceramic substrate connected by dense internal wiring. Because it is not practical or economical to redesign existing MCMs, there is need for an interface that brings optical fiber across the MCM boundary to the rest of the computer electronics. This need has been recognized by the industry for some time [73, 74], and various proposals have been presented to address this problem. Because of space limitations, we describe one typical approach to designing efficient, embedded optical fiber receptacles in current MCMs.

A typical MCM package design is illustrated in Figure 6.12. The top surface is usually taken up by heat sinks to provide effective cooling for the densely packaged chips. It is often not practical to interfere with this design by bringing optical fibers in through the top of the module. The bottom surface, not shown, is occupied by pin grid arrays and electrical connectors that bring signals and power into the module. Even if all of the signal pins are replaced by optical fiber, the power connections must remain; in fact, it is desirable to separate the signal and power connections from a noise perspective. It is therefore more advantageous to bring a new optical interface into the edge of the MCM package where it will not interfere with existing structures and can easily interface with chips on the edge of the MCM substrate. This technique is shown in Figure 6.12, in which a small bundle of optical fibers enters through the MCM side wall. One of the most significant problems is the relatively tight tolerances that must be maintained on optical fiber alignment with transmitters and receivers; the fiber must also be strain-relieved at the MCM interface. Some of the alignment problems can be overcome by embedding the fibers in preferentially etched V-grooves in a silicon substrate [76]; a mechanical boot may then be used for strain relief. A recent patent in this area [75] uses both these effects to bring a few optical fibers through the side wall of an MCM, as illustrated in Figures 6.13 and 6.14. Note that the small bundle of fibers (*23*) passes through the strain-relief block (*22*) in the MCM side wall at an angle sharp enough to provide strain relief; the fibers

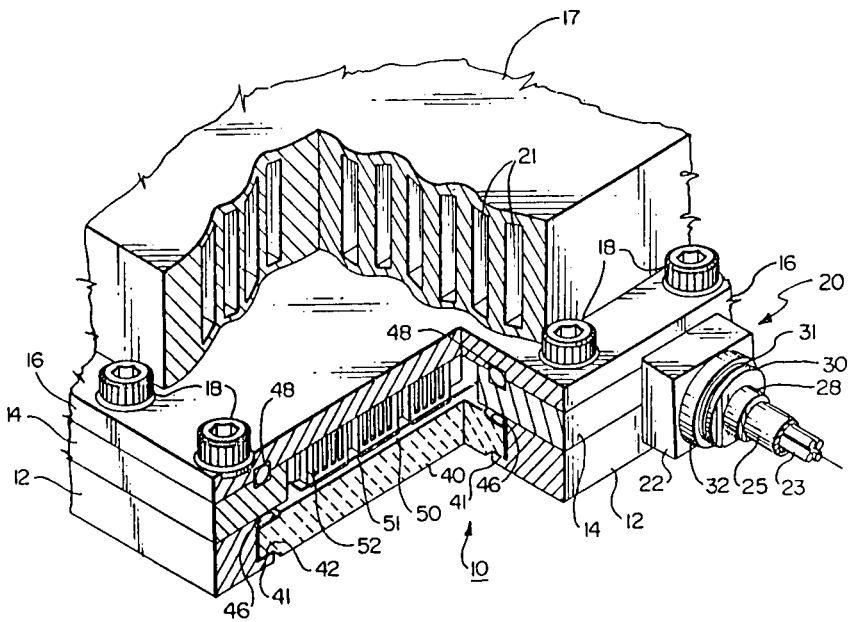


FIGURE 6.12 External view of a multichip module (MCM) package with an embedded permanent optical fiber attachment [75]. Several fibers (*23*) are bundled into a cable and brought into the side wall of an MCM, where fibers are permanently attached.

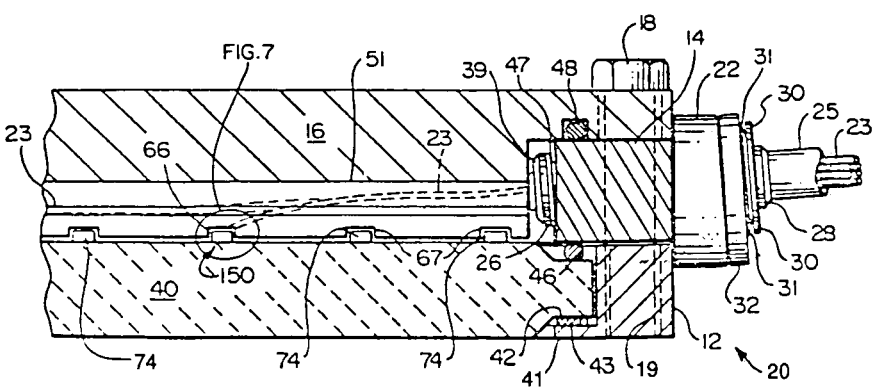


FIGURE 6.13 Illustration of multichip module package cross-sectional view of optical fiber strain relief [75].

are then attached to silicon chips (*150*) inside the module. This technique brings fiber in through a liquid-tight seal (*30, 31*) and is intended to be compatible with MCMs that rely on liquid cooling, such as the IBM Thermal Conduction Module (TCM). Once inside the module (see Fig. 6.14), the fibers (77) are guided by silicon V-grooves on a miniature optical bench (*80*) which provide alignment to an electronic chip (*84*) containing transmitters (*88*) and/or receivers (*100*). The arrows (*123, 125*) in Figure 6.14 indicate the

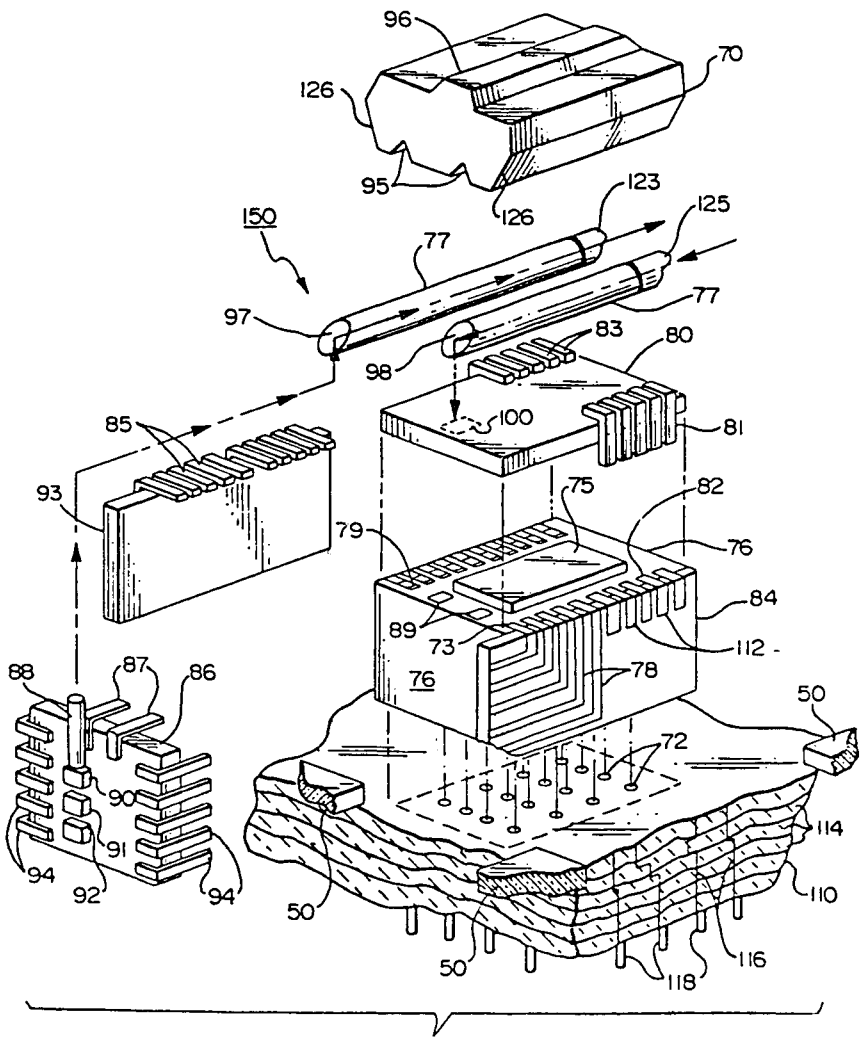


FIGURE 6.14 Illustration of multichip module package cross-sectional view of optical fiber silicon V-groove alignment technique with miniature optical bench [75].

direction of light. The fiber end face is cleaved at an angle (*97*) to deflect light into a receiver or from a transmitter into the fiber. Although we will not describe details of the optoelectronics used, this method provides one way to interface fiber with chips inside an MCM.

This approach is adequate for optical clocking in which only a few fibers are required and they may be permanently attached to the MCM. Mechanical limitations prevent this design from being scaled up to allow for a multifiber bus, which may require dozens of fibers and which must also be separable from the MCM interface for channel reconfiguration and maintenance. These problems have been addressed in another patent application [77], which describes the design of a scaleable, pluggable fiber-optic interface to multichip modules. This design is illustrated in Figure 6.15, which shows a ribbon fiber (*15*) and pluggable connector (*8*) engaged with a receptacle embedded in the MCM substrate. The pluggable connector maintains fiber alignment using silicon V-grooves; the fibers may be slightly bent to facilitate passing through available openings in the MCM superstructure. The receptacle shown in Figures 6.16 and 6.17 is a key feature of the design; it may be aligned with the substrate using the same coordinate grid as the electronic chips and bonded to a well in the substrate with lithographic precision. Silicon V-grooves are used in the receptacle (*18*) to align the fibers; a cylindrical lens (*17*) runs along the length of the receptacle to focus light across the interface and improve coupling efficiency between the fibers. Fibers (*21*) embedded within the MCM multilayer substrate collect the light and by cleaving the end faces (*26*), can redirect the light to illuminate detectors (*36*) placed on the lower surfaces of chips (*6*). By using this technique, up to 32 fibers can be connected in parallel within a $\frac{1}{2}$-in. long space. By using multiple connectors

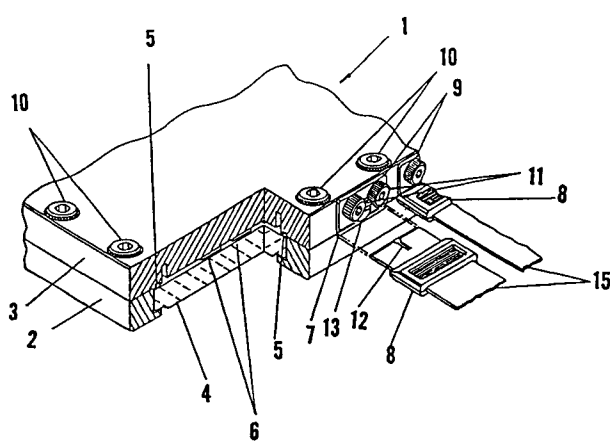


FIGURE 6.15 Pluggable connector design for parallel fiber optic interface to electronic module [77].

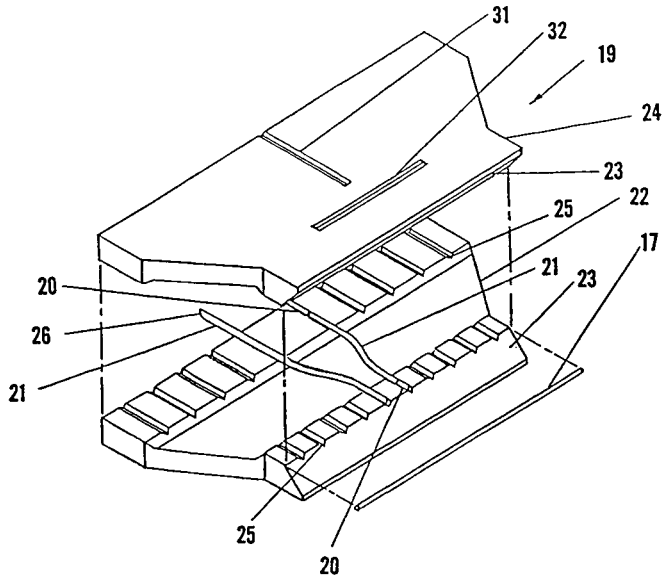


FIGURE 6.16 Exploded view of parallel fiber connector using V-groove alignment [77].

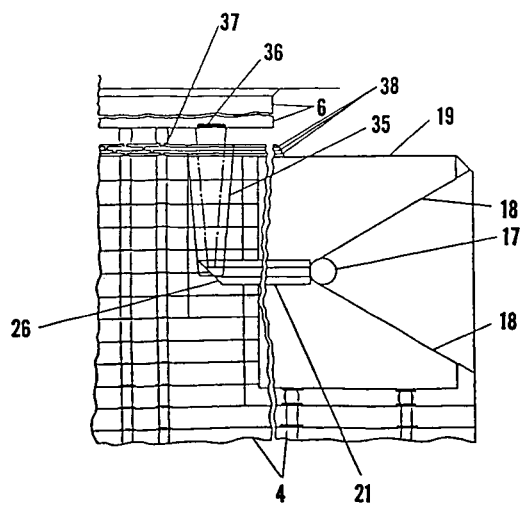


FIGURE 6.17 Design of an embedded receptacle for a parallel fiber-optic interface to an electronic module [77].

spaced around the perimeter of an MCM, thousands of fibers could be used to form a high bandwidth data bus that dramatically increases the I/O of the MCM without introducing additional noise and crosstalk problems. Designs such as this can also be modified to accommodate miniature coaxial cables or hybrid interconnection arrays. Optical fiber MCM interfaces represent one possible way to address emerging requirements for high bandwidth, large I/O count multichip packages.

The use of parallel fiber interconnects is not limited to MCM packaging; other designs have been proposed for multifiber connectors, such as the MAC connector from AT&T and the MT connector from U.S. Conec. As cable plants continue to grow in complexity, cable management becomes a significant problem; architectural restrictions such as fire walls and building codes also limit the number and type of cables that can be installed in a building. It is therefore desirable to bring many fibers together in a small space; multifiber connectors accomplish this while maintaining fiber alignment by using a larger alignment feature, such as a set of pins, to guide the fiber connectors together. In this way, only the guide pins must have critical tolerances; the rest of the connector is relatively inexpensive. The MT connector has been proposed as an industry standard and is already in use by several telecommunication carriers in Japan. Multifiber interconnects are thus expected to play an increasingly important role both within and between electronic circuit packages in the future.

6.5 ATM/SONET: Emerging Telecom/Datacom Standard

In recent years, there has been great progress made toward realizing broadband optical networks through the establishment of international standards [16, 58]. Standardization is a critical factor in future network development, as it ensures that different types of data communication equipment will interconnect and interoperate on some basic level while providing a path for growth and migration to higher performance systems. Early data communication networks such as fiber-in-the-loop (FITL) faced the challenge of initially providing narrowband services (about 1 Mbit/s) while providing an architecture that could evolve to enable more advanced services at higher data rates. Future systems will also require that different types of voice, data, and digital signals be processed efficiently on the network; this implies sharable and scaleable bandwidth requirements. Although many existing systems have partially addressed these problems and will continue to be in use for many years, there is a need to move network services to a more comprehensive platform. At this point, it is difficult to speculate on how these services will evolve and how they will be introduced into the network architecture; there is considerable work ongoing to explore alternatives such as hybrid fiber/coax systems or gateway converters that interconnect fiber loops designed around the different standards presented in the previous sections.

There is, however, growing acceptance of one preferred link embodiment, namely, the link protocol of asynchronous transfer mode (ATM) operating over a physical layer based on the synchronous optical network (SONET) standard (the apparent contradiction of a synchronous physical layer with asynchronous transmission protocol will be explained shortly). These have already been embraced by the telecommunication industry and are being accepted (with a strong sense of urgency) by computer and LAN providers as the basis for future broadband integrated services digital network (B-ISDN) applications. A number of suppliers have already announced ATM support for network products available today, and many more plan to have products available in the next 5 to 10 years. Although ATM/SONET is not limited to fiber-optic installations (indeed, ATM is not even a physical layer protocol), these emerging standards deserve consideration as they will probably shape the next generation of optical data link technology. In this section, we review the basic technology behind ATM and SONET and discuss their potential advantages and implications for developing networks. For a more detailed treatment of these standards, the reader is referred to both the standards themselves and a wide range of literature available in this area [78–80]. This section is intended to provide the necessary background for reading and understanding these emerging standards. Because of their large potential impact on both telecommunications and data communications, we will devote some time to explaining the higher protocols and frame structure of ATM transmission.

The ATM technology may be thought of as the ultimate evolution of packet switching, although it incorporates aspects of both circuit and packet switching. All binary data, regardless of the source (video, voice, digital, etc.) is broken down into the fundamental unit of a 53-byte cell, as shown in Figure 6.18; the cell may be viewed as a short fixed-length packet. Each byte is 8 bits long; 5 bytes constitute the cell header and contain routing and control

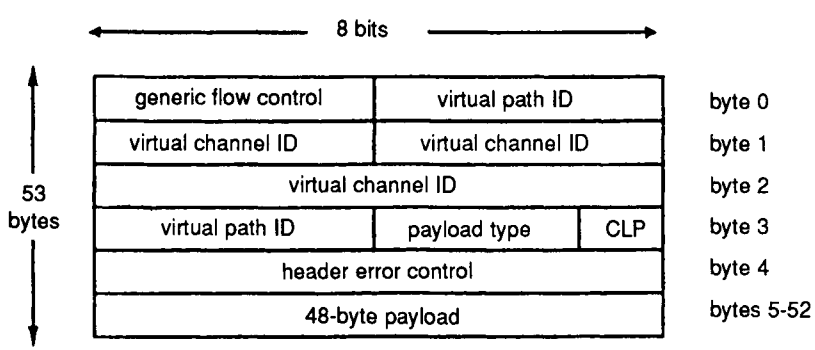


FIGURE 6.18 The structure of an ATM data cell.

information, providing a 48-byte data payload. The header structure and function have been standardized by the International Consultative Committee on Telegraph and Telephone (CCITT) [78]. The header information consists of user access and control information, an address routing field, a field to identify the type of payload, an error-correcting code for header information, and a cell-loss priority. These fixed-length cells are multiplexed onto the physical transmission media, for example, optical fiber using SONET. One interface can support many connections and each connection uses only the number of cells required to transmit its data. In this way, link bandwidth is allocated independently to each connection.

Routing information (cell addresses) are contained in the virtual path and virtual channel ID fields in the header. This routing information establishes a logical connection between any two nodes (users) on the network. Before two nodes can communicate, a connection between them must be established which may pass through many intermediate nodes. As the path is established, each node contributes to forming a virtual channel indicator which is placed in the cell header. Once a path is established and data transfer begins, the virtual channel header is used at each intermediate node to switch incoming cells to the appropriate outgoing link; the process is repeated at each node, and the connection is carried forward until it reaches its destination. Note that the label addressing allows flexible use of link bandwidth for different kinds of data, characteristic of packet switching. At the same time, the fixed cell length allows the use of synchronous switching and multiplexing techniques. This combination makes it possible to provide simple implementations of high speed link adapters and hardware.

Other information in the header includes the payload type field, which shows whether link traffic belongs to a user or to the link maintenance system; this feeds into higher protocol levels and is also used to signal congestion in the network. Information in the ATM header is verified by a cyclic redundancy check (CRC) at each intermediate node as the cell travels through the network. If an intermediate node detects an error, the cell is discarded without notification to either of the users. For some types of data, such as still images, this can cause unacceptable distortion, as the effect is cumulative over successive scan lines in the image. Because the cells are multiplexed over the link, they may also arrive in the wrong order. For some kinds of link traffic, higher levels of the protocol may check the data payload for errors as well. Error checking of the header is one of the few functions performed by the intermediate nodes; usually, ATM eliminates the involvement of intermediate nodes once the virtual channel is established. Given the performance, reliability, and low BER of today's fiber-optic links, it is considered unnecessary overhead to replicate link performance checks at each node. The advantage of this approach is that it enables straightforward all-hardware implementation of ATM processing so that recent advances in high speed semiconductors can directly benefit the channel; software overhead for data flow control at

every node is eliminated. For some applications, such as backup of critical computer data, this overhead reduction and adoption of unacknowledged transmission is a potential weakness of ATM as compared with more robust channel protocols, such as FCS. Cell-loss priority in ATM is defined by a separate field in the header. An ATM adaptation layer usually resides above the ATM protocol layer; this performs higher order functions such as matching the ATM channel characteristics to the type of link traffic (reassembly of variable-length data into fixed-length ATM cells, etc.). Details of these protocols are beyond the scope of our current discussion.

Cells are generated asynchronously; as the data source provides enough information to fill a cell, it is placed in the next available cell slot. There is no fixed relationship between the cells and a temporal frame reference, as in conventional time division multiplexing schemes. The flow of cells is driven by the bandwidth needs of the source, rather than by a master clock. If the source data comes in bursts, as in many client/server applications, several sources may share a common link by multiplexing data from one source during the idle intervals of other sources. This flexibility allows ATM to support both constant and variable bit rate services, such as switched multimegabit data service (SMDS) and frame relay, through the ATM adaptation layer. This is an effective method of multiplexing and switching without requiring that bandwidth be allocated in fixed multiples. ATM provides bandwidth on demand, covering the range of 25.6 to 2488 Mbit/s. ATM speeds were initially intended to match SONET rates of 51, 155, and 622 Mbit/s, with an FDDI-compliant rate of 100 Mbit/s. In order to provide compatibility with existing installed copper channels, including some telecommunication applications and 16-Mbit/s token ring LANs, a 25-Mbit/s speed has recently been advocated by several ATM chip manufacturers as a proposed addition to the industry standard. Used with an effective switch fabric, this provides efficient utilization of transmission equipment and allows the platform to be scaled up to higher data rates in the future. The combination of data transmission options available is sometimes described as a pleisosynchronous operation, meaning that it combines some features of multiplexing architectures without requiring a fully synchronous implementation. Such an architecture is well suited to the requirements of multimedia, distance-independent transmission over LANs, MANs, and WANs, and connectivity between workstations, controllers, and routers.

Asynchronous transfer mode technology can operate over many physical layers; it is currently planned for the telecommunication standards DS1 and DS3, which form the backbone of the U.S. phone network. It is generally agreed, however, that the preferred embodiment will use the SONET transmission hierarchy developed for telecommunications [79]; this is defined by ANSI standard T1.105-1988 and T1.106-1988, as well as by several CCITT recommendations [81]. The ATM transmission speeds are well matched to multiples of the allowed SONET data rates, whereas other

standards would require additional data rate conversions for speed matching. SONET is expected to be incorporated into existing networks where it will displace existing point-to-point implementations in the common carrier network; eventually, it should allow the carriers to provide a range of new B-ISDN services and facilitate the merging of voice and computer networks. A number of commercial products are currently available, or have been announced, supporting the SONET protocol.

SONET is really a family of standards covering data rates, formats, and optical parameters for high speed transmission. It ensures that equipment is interoperable in the sense that data can be exchanged between any two devices conforming to the standard. A standardized interface is provided for SM fiber at data rates from 51.84 Mbit/s (OC-1) up to 2488.32 Mbit/s (OC-48). The standard provides for six intermediate data rates, OC-3, OC-9, OC-12, OC-18, OC-24, and OC-36, for which the numerical part of the OC-level designation indicates a multiple of the fundamental 51.84 Mbit/s data rate. SONET also contains provisions to carry sub-OC-1 data rates, compatible with existing telecommunications equipment; these are called virtual tributaries (VT). They support transmission at 1.544 Mbit/s (DS-1), 2.048 Mbit/s (CEPT-1), 3.152 Mbit/s (DS1C), and 6.312 Mbit/s (DS-2); the virtual tributaries are designated VT1.5, VT2, VT3, and VT6, respectively. The SONET physical layer optical specifications for various line rates are summarized in Table 6.3.

The basic SONET frame structure is shown in Figure 6.19; it consists of an array of 9 rows, with 90 bytes per row. This block is known as the synchronous transport signal-level 1 (STS-1) frame. In an OC-1 system, the STS-1 frame is transmitted once every 125 μs (810 bytes per 125 μs yields the 51.84 Mbit/s data rate). The first three columns are overhead which provide functions including identification, framing, protection switching information, error checking, and an embedded operations channel operating at 768 kbit/s. Note that the data payload may be split across two consecutive 125-μs intervals; part of the overhead is a pointer that identifies the start of the payload envelope with respect to the overall framing structure. Because the payload floats within the STS-1 frame, there is no need for a 125-microsecond buffer at the receiving end of the link; the rules for pointers are quite complex (they also allow for the different VT options) and are described more fully in Reference 79.

The remaining 87 bytes of an STS-1 frame are data payload; 87 bytes per 9 rows per 125 microseconds yields 50.112 Mbit/s effective data throughput rate. The full payload, or 87 bytes, supports a single DS-3 signal; VT1.5 occupies 3 columns, or 1.728 Mbit/s, to accommodate DS-1, and similarly VT2 occupies 4 columns, VT3 six columns, and VT6 12 columns. Higher speeds can be achieved either by concatenation of STS-1 frames or byte-interleaved multiplexing of these frames. For example, an STS-Nc frame is formed by an N by 90 byte and 9-row array (the c stands for concatenated).

TABLE 6.3 Physical layer optical specification for various SONET line rates, intermediate length.

	Line Rate								
Parameter	**OC-1**	**OC-3**	**OC-9**	**OC-12**	**OC-18**	**OC-24**	**OC-36**	**OC-48**	**Unit**
Transmitter type	MLM/LED	MLM/LED	MLM/LED	MLM/LED	MLM	MLM	MLM	MLM	
$\lambda_{Tmin} - \lambda_{Tmax}$	1260–1360	1260–1360	1260–1360	1260–1360	1260–1360	1260–1360	1260–1360	1265–1360	nm
$\Delta\lambda_{rms}$	80	40/80	19/45	14.5/35	9.5	7	4.8	4	nm
$\Delta\lambda_{20}$	NA	NA	NA	NA	NA	NA	NA	NA	nm
SSR_{min}	NA	NA	NA	NA	NA	NA	NA	NA	dB
P_{Tmax}	−14	−8	−8	−8	−8	−5	−3	−3	dBm
P_{Tmin}	−23	−15	−15	−15	−15	−12	−10	−10	dBm
r_{emin}	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	dB
Optical Path									
System ORL_{max}	NA	NA	NA	NA	20	20	24	24	dB
D_{SRmax}	NA	NA	13/NA	13/NA	13	13	13	12	ps/nm
Attenuation	0–7	0–7	0–7	0–7	0–7	0–7	0–7	0–7	dB
Max. Reflectance between S and R	NA	NA	NA	NA	−25	−25	−27	−27	dB
Receiver									
P_{Rmax}	−14	−8	−8	−8	−8	−5	−3	−3	dBm
P_{Rmin}	−31	−23	−23	−23	−23	−20	−18	−18	dBm
P_O	1	1	1	1	1	1	1	1	dB

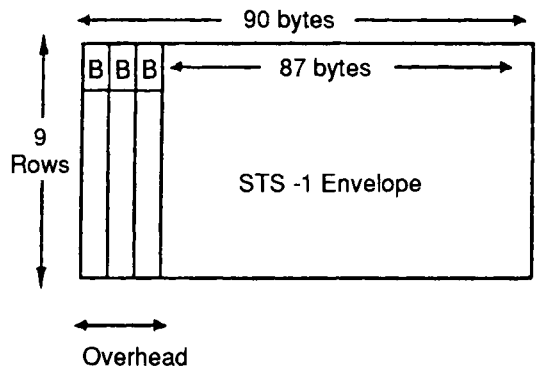


FIGURE 6.19 The structure of a SONET STS-1 frame. B denotes byte or octet.

The first three columns of an STS-Nc frame are taken up by overhead and the remainder provides a wide data pipe for high speed transmission. A nonconcatenated byte multiplexed STS-N frame is formed by taking the individual overhead of N frames of STS-1 data and interleaving the bytes to form the first N by 3 columns of a new data frame (which is still 9 rows wide). Independent payloads from the N individual frames are combined in byte-interleaved form to make up the payload of an STS-N frame.

Some typical examples of the SONET link length limitations are given in Table 6.4. The interested reader may wish to use this standard as a guide to apply some of the link budget models discussed earlier in this chapter and ascertain the maximum unrepeated link spans allowed by SONET. The strength of ATM/SONET is its versatility in providing a wide range of applications, particularly those which are bandwidth-intensive, such as video servers, channel extenders, and the central office LAN/WAN. Other applications, such as still-picture transfer for medical images or document records, can also be used on a B-ISDN link. The current interest in and proliferation of ATM/SONET are expected to set a significant technology direction for the next decade. In the following section, we discuss other potential applications and technologies that may make significant contributions to the future of data communications.

Finally, we must mention a recent proposal by IBM to extend the performance of current ATM technology. Known as packet transfer mode (PTM), this proposed standard is really a superset of the ATM requirements [80]. PTM addresses certain weaknesses in the ATM standard, such as the lack of true packetized voice and video standards, multipoint connection problems, and a full internal switch signaling protocol. This allows variable length frames to be sent over an ATM network, which was originally designed for fixed-length data frames and works best in constant bit rate applications. PTM can reduce the overhead on data packets by 20 to 50% (although voice

TABLE 6.4 Length limits due to dispersion and attenuation of typical SONET links [106]. Copyright 1993 IEEE.

Bitrate (Gbit/s)	Dispersion Limits[a] (km)			Attenuation Limits[b] (km)	
	1550 nm		1300 nm		
	(SMF)	(DSF)	(SMF)	1550 nm	1300 nm
10	58	283	400	47	31
20	14.5	70	100	37	34
40	3.6	18	25	27	18

[a] Dispersioin limitations for 1 dB eye closure penalty /7/:
Assumed dispersion: DSF (1550 nm), SMF (1300 nm): D = 3.5 ps/nm/km
SMF (1550 nm): D = 17 ps/nm/km

[b] Attenuation limitations calculated with following assumptions:

Optical transmitter power:	−3 dBm
Receiver sensitivities:	−20 dBm (at 10 Gbit/s)
	−17 dBm (at 20 Gbit/s)
	−14 dBm (at 40 Gbit/s)
Fiber attentuation:	0.3 dB (at 1550 nm)
(incl. splices/connectors)	0.45 dB (at 1300 nm)
System margin	3 dB

traffic does not realize these gains). Because the PTM has no defined upper limit on frame size, an ATM frame can be embedded within a PTM frame, allowing the two to coexist on a single network. However, PTM requires some extra framing information, such as bits that define the size of a data cell; thus, all PTM networks can accommodate ATM, no ATM network can adapt to PTM. Currently, PTM is a proprietary standard that has not received strong support from the data communication and telecommunication industries; it remains to be seen whether PTM will be adopted as a complement to existing ATM standards.

6.6 Emerging Applications and Technologies

Increasing demands are being placed on existing data and telecommunication systems by both operators and customers. These include the demand for providing more broadband services, flexible allocation of system bandwidth, intelligent network control, and built-in high security, reliability, and channel performance. Future optical data links and networks will have to satisfy the needs of large and diverse markets, including traditional data and voice communication, client-server and networked computer systems, and emerging markets such as interactive and on-demand video service. These trends are driving the development of higher capacity fiber-optic systems; at the same time, system developers are migrating toward hybrid copper/fiber

channels for near-term applications that can exploit both the high performance of optics and the relatively well-established coaxial transmission lines already installed in many locations. For applications in terrestrial and undersea lines, optical fiber has been accepted over alternatives such as microwave and satellite because of its high bandwidth and low overall cost. Because there are many different types of data and voice communication channels available today, system planners are faced with the task of integrating their existing systems with new product offerings. Realistically, the evolution of data communications must move away from the small, proprietary systems of today, each operating with a different protocol, toward more open system architectures that interconnect and interoperate with larger computer networks. Just as the data and voice communication systems are merging, so is the distinction between high and low range products beginning to blur as new applications cross traditional product lines.

Next generation computer systems are expected to make increasing use of fiber optics. Worldwide acceptance of fiber-optic solutions has been due in part to the general acceptance of standards such as ATM for digital access and switching protocols and SONET for link transport. A large number of vendors have already announced their support for this strategic direction in fiber-optic systems development. Additionally, government agencies are recognizing the importance of an information infrastructure for their future competitiveness. In the United States, agencies such as the National Science Foundation (NSF), Advanced Projects Research Agency (ARPA, formerly a defense-related contractor), and the National Institute for Standards and Technology (NIST, formerly the National Bureau of Standards) have begun to support research aimed at future broadband networks. Special projects such as the National Research and Education Network (NREN) and planned upgrades to the existing electronic mail facilities of the Internet are just part of the efforts to develop a national information infrastructure (NII), the so-called "information superhighway" which is intended to become a nationwide data communication network. Similar projects are underway in both Europe (the RACE program) and Japan (currently unnamed). It is clear that policy makers expect digital communication to play a major role in national productivity, and fiber optics is a key enabling technology to realize this vision. Optical communications should continue to expand in response to both the promise of new applications and the push of emerging technologies. In this section, we discuss both emerging technologies that increase the performance of fiber optics and some of the applications that will drive the field in years to come.

6.6.1 High Data Rates and Wavelength Multiplexing

One of the critical areas for future development is higher data rate channels; recall that even the highest speed existing systems (2 to 3 Gbit/s)

utilize only about 1% of the 50 THz bandwidth theoretically available in an optical fiber link. State-of-the-art technologies have been explored in so-called "gigabit testbeds" partially sponsored by the government for applications such as the High Performance Computing and Communication (HPCC) initiative. As we have seen in previous sections, gigabit transmission rates are becoming increasingly common in commercial products that are available today. Because of the strong increases in traffic, 2.5-Gbit/s systems have already been deployed (STM-16, OC-48) by telecommunication companies for interoffice and long haul links. Field trials are currently being conducted on the next step in the digital hierarchy, 10-Gbit/s systems (STM-64 or OC-192). These products could be available in a few years for some applications. In research laboratories, systems with 100-Gbit/s capacity and beyond are being explored today, and there are several "hero experiments" in the literature of this high speed, long distance transmission [82–84].

To achieve higher data rates, our first consideration concerns a suitable optical source. The simplest modulation technique involves intensity modulation of a semiconductor laser diode by controlling the injection current; indeed, this approach is commonly used in 1-Gbit/s applications today. Laser diodes have been demonstrated in the laboratory that have modulation rates of 10 Gbit/s [85], and sample links have been operated with bit rates as high as 16 Gbit/s [86]. This approach is limited by spectral line width broadening of the laser caused by chirping effects; the mechanism is the same as the one mentioned earlier in connection with nonlinear optical noise effects. This spectral broadening causes dispersion problems in the link; additional penalties of 1 dB have been observed on 10-Gbit/s links after about 3.5 km, as compared with perhaps 0.1 dB penalty at lower bit rates [86]. This limitation leads to the investigation of other modulation techniques. For example, by phase modulating the laser dioide and using an interferometer to convert from phase to intensity modulation, DC modulation up to 20 Gbit/s has been reported [82].

The chirping problem can also be reduced by using external cavity modulators, such as electro-optic or acousto-optic interferometers. These can be fabricated as integrated optics devices in crystals such as lithium niobate. Although commercial products have been demonstrated, performance is typically limited by the need for complex, high power electronic drives (on the order of 5 to 10 V) and the high optical insertion loss (3 to 10 dB in some cases). The modulation can also be done using multiple quantum well or GaAs devices, although less efficiently; lasers with monolithically integrated modulators in these materials are under development to address the insertion loss problems [87]. The modulator voltages, however, are not within the range of integrated driver circuits, and hybrid or discrete circuit drivers remain the only practical alternative. The lack of integrated laser drivers could be a serious obstacle to further development of this approach. Another

alternative is the use of gain-switched or mode-locked lasers, which take advantage of optical pulse compression techniques [88]. Even using relatively low bandwidth, low voltage laser drivers, optical pulses as short as 6 ps have been produced [88] using gain switching (for a discussion of the physics of gain switching and mode locking, see Ref. 89). Mode locking techniques have even greater potential, as their use of an external laser cavity almost eliminates the effects of laser chirping. Pulse widths in the range 1.2 to 15 ps have been reported, and mode-locked laser pulse trains are being used today for laboratory experiments up to 100 Gbit/s [90]. Near transform-limited pulses can be produced by this method [91].

Operation of high speed channels also requires high speed electronics to support digital multiplexing (serializing/deserializing of data), decision circuits, and other signal processing functions. Realization of these circuits has been demonstrated in the laboratory for 10-Gbit/s systems. For example, using silicon bipolar technology with 0.8-μm line widths, a 2:1 multiplexer, decision circuit, frequency divider, and demultiplexer have all been demonstrated to at least 25 GHz [92]. Similar results have been published using AlGaAs/GaAs technology [93] and Si/SiGe material systems [94]. Given recent prototype demonstrations, there seems to be no reason why these technologies cannot be extended to more than 50 Gbit/s. Hybrid analog circuits with discrete optoelectronic components have also been used as low-noise and high-gain amplifiers and clock recovery circuits up to 10 to 15 Gbits [95]. All of the essential building blocks for high speed electronics supporting digital fiber-optic links have been demonstrated and should be available in the near future to support the requirements for such links.

One alternative to developing higher speed modulators is the use of multiplexing techniques to combine several lower data rate signals into a high bandwidth channel. This approach has the advantage of not requiring high speed modulators to be installed for every device on a network; most applications will tolerate much slower speeds at the device interface compared with the requirements on the data stream. There is a wide range of multiplexing approaches available; one example is time-division multiplexing. Using optical switches or star couplers, several lower data rate optical pulse trains can be multiplexed to create higher data rate channels without requiring the associated electronics to operate at the higher data rate. Although this method can overcome the need for high speed laser modulation, it is interesting to speculate on combining this approach with high speed optical modulators; a train of 10-GHz pulses could be multiplexed to create a 40 to 50 Gbit/s operation. Indeed, multiplexing is probably the only way to reach such speeds in the near future. Considerable development work remains to be done before assessing the cost trade-offs of high speed sources versus multiplexers for different applications.

One approach which deserves special attention is wavelength-division multiplexing (WDM), or optical frequency division multiplexing (OFDM).

Because optical signals at different wavelengths do not interfere with each other, it is possible to send several data streams along the same fiber, each at a different wavelength. This can be achieved using either a frequency-tunable laser source or several lasers at slightly different wavelengths. The received signals can be separated again using passive star couplers and optical filters, or tunable optical receivers. This approach is seen by many as the preferred path to realizing the full 50-THz potential of the optical channel; note that the various optical amplification techniques to be discussed shortly have sufficiently large bandwidth to be fully compatible with most WDM techniques. Currently, optical networks use a single wavelength, which leads to the requirement for high bit rates if the network services many customers, each with a high bandwidth requirement. An alternative is the use of many different wavelengths, each assigned to a different customer or node on the network. The number of nodes is limited only by the spacing of adjacent wavelength channels and the selectivity of the receivers. Transmission of TV signals to hundreds of customers using this technology has been in progress in Europe for more than a year as a technology testbed [91]. The analogous situation in computer networks has great potential, as does the extension of this system to concepts such as high definition television (HDTV) and fiber-to-the-home services (FTTH). This approach requires active components for tunable receivers. For some time, experimental systems have also been in operation using passive components such as gratings. One example is the demonstration WDM link operated by IBM in Hawthorne, New York. This has led to one of the first WDM products in the data communication industry, the Muxmaster high speed data multiplexer. This product uses wavelength-tunable distributed feedback lasers and diffraction gratings to combine light at different wavelengths near the 1550 nm transmission window. A maximum of 10 full duplex (bidirectional) channels may be provided, each at a slightly different wavelength. By placing a Muxmaster at either end of a fiber link, multiplexing occurs independent of the data stream or protocols at either end. Distances of up to 50 km and speeds of up to 1.2 Gbit/s can be provided, and different protocols can be sent on the same link without interference. The main obstacle to wider adoption of this technology is the cost of multiplexers and demultiplexers rather than the performance; with increasing market demand, WDM systems are expected to play a vital role in future high speed applications.

6.6.2 Optical Amplifiers

It has been demonstrated [94] that bipolar silicon technology is capable of multiplexing and other signal processing operations up to about 40 Gbit/s. Electronic time division multiplexing can probably be used in 10-Gbit/s systems; other techniques such as WDM or frequency-division multiplexing (FDM) are also options for increasing speed. High speed alone is not

sufficient, however; a 100-Gbit/s link that is limited to less than a meter has very limited application, except possibly as a component of an intrasystem data bus. The real goal is to increase the bandwidth-distance product of the link; we recognize that multigigabit modulation is not a very efficient optical to electrical (O/E) conversion process, and the signal degrades rapidly with distance. In existing long-haul links, full 3R signal regeneration (reshaping, retiming, and reclocking) must be performed about every 40 km; it would be cost effective to reduce the number of repeaters required for such distances. Of course, this is a link budget design issue; it becomes increasingly difficult to guarantee performance for long distance, high speed links while still taking into account all manufacturing, aging, and other tolerances. In this section, we discuss some of the limitations on increasing unrepeated distances beyond 40 km.

The main physical limitations for high speed, long haul data link results from the transmission properties of optical fiber—attenuation and dispersion. First, consider the limitations of fiber attenuation; nearly all long distance systems previously installed use SM fiber operating near 1300 nm. Theoretically, optical fiber loss can be reduced to about 0.05 dB/km, which is the Raleigh scattering limit (recall that Raleigh scattering loss is proportional to the inverse fourth power of wavelength). The existing installed fiber base, however, has a typical loss of 0.2 to 0.5 dB/km. To interoperate with this fiber, some form of amplification that does not require O/E conversion is required. By selectively doping optical fiber, it is possible to obtain amplification of optical signals at one wavelength by pumping the fiber with light at another wavelength, similar to an optically pumped laser. These are known as optical fiber amplifiers; they allow the data signal to be amplified without performing O/E conversion, and thus hold great potential for relieving power budget constraints on the distance of an optical link. We will not attempt a detailed treatment of fiber amplifiers here; there are many references available on the subject [97, 98].

Doped fiber amplifiers for the 1300-nm transmission window use either praesmodymium [99] or neodymium [100]. These are still topics of research, with great potential but currently limited performance; additionally, their preparation involves the use of highly toxic dopants. Nonlinear optical amplifiers using either the stimulated Raman or Brillouin effects discussed earlier have been demonstrated [101]. Although they provide broadband amplification, these approaches typically require high optical pump power or precise control of the pump wavelength; it is therefore questionable whether they will ever be practical for field applications. The high power required (on the order of 500 mW for stimulated Raman scattering, 1 W for Brillouin) are also not likely to be reached by higher bit rate systems. Semiconductor optical amplifiers (similar to traveling wave optical amplifiers or lasers) can also be used to amplify optical signals, and although some are commercially available, they exhibit undesirable properties such as spectral broadening, gain ripples,

polarization sensitivity, and fairly large insertion loss. The advantages of semiconductor amplifiers is their broad bandwidth and ability to work with wavelengths throughout the 1300 and 1550 transmission windows. There is ongoing work that may solve the problems associated with these amplifiers [102], which would make them an attractive option for the 1300-nm window. By far the most mature optical amplification technology operates in the 1550-nm window, using erbium-doped fiber amplifiers [97, 98]. Indeed, the advent of these amplifiers has made the 1550-nm window considerably more important for future long distance applications. Erbium-doped amplifiers are commercially available and offer previously unheard of increases in transmission distance; recently, a cascade of 274 in-line optical amplifiers was demonstrated in a 10-Gbit/s link over 9000 km long, yielding a record value for the bit rate distance product of 90 Tbit/s-km [103]. More typically, a single erbium-doped amplifier provides enough gain to increase the unrepeated distance to about 80 km in a point-to-point link, and recent experiments providing from 1 to 5 Tbit/s-km have been reported [104]. Amplification also makes it possible to distribute a signal to many nodes on a network; fiber-optic distribution of TV signals at multigigabit rates to over 8 000 000 nodes has been demonstrated [105]. From this work, it can be seen that optical amplifiers virtually eliminate the link budget restrictions, even in multigigabit systems. In summary, these optical amplifiers offer the following advantages:

- amplification that is independent of the bit rate and light polarization;
- broadband amplification (1530 to 1560 nm), which facilitates the introduction of WDM techniques:
- high output power (typically about 15 dBm);
- high small-signal gain (typically 30 dB); and
- low noise figure (3 to 5 dB) and little spectral broadening, allowing amplifiers to be cascaded or used as receiver pre-amps.

6.6.3 Solitons

Optical fiber amplifiers operate near the dispersion minimum of the optical fiber; this is important, because after attenuation is overcome, dispersion becomes a limiting factor on distance. This was illustrated in Table 6.4 [106], which shows the dispersion and attenuation limits for a typical link configuration at 1300 and 1550 nm; note the importance of dispersion limits at higher data rates. There are several approaches to reducing dispersion limits in a system. One approach [107] involves transmitting a frequency-modulated signal, which is converted into amplitude modulation by dispersion in the optical fiber. This so-called dispersion-supported transmission (DST) actually required a certain amount of fiber dispersion in order to operate properly. Rather than taking advantage of existing dispersion, most other techniques

attempt to eliminate its effects. As discussed earlier, the line width broadening of a chirped semiconductor laser can be overcome by various external modulation techniques; an optical signal that is not chirped suffers less dispersion penalty. Indeed, by using a combination of prechirped laser light and fiber with the proper nonlinear refractive index, it is possible to obtain a negative dispersion penalty; this is the same principle that is used in optical pulse compression. Basically, at high optical power levels, the fiber's self-phase modulation counteracts the laser chirp and compensates for dispersive pulse broadening; stable optical pulses, known as solitons, result which can propagate without suffering dispersion. A full mathematical treatment of solitons and their properties is beyond the scope of this chapter, several excellent references are available [108–111]. Soliton generation, then, is a nonlinear optical process that requires fairly high optical power levels to be initiated. The required optical power levels can be obtained with erbium-doped amplifiers, and experiments with soliton propagation on dispersion-shifted fiber have demonstrated multigigabit transmission over hundreds of kilometers [112].

6.6.4 Future Data Communication Applications

There are many other potential applications for which fiber optics offers advantages over existing data communication solutions. The application of massive parallelism to relatively low-cost processors that is currently being used in mainframes has also been adopted by many supercomputer architectures. Although most commercial supercomputers achieve parallelism by using shorter distance interconnections at lower speeds, future growth of these systems is expected to require increases in speed and distance between processors. As cross talk and electrical noise limits are approached and optical transceivers become increasingly cost competitive with copper, supercomputer designs are also expected to adopt fiber interconnections to solve data processing bottleneck problems. The interconnection network is critical to the success of massively parallel architectures, as it provides the means for processors, memory, and I/O devices to communicate with each other. Recent research into the protocols and requirements for these systems [113] indicates that the software problems associated with large scale parallelism, including message contention and efficient bandwidth utilization, can be overcome by using novel packet switching and routing techniques. Thus, although the problems associated with large distributed architectures should not be minimized, they also should not be an obstacle to the adoption of advanced fiber-optic hardware.

It is interesting to speculate on the future of such machines; as current mainframes experience decreasing cost/performance ratios, supercomputers based on highly parallel workstations are finding applications in many commercial areas as well as in the conventional scientific/technical

applications. There is an ongoing need for the performance, security, and reliability of the mainframe, which implies that this class of machine will continue to dominate data processing environments well into the next decade. Mainframes are also evolving to new parallel processor architectures, which is expected to improve significantly the mainframe competitiveness and extend their usefulness. At some point, the parallel architecture of the mainframe and supercomputer may be sufficiently similar that supercomputers supplement or replace the mainframe in many applications; if this occurs, the need for high bandwidth data links in large (128 or more node) scaleable parallel machines will only become more critical. Furthermore, if the nodes in a future supercomputer are simply low-cost RISC-type processors, it may be possible to interconnect nodes over hundreds of meters or even kilometers and form a distributed, parallel coupled architecture. Many commercial offices already have networks installed to share data among their workstations; future offices may be able to convert this connectivity into a supercomputer when necessary, and back into a shared client/server LAN for different applications requiring multiple users. In either case, high bandwidth interconnects are expected to play a critical role in the developing supercomputer market over the next decade.

Another possible future application is optical data buses within the machine rather than between machines; optical interconnects have been discussed and researched for some time and there are several commercial applications being developed that will take advantage of this technology. We have already discussed the potential of optical clocking and optical data buses, as well as some of the European companies that are planning to commercialize this technology [71, 72]. Current applications are limited, in part, by processor electronics; few digital designers have a use for a 500-MHz clock signal. As processors, multichip integration, and packaging become more sophisticated [77], optics is anticipated to become important to these designs.

In general, fiber-optic technology will be driven by developing applications such as multimedia, which plans to deliver a mixture of high fidelity motion pictures and sound quality to the desktop. Advanced client/server applications are becoming more common in large data processing facilities such as banks and credit card companies; these, too, will demand increasing data throughput in the future. Following the deregulation of the telecommunication industry, many cable television operators are also installing large fiber networks, and are planning to offer a mix of interactive services, home shopping, and movies on demand. In addition, other bandwidth-intensive services have emerged, including broadband integrated services digital networks (B-ISDN) at speeds of 150 Mbit/s and higher. As LANs grow and interconnect more frequently, backbone services such as switched multimegabit data services (SMDS) become more important; in MANs and WANs, techniques called fast packet switching are also emerging. These include high speed cell relay, as used in ATM and the IEEE 802.6 protocol

for WANs, as well as frame relay techniques that operate on variable-length data frames, such as the CCITT 1.122 interface recommendation. Large volume applications such as these should drive down the cost of current technology and fuel interest in future network designs based entirely on optical fiber. Although fiber will probably never replace coaxial cable or other media, it will enable new applications that might not have been possible otherwise, and by so doing, will be a significant factor in the deployment of future data communication systems.

References

1. S. E. Miller and I. P. Kaminow (Eds.) (1988). *Optical Fiber Telecommunications II.* New York: Academic Press.
2. P. Green (1993). *Fiber Optic Networks.* Englewood Cliffs, NJ: Prentice-Hall.
3. S. S. Walker (1986). Rapid modeling and estimation of total spectral loss in optical fibers. *IEEE Journal of Lightwave Technology* 4, 1125–1131.
4. F. C. Allard (1990). *Fiber Optics Handbook for Engineers and Scientists.* New York: McGraw-Hill.
5. E. J. Frieble, E. W. Taylor, G. T. De Beaureagard, J. A. Well, and C. E. Barnes (1988). Interlaboratory comparison of radiation-induced attenuation in optical fibers (Part I). *IEEE Journal of Lightwave Technology* **6**, 165–171. (Part II) (1990) *IEEE Journal of Lightwave Technology* **8**, 967–988.
6. E. J. Frieble, C. G. Askins, and M. E. Gingerich (1984). Effect of low dose rate irradiation on doped silica core optical fibers. *App. Opt.* **23**, 4202–4208.
7. E. J. Frieble and M. E. Gingrich (1981). Photobleaching effects in optical fiber waveguides. *App. Opt.* **20**, 3448–3452.
8. T. E. Tsai and H. B. Lin (1990). Photoluminescence induced by 6.4 cV photons in high purity silica. *Proc. Mat. Res. Soc. Symposium* **172**, 125–131.
9. J. B. Haber, E. Mies, J. R. Simpson, and S. Wong (1989). Assessment of radiation-induced loss for AT&T fiber optic transmission systems in the terrestrial environment. *IEEE Journal of Lightwave Technology* **6**, 150–154.
10. N. A. Amos, A. D. Bross, and M. C. Lundin (1990). Optical attenuation length measurements of scintillating fibers. *Nucl. Inst. Methods* **A297**, 396–403.
11. M. Kyoto (1992). Gamma ray radiation-hardened properties of pure silica core single-mode fiber and its data link system in radioactive environments. *IEEE Journal of Lightwave Technology* **10**, 289–294.
12. M. Young (1973). Geometric theory of multimode optical fiber-to-fiber connections. *Opt. Comm.* **7**, 253–255.
13. D. L. Bisbee (1971). Measurements of loss due to offsets and end separation of optical fibers. *Bell System Technical Journal* **49**, 3159–3168.
14. T. C. Chu and A. R. McCormick (1978). Measurements of loss due to offside, end separation, and angular misalignment in graded index optical fibers excited by an incoherent source. *Bell System Technical Journal* **57**, 595–602.
15. D. Gloge (1975). Propagation effects in optical fibers. *IEEE Trans. Microwave Theory and Tech.* **MTT-23**, pp. 106–120.
16. Fiber Channel Standard. Physical layer and signaling specification. American National Standards Institute X3T9/91-062, X3T9.3/92-001 (rev. 4.2).
17. P. M. Shanker (1988). Effect of modal noise on single-mode fiber optic networks. *Opt. Comm.* **64**, 347–350.

18. *Fiber Optic Test Procedure FOTP-171.* Attenuation measurement by substitution method (also EIA-TIA-455-171). Available from Electronics Industry Association, New York.
19. H. Kosiorska, B. Wong, and C. Saravanos (1991). Statistical method for predicting fiber loss in the field using factory measurements. *Proc. OFC 91*, p. 138.
20. K. Abe, Y. Lacroix, L. Bonnell, and Z. Jakubezyk (1991). Modal noise interference in a short fiber section: Fiber length, splice loss, cutoff, and wavelength dependence. *Proc. OFC 91*, p. 139.
21. C. DeCusatis and M. Benedict (1992, May). Method for fabrication of high bandwidth optical fiber. *IBM Tech. Disc. Bulletin.*
22. D. Marcuse and H. M. Presby (1975). Mode coupling in an optical fiber with core distortion. *Bell System Technical Journal* **1**, 3.
23. J. J. Refi (1986). LED bandwidth of multimode fiber as a function of source bandwidth and LED spectral characteristics. *IEEE Journal of Lightwave Technology* **LT-14**, 265–272.
24. G. P. Agrawal (1988). Dispersion penalty for 1.3 micron lightwave systems with multimode semiconductor lasers. *IEEE Journal of Lightwave Technology* **6**, 620–625.
25. J. E. Midwinter (1977). A study of intersymbol interference and transmission media instability for an optical fiber system. *Opt. Quant. Elec.* **9**, 299–304.
26. K. Ogawa (1982). Analysis of mode partition noise in laser transmission systems. *IEEE Journal of Quant. Elec.* **QE-18**, 849–855; K. Ogawa (1985). Semiconductor laser noise: Mode partition noise. In *Semiconductors and Semimetals* (Ch. 8). R. K. Willardson and A. C. Beer, eds. Vol. 22C, *Lightwave Comm. Tech. Series.* W. T. Tsang, ed. New York: Academic Press.
27. K. Ogawa and R. S. Vodhanel (1982). Measurements of mode partition noise of laser diodes. *IEEE Journ. Quant. Elec.* **QE-18**, 1090–1093.
28. J. C. Campbell (1988). Calculation of the dispersion penalty for the route design of single-mode systems. *IEEE Journal of Lightwave Technology* **6**, 564–573.
29. W. Jiang, R. Feng, and P. Ye (1990). Improved analysis of dispersion penalty in optical fiber systems. *Opt. and Quant. Elec.* **22**, 23–31.
30. M. Ohtsu (1988). Mode stability analysis of nearly single-mode semiconductor lasers. *IEEE Journ. Quant. Elec.* **24**, 716–723.
31. M. Ohtsu and Y. Teramachi (1989). Analysis of mode partition and mode hopping in semiconductor lasers. *IEEE Journ. Quant. Elec.* **25**, 31–38.
32. R. Olshansky (1976). Mode coupling effects in graded-index optical fibers. *App. Opt.* **14**, 935–945.
33. D. Duff (1989). Measurements and simulations of multipath interference for 1.7 Gbit/s lightwave systems utilizing single and multifrequency lasers. *Proc. Optical Fiber Conference* p. 128.
34. J. Radcliff (1989). Fiber-optic link performance in the presence of internal noise sources. *IBM Tech. Report.* Endicott, NY; Glendale Laboratories.
35. L. L. Xiao, C. B. Su, and R. B. Lauer (1992). Increase in laser RIN due to asymmetric nonlinear gain, fiber dispersion, and modulation. *IEEE Photon. Tech. Lett.* **4**, 774–777.
36. B. W. Hakki, F. Bosch, and S. Lumish (1989). Dispersion and noise of 1.3 micron lasers in microwave digital systems. *IEEE Journal Lightwave Technology* **7**, 804–812.
37. AT&T Technical Report (1985). *Methods of Optical Data Link Module Testing* (p. 14).
38. A. A. Manalino (1987). Time-domain eyewidth measurement of an optical data link operating at 200 Mbit/s. *IEEE Trans. on Inst. and Meas.* IM-36, 551–553.
39. P. Trischitta and P. Sannuti (1988). The accumulation of pattern-dependent jitter for a chain of fiber-optic regenerators. *IEEE Trans. Comm.* **36**, 761–765.
40. S. D. Walker, R. B. P. Carpenter, and P. Cochrane (1983). Jitter accumulation with SAW timing extraction at 325 Mbaud. *Elec. Lett.* **19**, 193.
41. CCITT Recommendations G.824, G.823, O.171, and G.703 on timing jitter in digital systems, 1984.
42. F. M. Gardner (1979). Phaselock Techniques, 2nd ed. New York: John Wiley & Sons.

43. G. P. Agrawal and T. M. Shen (1986). Power penalty due to decision-time jitter in optical communication systems. *Elec. Lett.* **22**, 450–451.
44. K. Schumacher and J. J. O'Reily (1986). Power penalty due to jitter on optical communication systems. *Elec. Lett.* **23**, 718–719.
45. T. M. Shen (1986). Power penalty due to decision-time jitter in receivers using avalanche photodioides. *Elec. Lett.* **22**, 1043–1045.
46. K. Schumacher and J. J. O'Reily (1989). Distribution-free bound on the performance of optical communication systems in the presence of jitter. *IEEE Proc.* **136**, 129–136.
47. R. J. S. Bates (1983). A model for jitter accumulation in digital networks. *Proceedings of IEEE Globecon '83* (pp. 145–149), San Diego, CA.
48. P. R. Trischitta and E. L. Varma (1989). *Jitter in Digital Transmission Systems.* Boston: Artech House.
49. C. J. Byrne, B. J. Karafin, and D. B. Robinson, Jr. (1963). Systematic jitter in a chain of digital regenerators. *Bell System Technical Journal* **43**, 2679–2714.
50. R. J. S. Bates and L. A. Sauer (1985). Jitter accumulation in token-passing ring LANs. *IBM Journ. Res. Dev.* **29**, 580–587.
51. C. Chamzas (1985). Accumulation of jitter: A stochastic model. *AT&T Tech. Journ.*, vol. 64.
52. R. Stolen (1979). Nonlinear properties of optical fiber. In *Optical Fiber Communications* (Ch. 5). S. E. Miller and A. G. Chynoweth, eds. New York: Academic Press.
53. D. Cotter (1982). Observation of stimulated Brillouin scattering in low loss silica fibers at 1.3 microns. *Elec. Lett.* **18**, 105–106.
54. H. J. A. da Silva and J. J. O'Reily (1989). System performance implications of laser chirp for long haul, high bit rate direct detection optical fiber systems. *Proc. IEEE Globecom*, pp. 19.5.1–19.5.5.
55. T. Kurashima, T. Horiguchi, and M. Tateda (1990). Thermal effects of Brillouin gain spectra in single-mode fibers. *IEEE Photon. Tech. Lett.* **2**, 718–720.
56. R. H. Stolen and C. Lin (1978). Self-phase modulation in silica optical fiber. *Phys. Rev.* **17**, 1448.
57. R. A. Fisher and W. K. Bischel (1975). Numerical studies of the interplay between self-phase modulation and dispersion for intense plane-wave laser pulses. *Journal of Applied Physics* **46**, 4921.
58. American National Standards Institute. Fiber Distributed Data Interface Physical Layer Medium Dependent (PMD). ANSI Standard X3T9.5.
59. N. R. Aulet, D. W. Boerstler, G. De Mario, F. D. Ferraiolo, C. E. Hayward, C. D. Heath, A. L. Huffman, W. R. Kelley, G. W. Peterson, and D. J. Stigliani, Jr. (1992). IBM Enterprise Systems multimode fiber optic technology. IBM Journ. Res. & Dev. **36**, 553–577.
60. *ESCON I/O Interface Physical Layer Document* (IBM Document No. SA23-0394). IBM Corp., Mechanicsburg, PA.
61. I. D. Allen and E. Frymayer (1993, May). FCS can break the datacom gridlock. *Elec. Design*, p. 87.
62. *Coupling Facility Channel I/O Interface Physical Layer Document* (IBM Document No. SA23-0395). IBM Corp., Mechanicsburg, PA.
63. D. H. Hartman (1986). Digital high-speed interconnects: A study of the optical alternative. *Opt. Eng.* **25**, 1086.
64. M. R. Feldman, S. C. Esener, C. C. Guest, and S. H. Lee (1988). Comparison between optical and electrical interconnects based on power and speed considerations. *App. Opt.* **27**, 1742–1751.
65. R. K. Kostuk, L. Wang, and Y. T. Huang (1988). Optical clock distribution with holographic optical elements. *Proc. SPIE* **977** (real-time signal processing XI), 24–36.
66. J. M. Treubella, J. D. Gelorme, B. Fan, A. Speth, D. Flagello, and M. M. Oprysko (1991). Photopatternable epoxy ridge waveguides. *Proc. Optical Fiber Conference 91*, p. 24.

67. S. Imamura, R. Yoshimura, and T. Izawa (1991). Organic channel waveguides with low loss at 1.3 microns. *Proc. Optical Fiber Conference 91*, p. 23.
68. P. J. Delfyett, D. H. Hartman, and S. Z. Ahmad (1991). Optical clock distribution using a mode-locked semiconductor laser diode system. *IEEE Journal of Lightwave Technology* **9**, 1646.
69. B. D. Clymer, and J. W. Goodman (1986). Optical clock distribution to silicon chips. *Opt. Eng.* **25**, 1103–1108.
70. *IEEE Journal of Lightwave Technology*. (1991). Special issue on optical interconnects for information processing. Vol. 9.
71. *IEEE Journal of Lightwave Technology* (1988). Selected areas in communication: Special issue on broadband packet communications. Vol. 6, No. 9.
72. *IEEE Journal of Lightwave Technology* (1987). Selected areas in communication: Special issue on switching systems for broadband networks. Vol. 5, No. 8.
73. J. D. Crow (1991, March). Optical interconnects speed interprocessor nets. *IEEE Comm. Mag.*, pp. 20–25.
74. J. A. Fried (1986). Optical I/O for high speed CMOS systems. *Opt. Eng.* **25**, 1132–1141.
75. M. Ecker and L. Jacobowitz (1992). *An Apparatus and Method for an Optical Fiber Interface.* U.S. Patent No. 5,155,786 (Oct. 13, 1992), IBM Corporation.
76. J. Crow (1977). GaAs laser array source package. *Opt. Lett.* **1**, 40–42.
77. C. DeCusatis, M. Ecker, and L. Jacobowitz (1994). *Substrate-Embedded Pluggable Receptacles for Connecting Clustered Optical Cables to a Module.* U.S. Patent No. 5,333,225 (July 26, 1994), IBM Corporation.
78. D. A. Irwin (1989). An introduction to SONET. *IBM Technical Report 29.0978.*
79. American National Standards Institute (1988). Digital hierarchy optical rates and formats specification (ANSI T1.105-1988).
80. W. Goralski (1993, November). IBM's spin on ATM. *Network World*, p. 55.
81. CCITT Recommendation G.707, "Synchronous Digital Hierarchy Bit Rates"; CCITT Recommendation G.708, "Network Node Interfaces for the Synchronous Digital Hierarchy"; CCITT Recommendation G.709, "Synchronous Multiplexing Structure." 1989.
82. K. Hagimoto (1992). 20-Gbit/s transmission using a simple high sensitivity optical receiver. *Proc. OFC 92*, p. TuI3.
83. P. A. Anderson (1992). 32-Gbit/s optical soliton data transmission over 90 km. *IEEE Photon. Tech. Tech. Lett.* **1**, 76–79.
84. M. Murakami (1992). 10-Gbit/s 6000 km transmission experiment using erbium doped in-line fiber amplifiers. *Elec. Lett.* **28**, 2254–2255.
85. P. Speier (1991). P. Speier *et al.* (1991). 10-Gbit/s MQW DFB SIBH lasers entirely grown by LPMOVPE. *Elec. Lett.* **22**, 863–864.
86. A. H. Gnauck (1989). 1-Tbit/s km transmission experiment at 16-Gbit/s using conventional fiber. *Elec. Lett.* **22**, 1695–1696.
87. O. Gautheron (1993). 252 km repeaterless 10-Gbit/s transmission demonstration. *Proc. OFC/IOOC 93*, p. PD25.
88. A. Takada (1986). Transform-limited 5.6 ps optical pulse generation at 12 GHz repetition rate from gain-switched distributed feedback laser diode by employing pulse compression technique. *Elec. Lett.* **22**, 1347–1348.
89. K. Iwatsuki (1992). 40-Gbit/s optical soliton transmission over 65 km. *Elec. Lett.* **28**, 1821–1822.
90. M. C. Wu (1991). Transform-limited 1.4 ps optical pulses from a monolithic colliding pulse mode locked quantum well laser. *Applied Physics Letters* **57**, 759–761.
91. S. Shimada (1993). Very-high-speed optical signal processing. *Proc. IEEE* **81**, 1633–1647.
92. A. Felder (in press). 25 to 40-Gbit/s Si ICs in selective epitaxial bipolar technology for future optical fiber communications.
93. H. Ichino (1991). 28-Gbit/s selector using AlGaAs/GaAs HBTs. *Elec. Lett.* **27**, 636–637.

94. A. Gruhle and H. Kibbel (1993). 91 GHz SiGe HBTs grown by MBE. *Elec. Lett.* **29**, 415–417.
95. M. A. R. Violas (1990). 10 GHz bandwidth low noise optical receiver using discrete commercial devices. *Elec. Lett.* **26**, 35–36.
96. P. J. Smith, D. W. Faulkner, and G. R. Hill (1993). Evolution scenarios for optical telecommunication networks using multiwavelength transmission. *Proc. IEEE* **81**, 1580–1588.
97. R. J. Mears (1987). Low noise erbium-doped fiber amplifier operating at 1.54 nm. *Elec. Lett.* **23**, 1026–1028.
98. M. Nakazawa, K. Suzuki, and E. Yamada (1992). 20-Gbit/s 1020 km penalty-free soliton data transmission using erbium-doped fiber amplifiers. *Elec. Lett.* **28**, 1046–1047.
99. S. F. Charter (1991). Amplification at 1.3 μm in an ER + 3-doped single-mode flurozirconate fiber. *Elec. Lett.* **27**, 628–629.
100. J. E. Peddersen (1990). Amplification in the 1300-nm telecommunication window in Nd-doped fluoride fiber. *Elec. Lett.* **26**, 329–330.
101. D. Cotter (1987). Fiber nonlinearities in optical communications. *Opt. Quant. Elec.* **19**, 1–17.
102. J. C. Simon (1990). Semiconductor optical amplifiers. *Proc. OFC 90*, p. ThJ1.
103. H. Taga (1993). 10-Gbit/s, 9000 km IM-DD transmission experiments using 274 erbium--doped fiber amplifier repeaters. *Proc. OFC/IOOC 93*, p. PD1.
104. C. Baack and G. Walft (1993). Photonics in future telecommunications. *Proc. IEEE* **18**, 1624–1633.
105. B. Wedding (1992). Breakthrough in multigigabit system technology due to optical amplifiers. *Tech. Digest, Optical Amplifiers and Their Applications* (Paper FC4, pp. 196–199.
106. R. Heidemann, B. Wedding, and G. Veith (1993). 10-Gbit/s transmission and beyond. *Proc. IEEE* **81**, 1558–1568.
107. B. Wedding (1992). New method for optical transmission beyond the dispersion limit. *Elect. Lett.* **28**, 1298–1300.
108. M. Nakazura, K. Suzuki, and E. Yamada (1990). Distortion-free, single-pass soliton communication over 250 km using multiple erbium-doped optical amplifiers. *Proc. OFC 90*, p. PD5.
109. N. A. Olsson (1990). 4-Gbit/s soliton data transmission over 136 km using erbium-doped fiber amplifiers. *IEEE Photon. Tech. Lett.* **2**, 358–359.
110. D. Grischkowski and A. C. Balant (1992). Optical pulse compression based on enhanced frequency chirping. *Applied Physics Letters* **41**, 1.
111. A. M. Johnson, R. H. Stolen, and W. M. Simpson (1984). 80 X single-stage compression of frequency doubled Nd-yttrium aluminium garnet laser pulses. *Applied Physics Letters* **44**, 729–731.
112. A. Hasegawa, and Y. Kodama (1981). Signal transmission by optical solitons in single-mode fiber. *Proc. IEEE* **69**, 1145–1150.
113. S. Konstantinidou (in press). Architectural requirements of parallel scientific applications with explicit communications. *Comp. Science.*

This Page Intentionally Left Blank

Chapter 7

Bringing a Product to Market

Thomas B. Kellerman
IBM Corporation
Optical Link Products
Endicott, New York

7.1 Introduction

In the preceding chapters we discussed many promising theories in and compelling possibilities about how to use optoelectronics. The ultimate goal of a business is to select and then transform some of these ideas into products whose sales result in bottom-line profits for the investors and many delighted customers. Successfully bringing these ideas to the market requires careful consideration of a number of factors.

Until recently, engineers and researchers created most optoelectronics products in isolation in a laboratory deep in the bowels of usually one company with little thought of the customer. A company's engineers skillfully fashioned the next generation of optoelectronics ideas into products and then turned them over to the sales department to sell. In some cases, there was a defined need for the product; but in

Optoelectronics for Data Communication

a large number of situations, the salespeople had to create the need usually by proclaiming how scientifically advanced the product was. The emphasis was clearly on pushing the next technological breakthrough into the marketplace as rapidly as possible.

A classic example of this was the early development efforts in high speed gigabit laser modules. Initially, the development of these high speed modules took place in the laboratory. The early optoelectronics companies discovered ways to develop transceivers that could operate at speeds in excess of 1 billion bps. These products were indeed impressive and would make any engineer proud. Companies assumed, without question, that customers would line up to purchase the world's fastest gigabit link. However, the early products were very expensive, customers did not have an immediate need for them, and there were very few applications at the time that could readily use either the increased bandwidth or increased distance offered by these early modules. Customers could not afford the high price tag for just technology breakthroughs.

Recently, the price of these modules has come down. They are being tailored more to specific customer applications. Some of the components that are used to make up these early modules are now being made more economically. They are now being repackaged at customer request into lower cost packages. The lower cost packages may not meet every customer's long distance requirements; however, this is far from a problem with some rather significant applications. Moreover, with the lower cost packages and lower prices, more applications are now possible that once were too cost prohibitive. One such application is remote storage devices connected via fiber-optics links. This type of product came as a result of market analysis and working with the customer. Whereas the original modules were an example of creating a product without a customer, the later ones are examples of creating products with customers.

Companies are now starting to understand the importance of being market driven from the very beginning. This is not just a slogan but an attitude in which companies give very careful examination to what is needed by the customer as told by the customer, how the product can be made most efficiently even if it involves working with partners, and how the product or service can be sold profitably by the producing company. Companies now tend to work together, especially with products that require large capital up-front investments. Many product developments are made jointly by groups of companies, by comapnies and universities, and even by groups of companies along with customers who commit up front to use the product. Sometimes the government participates in initial funding as well.

This chapter presents a methodology for converting ideas into products and for marketing and selling the product once developed and manufactured. This chapter explores the roles and interaction of the various players: the customers, partners, and departments or people within a company. The

various decisions and other considerations are presented. The discussion includes the preparation of the all-important business case with its financial considerations. The results of a well-developed business case may reveal some unpleasant realities about what seem to be wonderful ideas. The chapter also discusses the identification of key "tracking" factors during the business case process. If these factors are religiously tracked, they can keep a product development focused and thus make a product a success or cause management to terminate development. Finally, the chapter discusses ways to measure the success of a product in the marketplace.

7.2 Organization

There is no one right way to organize people in a company to select from the many ideas, refine the selected ideas, then plan, design, develop, manufacture, and sell the resultant optoelectronics products. Sometimes the company is organized with separate departments, each representing a specific skill. Each of these departments may be involved in several different products using very different technologies.

With the flattening out of organizations and the elimination of middle management, several tasks that formally had entire departments assigned to them with separate management, now are combined in one department where a team works on only a few closely related products.

In the past, a person or department carried out many of the required steps sequentially. Now, with increasing pressure for shortening the time to

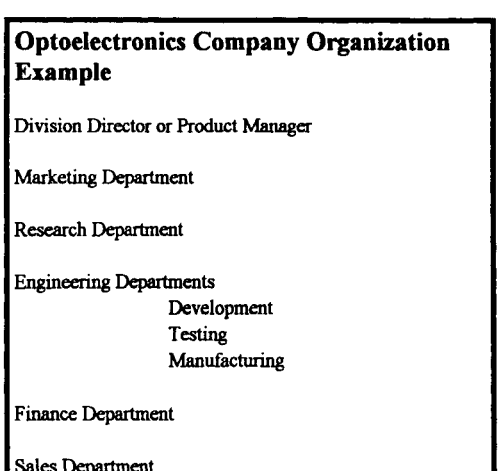


FIGURE 7.1 An organizational structure: Departments of an optoelectronics company.

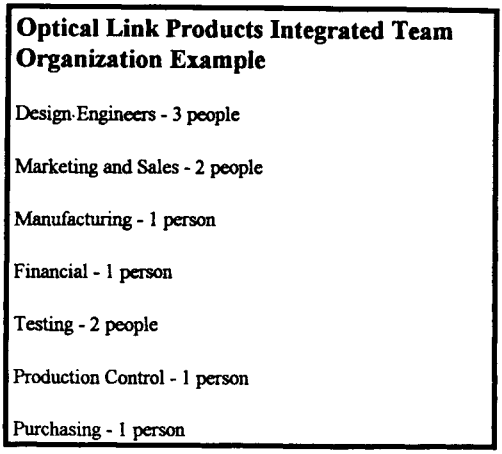


FIGURE 7.2 An organizational structure: Integrated team organization.

market, people or departments carry out much of the work in parallel. There is also an increasing trend toward prudent subcontracting of some tasks. Whether it is a set of departments or individuals in one department, each has different responsibilities for a product somewhere along the product's life cycle. Each plays a key role in this cycle. It is important that all of the departments or teams within a company work together and be measured to the same goal. In some companies, there is a tendency to downplay the role of departments or people that do not have profit responsibility or that do not contribute directly to the company's bottom line. This is dangerous because it can lead to inefficiencies and harmful interdepartmental rivalries.

7.2.1 Role of the CEO or Divisional Director

A company's CEO or divisional director sets forth the company's or division's charter or mission. As part of a strategic planning process, he or she decides in what areas the company should participate. A company may choose to specialize in only a few markets or in a specific niche market. For example, although there are many potential uses for laser transceiver modules, a company may decide that it only wants to specialize in the high speed gigabit products. Another company might specialize in laser weapons or medical devices and yet another company may decide that they are only going to develop links that are used in data communication and stay out of the telecommunication market and the other possible markets for this product. A company may decide to specialize because they want to be known as the best-of-breed provider of this particular area of the market and gain a reputation as such. Also, most companies have limited resources, including

money and people, and feel that they cannot compete well in every possible market.

A company director or CEO may also make the decision to find a partner to share the development of a product. This partner might be a university, another company, the government, or even a customer. The partnership can be very advantageous if the partner can bring expertise or funding to the total product. The partner can also help ensure a market for the product. Another potential benefit of partnership is in the case in which the partner is best-of-breed in a key piece of technology needed for the product and the company is either weak or does not want to invest resources in that specific area.

A company's CEO or division director articulates overall goals and targets for the company to achieve. These may include such things as going after major improvement in time to market, going after profit, going after market share, or going after the first gigabit link to reach commercial status. He or she may also specify financial objectives such as profit margin, revenue, or market share.

The company's CEO or director is the final decision maker in deciding where to ration the firm's capital and how to allocate the company's resources. The other groups within the company prepare the data required for him or her to make this decision.

7.2.2 Marketing

The marketing department's primary focus is the marketplace: the present, future, and even past to the extent that marketing can learn from past experiences of what works and what does not. The marketing group performs an analysis of the potential markets for a company's products.

The marketing group meets extensively with potential customers to determine their needs and wants accurately. Also, most importantly, the marketing group identifies who their customers really are, as it is not always obvious. Marketing gathers and organizes feedback about their products and customers. They conduct market surveys to determine the size of their potential markets. Marketing can develop new surveys or make use of existing surveys. Marketing proposes possible products to new and existing customers. The proposals may be for new potential products or for variations to existing products. Marketing asks the customers whether they would buy the proposed products and/or what changes would make the products more appealing. They can also conduct target marketing surveys. For example, marketing could try to determine the total number of low-end users and what percentages of this number could use a particular product slated for the low-end market. In this way, marketing can determine the potential market, the size of the potential market, and even identify customers in the market for the product.

Not only must marketing gather specific customer requirements as part of the market analysis, importantly, marketing must also learn about the competition. This can help marketing decide how to price and what features to highlight in the product. Understanding the competition is greatly aided by the company's performing a thorough competitive analysis of competing products. Competing products may not only include similar products, but there may also be competing technology solutions for the same customers. For example, in the cost sensitive low-end market, fiber often has to compete with existing copper technology for the same customer. The competitive analysis often involves the help of the engineering and research departments to perform analysis of the design and actual testing of the competing products. Engineering may also perform reverse engineering and construction analysis of a competitor's product to see how it performs.

There are several sources available to help collect the necessary data and to help perform the actual market analysis using the data. An obvious source is the good customers the company already has. Another source for market analysis data is to select customers that appear in optoelectronics trade magazines and to call these customers. Another possibility is to gather the company's main customers together in a group meeting to brainstorm. During these meetings, the customers and the company have an opportunity to discuss their goals and strategies. From these discussions, it is possible to discover ways that the company's technologies could be synergistic with where the customer wants to go. The idea is that both the customer and the solution provider come out of these meetings with a knowledge of each other's strategic directions. Without meeting with the customer, the company risks presenting a solution to a customer that is not in the customer's plans. Another way to put a company in touch with customers is to attend, and in some cases participate in industry trade shows and conferences.

7.2.3 Research

The research group involves itself in the process of discovering new ideas that relate in most, but not in all cases to a firm's charter. The research group's office is the laboratory. Research equips the laboratory with the latest tools for developing the latest in new technologies. Marketing gives the research group information on what new products and improvements to existing products it should focus its attention. The research group also works closely with the engineering group on how to incorporate the research group's ideas into final products.

7.2.4 Engineering

The engineering arm of a company can be divided into different disciplines. The major disciplines are development, manufacturing, and testing.

7.2.4.1 Development

The development group is concerned primarily with the design and function of the product. The marketing group provides information as to what features and functions the customer would like to see in the final product. Development can access ideas from the research groups that might be used toward this end. Development may be given cost and performance targets and be asked to access trade-offs between these targets. The engineering group determines whether the proposed product from marketing is doable and may suggest alternatives. In a lot of cases, marketing may not even have the product completely defined or know what is even possible. This is one important reason for involving the engineering groups early in the process.

Development puts together a comprehensive plan for producing the product. This plan includes a product description that meets the customer's requirements, the costs required to produce the proposed product, the process design for development, and the product schedule. Development's sizing probably includes various scenarios, each scenario trading off one thing for another. For example, the performance achieved by a proposed product may be relaxed slightly if it results in a significantly reduced cost for producing the product. Marketing and engineering work very closely in the initial stages of a product design. They work together and with potential customers to come up with the optimal solution, one that gives the customer the best solution and that maximizes profit for the company.

7.2.4.2 Test Engineering

The test engineer tests the product in accordance with the development engineer's specifications. The test engineer may suggest improvements to the product to aid testing. Test engineering also performs the important function of qualifying the product. Once the product is qualified by test engineering, it is ready for manufacture in quantity. Qualification is discussed in a later section of this chapter.

7.2.4.3 Manufacturing Engineering

The manufacturing engineer is ultimately responsible for transforming development's designs and prototypes to a production level deliverable. He or she makes such decisions as to whether or not the product should be assembled in-house or at an outside vendor's location. This decision whether to produce a product internally or buy it externally from an outside source is known as the make/buy decision. The decision involves not only quantitative cost factors, but also qualitative factors such as ensuring product quality, maintaining long-term vendor relationships, and leaving facilities idle. These are some of the decisions that manufacturing must make to ensure that the product is available for customer demand.

Manufacturing may also take on the added responsibility of ensuring product quality. It is important for a company's image and integrity to ensure

that all products shipped to customers are of high quality, and it is vital that there is a structure in place to ensure that this is always the case.

7.2.5 Finance

The finance group is mainly concerned with the financial aspects of a project. Finance determines how well a company is making use of its assets and resources and forecasts how a company's future endeavors will influence these measurements. Finance is responsible for keeping careful track of the flow of money within a company. The financial group is responsible for managing cash flow, budgets, and tracking the financial measurements of a company. Additionally, the financial goup determines whether the company is meeting its financial targets.

For the most part, finance is not directly involved with the technical issues of a proposed offering, instead it deals with all of the financial considerations. Finance takes the cost and price data and the offering proposal from engineering and marketing, respectively, and determines whether the business case for the proposed product or service makes sense. The company may have financial targets that if not met may mean that there is no point in even continuing. Finance usually has a limited budget with which to work; it must be selective in choosing where to recommend that management invest the company's resources. Fiber-optics products may tend to require a large up-front investment of capital, something finance is always a little leery of.

7.2.6 Sales

The sales group is responsible for selling the product. Some companies combine the sales group with the marketing team. In a company with only a few products, or specializing in similar products, this can be a good idea. In the case of a large multidivisional company that has many diverse and different product groups, this is not always practical. It is often a better idea to have one sales team that can sell the entire array of a company's products. In this case, the sales group needs to have knowledge about the entire portfolio of products that the company has for sale. It is up to the company's engineers and marketing team to educate the sales team on their particular offerings. In reality, when a customer is interested in a particular product from a particular marketing group within a company, the sales group often teams up with the marketing group for that particular customer. Fiber-optics products can have a long sell cycle. This can take much sales and market planning and also patience.

7.3 Ideas for New Products

Research and engineering groups can come up with breakthroughs in laboratories and drawing boards. Marketing, from its offices, can lust over

market surveys of huge opportunities in specific areas of fiber-optics development. Marketing can relate its interpretation of these opportunities to engineering and research and can strongly encourage these groups to focus in these opportunity areas. The customer is torn between wanting to lower costs and wanting to respond to his or her readings in the trade press on such spectacular possibilities as fiber-optics applications like video on demand. Potential business partners may want to share up-front costs of development of their own laboratory ideas. Management and missionaries can be equally well committed to their own special interests. There is no shortage of good ideas. Optoelectronics companies derive ideas from many sources, as shown in Figure 7.3.

7.3.1 Making a Good Idea an Excellent Idea

A careful and detailed analysis is required before a company makes the decision to invest a company's resources, or to divert its resources from a tried and true product line, into the development of a new product. Even the most attractive ideas may not survive answers to questions that may arise in such an analysis.

In fact, once an original idea is selected from competing sources, these sources then team together very closely to ensure that the resultant idea gets better defined so that it can survive a rigorous market analysis. Ultimately, the idea becomes the best that these people working together can deliver (Fig. 7.4).

After the idea is identified, whether it is a new discovery or a possibility in a research department or whether it is an idea from management or a

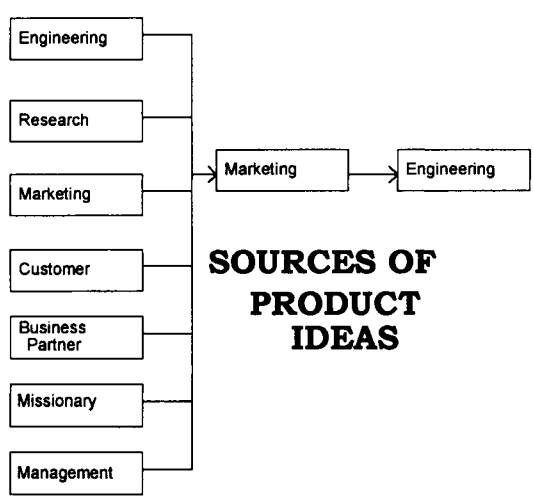


FIGURE 7.3 Sources of product ideas.

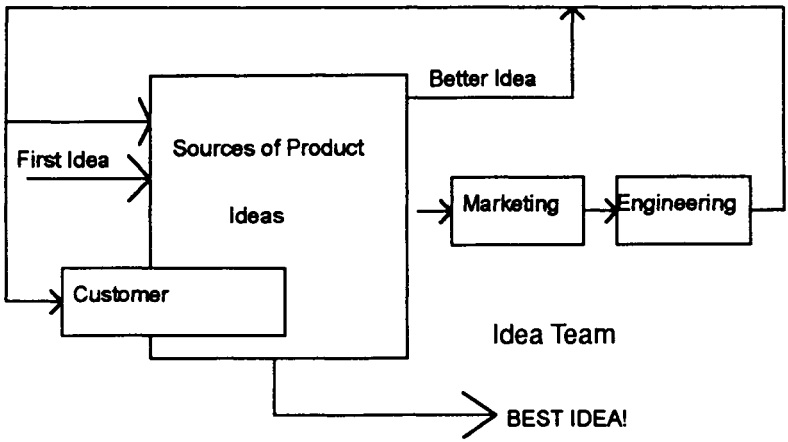


FIGURE 7.4 Idea team.

customer or partner, there needs to be extensive collaboration with the marketing group to go systematically through a responsible analysis. Marketing plays a key role in structuring the analysis, using input from all interested parties. Marketing also plays a critical role in preparing the results so that the appropriate people can decide whether the company wants to invest the money for work now, one year from now, or never.

The marketing group's function is to evaluate the market and determine what ideas to select and how these ideas need to be embellished to satisfy customer needs.

7.3.1.1 Working with Customers

Marketing is the prime contact with potential customers. A key purpose of that contact is to determine whether the idea can be incorporated into a product that can solve a customer's problem or need. Would the customer ever buy the product at any price? The time to have that dialogue with the paying customer is in the beginning.

A lot of times a company is tempted to develop a product based on the perceived and not on the actual needs of the customer. Other times, failure can be attributed to developing a product that is the average of perceived customers' needs so what results is a product that solves not any one customer's needs. The unhappy result of this is that not one customer buys it.

Although the marketing team is the main customer interface, it is not unusual for the engineers to accompany the marketing team on customer visits. This helps the engineering team better understand his or her customers. It also helps the marketing team to bring with them the design engineers of a potential product. There can be many surprises on these visits, most are pleasant, but some may be unpleasant.

Pleasant surprises usually involve the engineer realizing that the unreasonable requirements that they understood in the laboratory were really not the right ones and that the real ones could be articulated in such a way that makes implementing them both reasonable and possible. Sometimes the design can be substantially completed by engineers and marketing people working together in the customer's office. There is nothing like this to convince a customer that there is seriousness in the proposed venture. Also, when the engineer returns and is working on new products, actually talking to the customers and getting the feedback directly can make requirements understandable. It can significantly help engineers prioritize their efforts. For example, customer discussions can result in realistic specifications.

Often in the past, a product was overdesigned and thus unduly expensive. Testing became extremely difficult and expensive. As an example, suppose that you go into the customer's office because of a phone call with the data center manager. Judging from the initial conversation, it is clear that the customer does not know exactly what he wants. From the customer's perspective, he says that he needs to go a distance of 4 km and send data at a rate of 25 Mbps. What he does not tell you is that he is sending important financial transactions and needs a very low bit error rate. You arrive all ready to propose the new and exciting low-cost, lower spec link because you think that he only needs to go 4 km and the data rate is easily achievable by this product. Besides, most customers do not need the added cost or specs of the full-featured product. But it turns out, unfortunately at the customer's office, that the low-cost product has a bit error rate that is too high, so is short of the customer's requirements. Luckily, your development engineer, the person that designs and specs these links, is with you. Instead of going home with no sale, you, the customer, and the development engineer are able to "engineer" a new proposal right in the customer's office. The engineer is highly knowledgeable about the design and implementaton of fiber optics and has several different solutions that can be proposed to the customer. The engineer also suggests another idea: instead of paying for the expensive data link, the customer might be able to get away with the low-cost version, but perform error checking elsewhere to ensure that the financial data is intact and correct.

Not all ideas that please customers and bring profit to the bottom line require huge initial investments or major technology breakthroughs. Sometimes working with the customer results in product spin-offs, small modifications of existing products that yield very satisfactory results. One such example is simply coloring a fiber-optics connector to match a customer's logo colors. Another area of profit and customer satisfaction is defining offerings to be services rather than products. Such a service offering could be the testing of a fiber-optic link under certain conditions. Some of the test equipment that optoelectronics requires is quite expensive for any one company to invest in; a company may be only too willing to pay for a one-time testing service.

7.3.1.2 Working with Strategies

Although it is quite persuasive to develop ideas that are market driven and not primarily technology driven this decade, there are times when exceptions are made. A company may be embarking on a well thought out strategy to grow a market that does not currently exist.

In fiber optics, this could be in connection with the data super highway. In this case, there is no immediate customer, so marketing's job is to come up with a viable plan to enable the market. This is risky. A current trend is to form an alliance of partners so that the risk can be shared. It should be pointed out that before committing a company's resources, it should be pointed out that before committing a company's resources, it should be determined that there really is going to be a market for this product. Even though the proposed product does not have an immediate market, it is still beneficial to find business partners and customers to work with. By doing so, the company may gain useful insights as to how this new product could fit into a customer's future requirements and what market this product should be targeted to.

The first personal computers (PCs) are a good example of this. Although customers did not specifically ask for desktop computers, many customers now make use of them for a wide array of applications. PCs have improved customers' productivity immensely. Computer makers had to first make computers easy to use and show customers why PCs would help them. The computer manufacturers showed customers how they could perform their jobs better by using PCs. In other words, they enabled the market.

An example where a strategy backfired was in the development of low-cost plastic fiber optics for data communications. Years ago the differences in the costs of traditional glass fiber versus that of plastic were significant. Plastic cable assemblies were cheaper to buy and manufacture. It appeared that several applications that did not require the increased bandwidth of glass could substitute low performance, low-cost plastic. The disadvantages of decreased performance, in some cases, could be traded off with the lower cost which could enable fiber optic technology to enter the low end cost sensitive markets and compete head to head with copper.

This all sounds very good. However, recently, the costs of glass fiber cables have been coming down dramatically. This has virtually eliminated the market for plastic fiber for most applications for which in the past it would have been an ideal candidate. Part of the reason for the decrease in cost was that the number of producers of glass fibers increased and suppliers were forced to lower their margins and learn how to make cheaper glass fiber cables in order to compete.

Certain laboratory discoveries, although they may be excellent ideas, are either not in a company's best interests immediately or strategically, or do not make good business cases for a company. Sometimes the idea, itself, may be sold to another company or it may simply sit someplace and never be implemented into a product. An example of this would be an optoelectronics

company that did not want to develop laser products, such as guidance systems for weapons, even if the margins are good for this product and the company has technology that could be used in this area. Instead, the company may choose to focus strictly in the data communications arena.

7.3.2 Ideas for New Businesses

Optoelectronics is a growing market. More applications will use fiber optics because of innovative ideas, decreasing prices, increasing bandwidth requirements, and improved manufacturability. In addition to an optoelectronics product, there are a variety of ways to make money with optoelectronics.

Fiber Optics Businesses and Services
Design services
Component manufacturing (lasers, chips, fiber)
Assembly
Custom design work
Consultation
Testing
Installation
Sales and distribution
Basic research
Education and training

Testing services could include verifying a customer's link in the laboratory by simulating his or her data center. For example, a customer's long link with 10 breaks could be simulated and tested in the laboratory by using large spools of fiber and attenuators and then analyzing the throughput and determining the bit error rate.

Consulting services could include advising the customer how he or she should set up his or her machines with fiber optics to meet specific performance requirements. This service could include setting up the actual systems as well as selling the hardware.

The original equipment manufacturer (OEM) market is another area in which to sell the company's components. For example, a chip manufacturer may not only supply chips for their own transceiver modules, but also for another supplier's modules.

7.4 Preparing the Offerings Proposal

There are many other considerations to think of besides "the best idea."

- How much is the product going to cost to manufacture?
- Does the company have any of the manufacturing or testing facilities or skills to manufacture or test the product?
- How much can the customer afford to pay for the product?
- Does the company know how and where to sell this product?

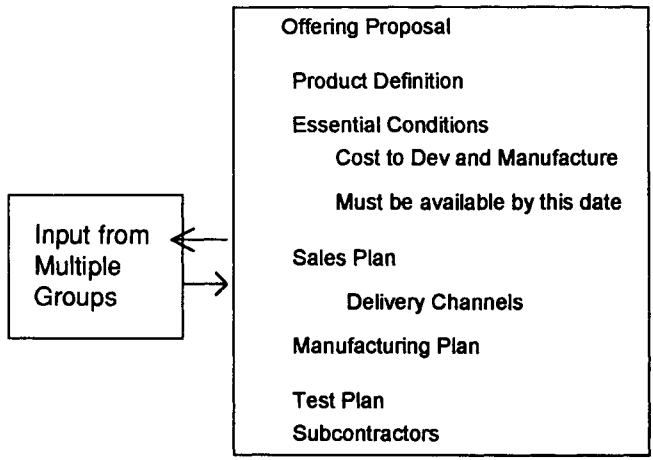


FIGURE 7.5 Offering proposal.

- Can a profit be made by any one company or group of partners developing, manufacturing, or selling this product?

To go from an excellent idea to a product or a service available in the fiber optics area, an offering proposal is prepared. It will contain proposed answers to some of the preceding questions. This offering proposal will then be the critical input to a business case.

7.5 Business Case

Once an offering proposal for a product has been defined, the next step is to determine whether or not to proceed with the development of the product. Although a company may have many good ideas, few may actually make it into products. The business case tells the company what risks are involved in developing the product, the costs, the market for the product, the expected revenue from the product, and the financial return that the company can expect from the product. The business case presents a financial model of the program and the expected payoff. It should tell the company if this is really the best use for the company's resources. At any given time there may be many different products competing for a firm's limited resources.

In addition to financial considerations, there are sometimes nonfinancial points to consider. Sometimes the ideas for products are good ones and have a very favorable business case but never see the light of day. This is because the products themselves are not within the goals of the company. That is, the company does not want to move in this direction, but would rather specialize in something else.

The business case, in order to be complete and successful, requires input and involvement from the different groups talked about in the previous section. They work to put together the best business plan so that the financial analysis comes out to be as optimal as possible. A favorable business case based on limited inputs may yield an unsatisfactory result. The key elements from each group that go into a business case are shown in Figure 7.6.

7.5.1 Inputs

The best way to show how a business plan evolves into a completed business case is to look at an example of one. We can run through the steps of a business case by using a hypothetical product. The product is a smart cable. A smart cable is a fiber-optic cable that has optics built into both connector ends. The cable is able to convert a computer's electrical signals into optical signals that can travel along the glass fiber. The ends plug into the standard electrical ports of a computer, specifically into a computer's serial port. This allows computers not already equipped with fiber optics to make use of fiber optics without having to purchase an adapter card.

The smart cable

Features

- Allows computers to take advantage of fiber optics
- Connects already existing computers by means of their serial port
- Is lightweight
- Offers a path for signals over significantly increased distance

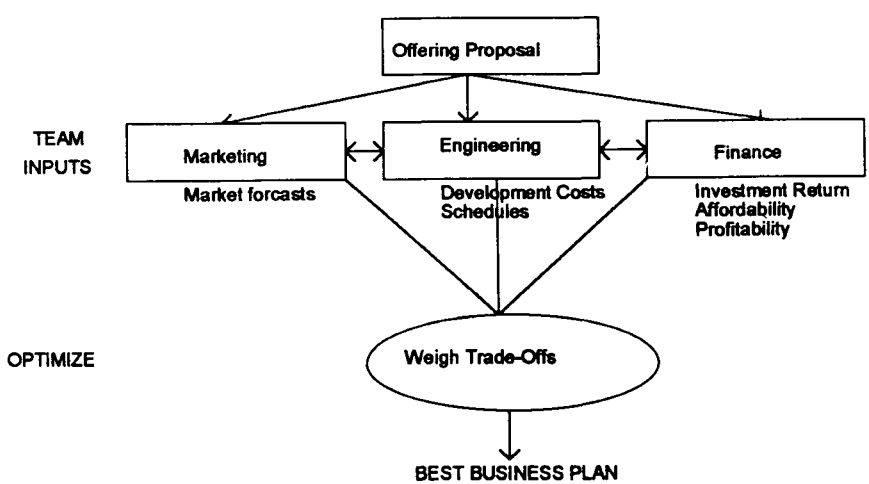


FIGURE 7.6 The business case.

Schedule

This product takes 1 year to develop, qualify, and manufacture.

Cost

Costs are broken down into nonrecurring fixed costs and variable costs. The fixed cost is the initial investment needed to start the program. In the case of the smart cable, there are fixed costs of $80,000 required. Engineering has determined that the product requires an up-front investment of $50,000. A mold and a manufacturing line are necessary to produce this product. Development has determined that approximately $30,000 is required for tooling to make the connector mold. Manufacturing has determined that this product is best made in-house initially, as careful examination of each manufacturing step is required. There is an up-front charge of about $20,000 to set up the manufacturing line to assemble this product. Test engineering has created a test plan, based on development's specifications that assures that the product meets certain minimum reliability requirements. The cost to qualify the product is estimated to be about $20,000. Included in this fee is the purchase of test equipment not currently in the existing test laboratory. Also, test engineering indicates that approximately 3 months is needed to test the product. This resource requirement is factored into the $20,000 cost. Development, test engineering, and marketing have all requested prototypes. Development estimates that the design and fabrication of prototype parts requires approximately $10,000. The preceding constitutes the fixed costs involved with producing the smart cable product. This is the initial investment that the company must make if it decides to go forward with this project. This input is very important to the financial group in the company, because the company needs to spend this money before getting back anything in terms of revenue.

In addition to the fixed costs, there are, of course, variable costs associated with this product. These costs include the costs of the actual materials, the cost to assemble the product, and any other direct costs attributed to the product. In most cases, as is the case with the smart cable, the larger the volume produced, the cheaper the cost to produce it. This is beause some of the product materials can be bought cheaper in bulk and the manufacturing line used to make this product can be loaded to capacity to save on overhead. This lowering of costs based on increasing volume is known as economies of scale.

The variable costs of the smart cable are determined by manufacturing engineering. They have developed a sourcing plan for this product and have determined the costs to produce the product in a production environment in volume. All of the costs are summarized in Table 7.1. The accompanying figure (Fig. 7.7) shows a graph of the total cost as a function of the number of units produced. As can be seen, the fixed cost is constant and needs to be overcome regardless of the number of units sold. The variable component of cost declines per unit as volume increases.

TABLE 7.1 Costs of producing smart cable.

Fixed Costs	
Tooling	$50,000
Qualification	$20,000
Prototypes	$10,000
Total fixed cost	$80,000

Variable Costs	
Number of Units	Cost per Unit
1	$80
5,000	$60
10,000	$50
15,000	$45

Total Cost

Number of Units	Fixed Cost	Variable Cost	Total Cost
1	$80,000	$80	$80,080
5,000	$80,000	$300,000	$380,000
10,000	$80,000	$500,000	$580,000
15,000	$80,000	$675,000	$755,000

The marketing and sales team have put together a sales forecast for this product. The forecast shows the worst case, the best case, and the most likely case for how the product will sell in the market. This forecast is based on the market analysis that has been performed. It is important that the marketing and sales team be realistic in their forecast, as this will determine whether the company makes money or not.

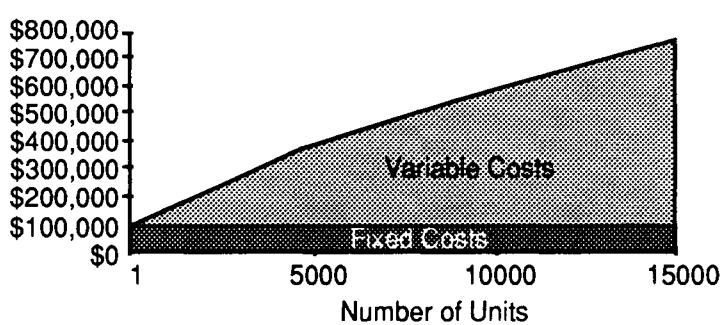


FIGURE 7.7 Total cost of smart cable.

Sales forecasts

Case	Low	Likely	High
Volume (units sold)	6,000	7,500	10,000

An important function of the marketing team is to price the product. The marketing team does this as part of the marketing analysis. It uses such things as customer surveys and competitive analysis to do this exercise. One way of pricing this product is to use the profit maximization approach. The number of smart cables that the company sells is a function of its price. We can start with a base price, estimate how many we can sell at this price, then vary the price and determine how much the volume and profit change. For example, a 10% increase in price may result in only a 4% decrease in the number sold. On the other hand, a 10% price decrease may result in 30% more volume and more profit. The goal is to find a point at which profits are maximized. At one end of the spectrum, the company can sell the smart cable for a very low price and although a lot of people may buy it, the margins will be extremely small. On the other hand, the company can charge such a high price that even though the margins are attractive, there are no customers willing to buy it. This is a common mistake when a company is trying to recover its investment too quickly and charges a price higher than the prevailing market is willing to pay. Somewhere along the pricing curve is the point of diminishing returns. This is the point at which further increases in price cause a reduction in volumes of a large enough magnitude to reduce the firm's profits. It is not always easy to determine what this price should be. Also, in some cases, to increase volume, a company may want to offer volume discounts to its big customers. In the case of the smart cable, the optimal average selling price is determined to be $65.

7.5.2 Breakeven Analysis

The financial group takes all the financial inputs from the team and puts together a financial analysis of the proposed offering. This involves costs from engineering and proposed pricing and sales forecasts from marketing. They can determine the gross profit that the company should make based on different volume scenarios. Table 7.2 shows a summary of this along with Figure 7.8.

TABLE 7.2 Breakeven analysis.

Volume	Revenue	Cost	Gross Profit
1	$65	$80,080	($80,015)
5,000	$325,000	$380,000	($55,000)
10,000	$650,000	$580,000	$70,000
15,000	$975,000	$755,000	$220,000

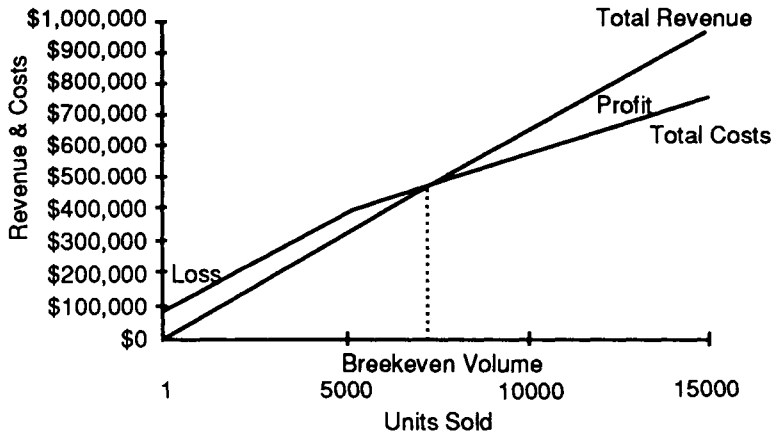


FIGURE 7.8 Breakeven analysis.

A breakeven analysis can be performed that shows how many of the smart cables the company must sell just to breakeven. Of course, the intent is to do more than breakeven. If any of the assumptions or inputs from the team are not accurate or change in the course of developing the product, the story can change. Therefore, it is a good idea to perform a sensitivity analysis. The sensitivity analysis shows the effect on the business case measurements if some of the variables are varied slightly. For example, what would happen if marketing cannot sell the smart cable for $65 and has to lower the price to $60?

Once the business case work is completed, the team can present the offering proposal with the business case to management for its decision.

The team presents the proposal, the offering plan, and business case to management (Fig. 7.9). Management may decide to accept or reject the proposal. However, it may also decide to request more information. All parties desire a quick resolution one way or another even if the answer is no.

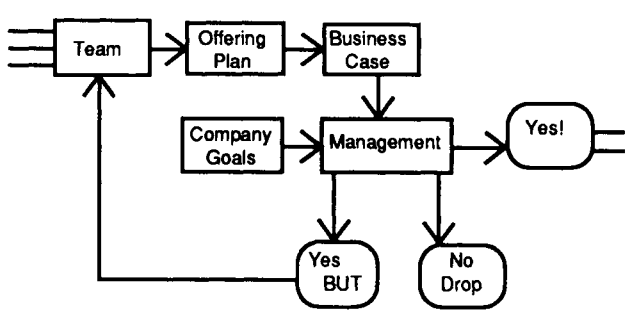


FIGURE 7.9 Decision to proceed.

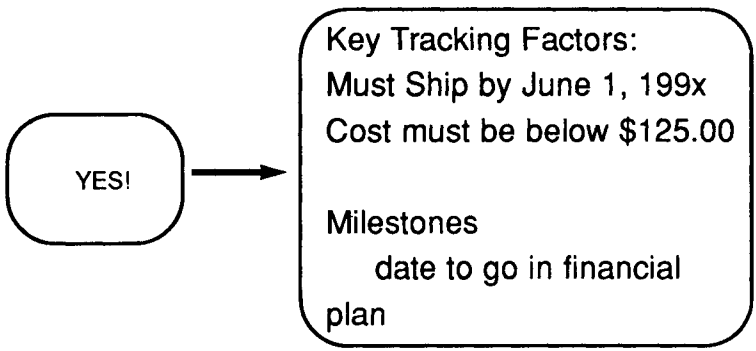


FIGURE 7.10 YES!

Valuable resources can be tied up in endless cycles through this loop; however, if the work is done well, approval can be quick and a yes can be achieved (Fig. 7.10).

The offering proposal now becomes the product or service plan.

7.6 YES!

7.6.1 Program Management

Once a product has been defined by the team and the measurements and milestones determined, it is important that the offering is tracked (Fig. 7.11). It is sometimes tempting for each group to go off and work on their own piece

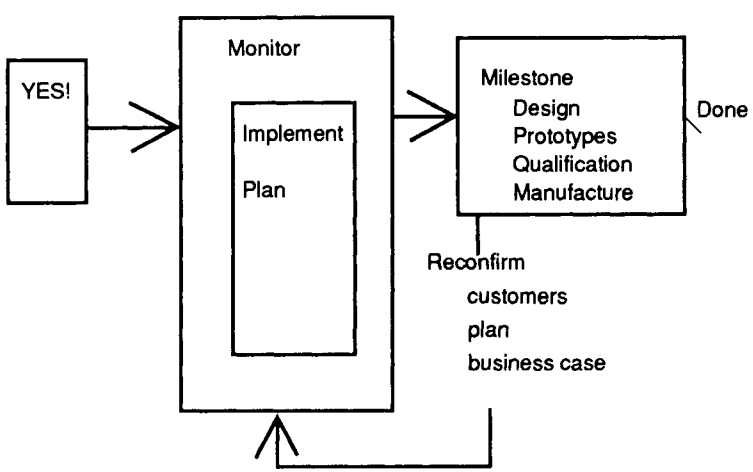


FIGURE 7.11 Program management.

without talking to the other team members. This is dangerous and can lead to problems integrating with the other team members.

Although no one in the organization likes to think about things going wrong during development, manufacture, and test, unexpected obstacles can crop up. The product can have unexpected schedule delays ranging from design difficulties to environmental acts of nature. A product can fail in testing. For example, a cable product may have unacceptable optical losses resulting from a bad process in the manufacturing line. It is then necessary to go back and try to debug the process until it is done right. There may be cost overruns. A chip set may cost considerably more than expected because of a lower initial yield than expected. Eventually, however, the design is completed, prototypes are built, and the product is ready to be qualified.

7.6.2 Qualification

In the qualification step, the product is tested in accordance to the environment that it is intended to be used in. This can include environmental stress testing, mechanical tests, functional (optical BER) testing, and other kinds of tests. Test engineering qualifies the product in accordance with the specification supplied by the product developers (development). The specification, of course, is based on the customer's requirements and is then translated into what specific tests need to be performed to meet them. The specification may have trade-offs where the team has decided to relax certain specifications, for example, to achieve cost benefits. Also, during testing, test engineering may come back to the team and propose changes to the product or explain where they feel the product is weak, or even suggest relaxing the specifications some more if the product does not meet them. Test engineering may feel that improving the product to a more stringent specification is not worth the cost and not necessary to meet the bulk of the market demand. Once the product is blessed by qualification, it is finally ready to be produced in a manufacturing environment. A line is set up to make the product in quantity. Test engineering finally approves the integrity and reliability of the product as it comes off the production line.

7.6.3 Manufacturing

After the product is qualified, it goes into manufacturing. Recall that manufacturing engineering has been involved since early in the process to ensure that the product has been designed for manufacturability. Manufacturing engineers, as part of the offering proposal and accepted business plan have concluded what is to be manufactured inside the company and what is to be manufactured outside the company. In general, new techniques and technologies are manufactured inside the company, especially in the early stages of production.

7.7 Selling the Product or Service

Once manufacturing has produced the product, the next step is to sell it actively. Recall that the overall offering plan includes a sales plan. As part of this plan, the sales team is continually working to seek out and identify any and all potential customers for the product. The team has put in place a plan that makes it easy for customers to find out about the product and to actually buy the product. There are many possible ways to promote the product. Most are in the sales plan; others are developed as the product is ready to be sold.

7.7.1 Identifying the Customer

The customers for optoelectronics can be roughly divided into four categories.

1. End users such as large computer installations. They may desperately need a fiber management system to contain their miles of cables. They may need to increase dramatically their physical bandwidth by installing a fiber-optic backbone.
2. An OEM customer. They may need gigabit chips to complete modules that they plan to repackage to sell to many other people.
3. Distributors. They may purchase inventory to sell.
4. Most importantly, the customers who participated in the beginning by helping identify and refine the requirements and design, and their partners and customers.

The sales team wants to make it extremely easy for all of these different kinds of customers to find out about the products and services.

7.7.2 Helping the Customer Find Out About the Product

Here are some things that the sales team does toward this end. The sales team

- places advertisements in trade journals
- participates in trade shows
- writes articles in trade magazines
- produces product catalogs
- places "cold calls"
- contacts existing customers

One good and obvious way is for the sales team to call existing customers to see whether they can use the product. There are also many trade shows that are attended by potential customers who are looking for optoelectronics technology. One very big show is the Optical Fiber Convention which is held every year. This shows includes both suppliers and end users. Another benefit

of attending these shows is the opportunity to size up the competition. Companies may not only just want to attend shows specifically geared to optoelectronics. For example, attending a super computing show might be a good way to identify customers that need to incorporate fiber optics in these high performance products.

Advertising is sometimes a good strategy especially when the company is new in optoelectronics and needs to get its name, products, and services known. The company should determine what markets it wants to be in and target its advertising toward decision makers who buy in these areas. For example, if the product is a cable management system, the company may want to put an ad in a cable installers' magazine describing how easily a data center manager can do his or her job with this product. If the company is selling chips, for starters, the company might put its ad in an electronics developers' magazine. Also, the company clearly wants an easy way for customers to contact it after reading the ad. An 800 telephone number and even an e-mail address or a card that potential customers can mail in are techniques that facilitate quick contact.

The sales team creates sales literature that can be sent out to any interested customer. This literature describes the company's products, lets the customer know how to get more information, and, most importantly, tells the customer how to place an order. In certain situations, a company may decide to start a direct mail campaign in which marketing packets are sent out to targeted customers with the hope of getting some "hits." The hit rate is also a good way to measure a product's acceptance in the marketplace.

Big customers identified as potential users may be cold called by the sales group. This may be necessary to educate the customer as to how the product can help them. Before the cold call, the sales team does extensive research on the customer and prepares a scenario for how their optoelectronics product can help that specific customer.

7.7.3 Measuring in the Marketplace

The product is out in the market place and the sales force is selling as hard as it can. It is important to measure exactly what is happening against the forecast. Is it as expected? How is the market reacting to the product? Does the right segment of the market know about the product? If the product is not living up to its expectations, it is important to identify this early and either change the sales strategy for the product, modify the product, or decide to focus resources on something more promising.

Marketing and sales keep track of all of their sales calls and all of the potential customers to whom they have attempted to sell the product. For a customer that did not buy the product, the marketers should ask themselves why and what could have been done to enable the sale. Perhaps, the customer was not really a possibility in the first place. How could a better identification

of customers be accomplished that would result in having a higher success rate? The sales team keeps records of how many sales are actually made and to whom. The actual sales figures should be compared with the forecast to see how well it tracks. Also, projections as to future sales should be determined to see how well the product is going to do in the future. Good news is, of course, if the product exceeds all volume expectations. However, this can take a turn for the worst if the company cannot keep up with demand. If the situation is not rectified soon, customers may become angry. The opposite situation exists when despite all that the sales team does sales fall woefully short of the forecast. Then it is time to reassess the business case and, in some cases, although a company's pride may be hurt, it is time to kill the project.

7.8 Conclusions

The goal of the optoelectronics company is to understand and recognize a customer's need, define products that realize the need consistent with the company's overall mission, determine how products can be made efficiently and effectively, and plan how the products can be sold profitably. Once the answer is yes to proceed to development, several departments or teams of people with different skills are brought in to deliver the product.

We have presented a methodology that can take good ideas, make them excellent ideas, and incorporate these ideas into complete offering proposals. Financial people can develop business cases. Management can decide on the basis of the goals of the company, the offering proposal, and the business case whether to proceed to development, manufacture, and sales. Critical to this methodology is working with the customer from the very beginning.

Optoelectronics is a new and exciting field and breakthroughs are continually happening. Companies must know how to select from a wide range of possibilities and then how to best utilize their resources to deliver the resultant products. Many departments and people are involved in a process to accomplish this.

For an optoelectronics company, investing significant resources in technology for technology's sake can be a big mistake. Not knowing the potential customer is also a big mistake. Commitment to the marketing concept needs to involve all of the people in a company. The marketing group may devote considerable resources into working with the customer and understanding the wants and needs of the customer. This effort is wasted if there is not total agreement from all departments. A company's engineering groups may create products that interest them, without guidance from the marketing group. Developing the next great breakthrough in laser link technology can be much more fun than figuring out how to put a purple logo on an existing connector or helping prepare a customer to use a testing service. However, the company's profits lie in investing resources to do the latter.

References

1. R. Emery, and J. Finnery (1991). *Principles of Finance with Corporate Applications.* New York: West Publishing Company.
2. D. Newman (1983). *Engineering Economic Analysis.* 2nd ed. San Jose, CA: Engineering Press.
3. Optical Fiber Convention (OFC).
4. R. Peterson (1989). *Principles of Marketing.* New York: Harcourt Brace Jovanovich.
5. J. Siegel, and J. Shim (1986). *Managerial Finance.* New York: McGraw Hill.
6. J. Siegel, and J. Shim (1991). *Financial Management.* New York: Barrons.
7. T. Vollmann, W. Berry, and D. Whybark (1988). *Manufacturing Planning and Control Systems.* 2nd ed. Homewood, IL: Irwin.

This Page Intentionally Left Blank

Chapter 8

The Future of Information Technology

George Cybenko
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755

8.1 Introduction

In an obscure 1940 experiment conducted at an American Mathematical Society meeting, George Stibitz demonstrated an early prototype of the information superhighway. From Hanover, New Hampshire, Stibitz controlled a digital relay computer at Bell Laboratories in New York by transmitting data over ordinary telephone lines. At that time, Stibitz was working as an engineer at Bell Laboratories designing switching circuits. Recognizing that when such switches are combined they could be used to perform computations, he built simple calculators that were used internally by the phone company to assist in filter design problems. The phone company had little interest in the fledgling computer business back then, but Stibitz did show that the telephone system could be used to control computers remotely. This demonstration, innocent and unheralded at the time, was a premonition of the

Optoelectronics for Data Communication

excitement and turmoil currently being created by the confluence of computing and communications technologies.

Today, more than 40 years later, it is a matter of fact not faith that computer, telecommunications, and even media technologies are merging. Developments regarding the national information infrastructure (NII), the information superhighway, and digital convergence are reported on a daily basis in the business, entertainment, and technology news. Companies in what were previously considered to be diverse industries are merging or forming strategic alliances. We are told that industry, government, education, and entertainment are about to undergo radical transformations because of these new technologies. The magnitude of these changes has been compared with the revolution brought about by the invention of the movable-type printing press by Guttenberg 500 years ago.

The information superhighway has its true believers and its sceptics. The ultimate reality will likely lie somewhere in between. I believe that exciting possibilities are awaiting us, but in order to realize their full potentials, we need to deal with some fundamentally new engineering and intellectual challenges. In the remainder of this chapter, I will sketch the major issues as I see them and then venture some predictions about how things will turn out in the short run. For the impatient reader, let me outline my main points now.

First of all, the information superhighway is not a single monolithic entity like the interstate highway system to which it is often compared. It is a metaphor for a wide assortment of technologies and services that will be brought to market shortly by telecommunications and computer companies. We constantly hear announcements of these new products, new services, and new technologies. How are we to make sense of them? Here is a useful litmus test: one should ask, does the research, new service, or new product promise to deliver more existing raw information or does it organize the information curently or soon to be available? I believe that the benefit of the former is becoming marginal unless radically new services are developed in the coming years. The critical problem is organizing, associating, and otherwise making sense of information. We have crossed an information watershed, one side of which is the demand for more information, and the other side is the demand for better information organization tools.

Second, the existence of technology for "making sense of information" assumes some concrete, quantifiable model of information that we presently do not have. We do not have reliable automatic methods for extracting meaning from documents nor do we even have a very useful vocabulary or taxonomy for the various types of information that do exist. Construction of the information superhighway is an engineering problem, but we do not yet have engineering models of the problems it is meant to solve.

Much of the information content proposed for the superhighway is already available in some other form so that the major new benefits are in packaging

and convenience of access, not novelty. The main exception so far is interactive content, in which a user can dynamically create new content. This brings us to my third point. Computers must play a more prominent role in this information revolution and must not be relegated to being just high tech "VCR" players of content. The communications component of the information superhighway does not provide enough novelty or capability alone to sustain an expensive national investment.

To pursue the point further, note that a major use of computers now is to simulate. Aerodynamicists use computers to simulate the properties of airfoils without actually building the airfoils and testing them in wind tunnels. Similarly, financial analysts can computationally simulate possible futures for financial markets. The ability to simulate, using information made available by new communications capabilities, must be made accessible and useful to the average citizen and professional for the superhighway to be successful.

These points can be summarized another way as follows. Much of the content that will move around on the information superhighway will be old hat initially. Our ability to access it quickly and conveniently will be a novelty. However, our ability to use it in radically new ways to improve our personal and professional lives remains relatively unexplored—herein are the major engineering and intellectual challenges. These challenges will be met successfully only if the simulation power of computing can be brought to full use in the information superhighway.

8.2 The Technology

The technology of the information superhighway rests on two major innovations of this century: integrated circuits and fiber optics. These technologies are major discontinuities in the evolutions of computing and communications, respectively.

Early digital computers used vacuum tube components as their switching elements. Older readers will remember televisions and radios with tubes that were typically an inch in diameter and about 3 inches high. Tubes burned out regularly and generated much heat, a sign of their significant power consumption. Transistors were invented in the late 1940s and were replacing vacuum tubes by the late 1950s. A single transistor replaces a single tube and in the 1950s a transistor was more than 100 times smaller and proportionately more energy efficient. Today, device sizes on a state-of-the-art integrated circuit are measured in fractions of a micron, some 100 million-fold reduction in size over vacuum tube technology.

This radical reduction in size and power requirements were completely unforeseen in 1950, the beginning of the digital computer era. Computer pioneers of the time were grossly misjudging the impact and ubiquity of computers. Howard Aitken of Harvard was such a pioneer, but in 1947 he

TABLE 8.1 Predicted trends in semiconductor computer technology. (Source: Semiconductor Industry Assn.)

Year	1993	1996	1999	2002
Line width (Microns)	0.5	0.35	0.25	0.18
DRAM capacity (Mbits)	16	64	256	1024
Speed (MH_3)	150	300	400	500+

stated that "there will never be enough problems, enough work, for more than one or two of these computers." Of course, he was talking about computers that occupied gymnasiums, cost small fortunes to build, and required a full-time staff of a dozen skilled technicians to operate. Integrated circuits and other innovations have made it possible for college freshmen to carry around with them notebook computers whose power and memory dwarf the capabilities of Aitken's machines. The invention of the integrated circuit resulted in radically new ground rules and created a fundamental discontinuity in what was possible. From one side of the discontinuity, it was impossible to predict what would be commonplace on the other side.

Table 8.1 shows predicted trends in semiconductor computer technology through the year 2002. As a reference point, note that today's supercomputers from Cray Research Incorporated have a clock speed of 166 MHz (6 nsec) and the 150 MHz clock has been achieved by stock microprocessors from Digital Equipment Corporation, for example. (The DEC Alpha processor has a top speed of 200 MHz.) In the DRAM arena, the entire contents of the *Encyclopedia Britannica* can be stored on eight 1024-Mbit (1 Gbit) memory chips.

Fiber optics makes for a similar story in communications. Traditional telephone wire service into our homes can move digital data around at the rate of about 64 000 bps, which is equivalent to 8000 typed characters a second. Here, I am speaking of so-called ISDN service, which is available in major metropolitan telephone service areas using existing telephone infrastructure. The copper wires themselves have a capability of moving about 1 million bps (125 000 characters), but the switching and multiplexing equipment limits that to the smaller number previously mentioned. Note that the smaller number of 64 000 bps was essentially determined by the requirements of high quality speech transmission, not data communications. The coaxial cable used for cable television has a capacity more than 250 million bps, sufficient for the 50 or so channels of television offered on most systems at about 5 million bps for each channel.

Optical fibers are theoretically capable of communication well over one million million bps. Written out, that is over 1 000 000 000 000 bps or 200 000 channels of television per fiber. This is theoretical capacity as tested in laboratories, not the capacity that is currently commercially available. Curiously, the orders of magnitude reduction in size when going from

vacuum tubes to today's integrated circuits is close to the orders of magnitude increase in the data transmission rate when going from voice quality telephone to what appears to be the capacity of optical fiber. However, in the integrated circuit arena, we have now about 20 years of commercial development and deployment of the technology which has brought us to the point where virtually every household has products that use relatively advanced integrated circuits. Integrated circuits are ubiquitous.

This is not the case yet with respect to optical fiber. Fiber is still primarily deployed in specialized applications. However, it will become ubiquitous and it will bring about changes. Similar to the ones we have witnessed in the transition from vacuum tubes to integrated circuits. But our perspective on how it will be used is closer to Aitken's 1947 predictions about computers rather than what reality will end up being. That is, most of our projections about applications of the information superhighway are extrapolations of our current use of communications technology.

As an example of misjudging markets for new services and technologies, consider the videophone and cellular telephones. Videophones have been repeatedly marketed with high expectations and have repeatedly been rejected by consumers. On the other hand, the cellular telephone was introduced in the mid-1980s with little fanfare, but the market response continues to astound the telecommunications industry with almost 25 million subscribers in 1994. Here is a situation in which mobility and ubiquity of existing services has been valued far more than expanded capability.

8.3 Technology versus Services

I have been arguing that it is difficult, if not impossible, to envision changes across a technological discontinuity. To pursue the point further, let me use another example involving digital libraries.

In 1963, John Kemeny, a future Dartmouth president and co-inventor with Tom Kurtz of BASIC, articulated his vision of the "Library of the Year 2000." In Kemeny's library, users sat at television-like consoles, interacting with a "National Library" using a telephone-like dial on which the user would dial codes for topics that were catalogued in a directory. The directory would be a paperbound document, always at the user's side, and would be compiled by experts drawn from the various areas of human knowledge. Information in the National Library would be structured hierarchically, and given a request for a topic, the Library's computers would retrieve titles applicable to the request. The user could then view the documents, one page of text at a time, as displayed on the console's screen. Kemeny went on to describe the cost of sustaining such a national facility and how that cost could be broken down in such a way as to make it a reasonable expense if shared by its institutional users.

The point is that Kemeny underestimated the progress of technology but overestimated the rate at which the corresponding services would be developed. This is eerily similar to Aitken's predictions. We must remember that in 1963, human–computer interactions meant punch cards and batch queues. The possibility of using rotary-dials for input, let alone video monitors, was radical then but seems quaint now. Kemeny's estimates for the speed at which data could be moved from the central computer site to users was surpassed by the early 1980s and completely missed the discontinuity created by fiber optics. Kemeny's library had consoles located at strategic locations in contrast to the relative ubiquity of personal computers and workstations today made possible by integrated circuits.

If the technology surpassed all expectations, why did services not do the same? This is a problem of growing proportions in the computing and communications industries. It is evident today in parallel computing in which advanced parallel computer hardware has surpassed the software base, which is now playing catch up. In the communications area, the technology underlying advanced services such as ISDN and videophones has existed for some time but the services themselves have lagged in acceptance.

On the bright side in the applications arena is the meteoric rise of Mosaic and the World Wide Web (WWW). The World Wide Web is a distributed information system on the Internet in which documents are linked to one another using a standardized formatting convention (hypertext markup language or HTML). Documents are very general entities, consisting of text, audio, video, and still images. Mosaic is a "browser" for WWW documents. It allows users to view pages and select "links" to other documents which may be resident on other computer systems at other sites giving the user the illusion that all the documents are local to his or her machine. This Internet application is remarkable because it emerged as a noncommercial enterprise and yet has generated enormous interest and activity. It offers a universal interface to the millions of Internet documents that have been created by thousands of authors asychronously. Commercialization of Mosaic is proceedingly intensely now and its descendents may ultimately realize Kemeny's future library.

It is wrong to reach the blanket conclusion that technology is easier to advance than services. Certainly, a major part of the explanation can be traced to the fact that our educational and research establishments have been devoted to developing technologies as opposed to the engineering aspects of services, applications, or, in the computing sphere, software systems. There may be a concerted change in this trend as the federal government now puts more focus on strategic, as opposed to basic, research and many industrial and national labs are cannibalizing their basic research activities into product and service development efforts.

Surely, some of the major successful applications of computing and communications today, things like electronic mail, facsimile transmission, word

processing, database systems, and spreadsheets were not much discussed 30 years ago. Again, the point is that technology tends to exceed expectations, whereas services and applications appear to lag or head into unforeseen directions, especially when there is a technological discontinuity.

8.4 Implications

One predictable consequence of a discontinuity follows from the economic law of supply and demand. When a discontinuity creates a great supply in some resource, that resource becomes a commodity and its unit value drops. This has happened with computers of all types when their computational power and storage has increased dramatically and industry standards have made components and software operationally interchangable. With little in the way of unique value-added applications to distinguish one vendor from another, computing systems have become essentially commodity items, driving profit margins down dramatically in recent years. The current financial woes of the computer industry can be traced in part to this fact.

What will be commodities when fiber optics becomes fully deployed? Bandwidth for one and possibly information.

One way to estimate the cost of bandwidth is to tie it to video-on-demand. Video-on-demand will simulate VCR capabilities for cable television subscribers: instead of renting and playing tapes, viewers will call up movies over the network which will then deliver them to their homes instantaneously. This is a service being test marketed now and may be extensively deployed nationally within 5 years. Today, the rental cost of a video tape is about $2.50 to be competitive, this service will likely have to be priced in the $2 to $5 range. In the middle range, this is about $0.03 per minute. The bandwidth required for TV quality compressed video is, say, 1.5 million bps, which is equivalent to about 100 simultaneous voice telephone signals or FAX transmissions at an equivalent cost of $0.03 per minute, which is significantly below toll phone charges now.

The point is that fiber optics and the corresponding video transmission services will force voice and FAX service charges to drop off the lower end of the tariff scale: bandwidth for those applications becomes a commodity and profit margins are microscopic if pricing is proportional to bandwidth.

This scenario applies to countries with advanced communications systems already in place, such as in North America, Europe, and Japan. The situation in the developing world still begs for bandwidth to support standard services that are still not extensively deployed. Table 8.2 shows the pending demand for standard telephone service and estimates of the corresponding investment required to support that service.

Information itself has the distinct possibility of becoming a commodity item as well, but this is a more complex issue. First of all, we have very

TABLE 8.2 Demand for telephone service in developing countries. (Source: International Telecommunication Union.)

Country	New lines added by 2000	% increase in lines	Investment ($ billions)
China	35.5	19.3	53.3
Russia	15.5	6.7	23.3
India	9.1	6.7	23.3
Brazil	6.8	6.4	10.2
Mexico	6.3	8.5	9.4
Thailand	4.3	16.7	6.6
Malaysia	3.1	11.9	4.6
Poland	2.7	6.7	4.0
Indonesia	2.6	13.6	3.9

primitive models of information as a market commodity. Information does not satisfy physical conservation laws. If someone sells a million barrels of petroleum, ownership of that petroleum changes hands and it ceases being the seller's asset. But if I transmit to you some information, generally speaking, I still have the original digital copy of that information and can resell it to another customer later. Moreover, unless there is some claim to intellectual property rights, other vendors can sell the same or similar information.

Many forms of information are essentially free already. People can obtain much information at public or institutional libraries without incurring personal cost. Broadcast network news and entertainment is free to the viewer. It is subsidized by advertisers. Therefore I believe that the marginal profits in the information supply business are not going to fare well unless the

TABLE 8.3 Corporate responses (782 respondents) about information technology (IT) priorities. (Source: Computer Sciences Corporation.)

1994 Rank	Issue	1989 Rank
1	Reengineering processes using IT	11
2	Aligning IT and corporate goals	2
3	Organizing and using data	6
4	Building cross-functional systems	7
5	Setting up IT infrastructure	5
6	Improve systems development process	13
7	Update obsolete IT	Unranked
8	Integrate IT systems	12
9	Improve IT personnel	8
10	Change basic technology for IT	Unranked
11	Develop an IT strategy	4
12	Cutting IT costs	14

information content is legally protected in some way. Entertainment information falls into this category but factual information does not. There are too many possible sources for factual information, that is, data, be it medical, educational, financial, or current events. A possible example of this may be the Mead Corporation, parent of Mead Data Central which operates LEXUS and NEXUS information services, which is auctioning off this unit to concentrate on its paper products business. On the doorstep of the coming information superhighway, this is either temporary insanity or insightful business forcasting.

Table 8.3 ranks corporate attitudes toward a variety of information technology (IT) issues. A rough interpretation of this list suggests that applications of information (services) lead, followed by basic technology and then cost cutting in this area.

8.5 Opportunities

I have painted a bleak future for the commodity side of things but a bright one for consumers. Integrated circuits have made computers commodities, fiber optics will make bandwidth a commodity, and the combination may make much information a commodity. Producing a commodity does not mean the end of the world for an industry. Take potable water for example. What it does mean is that production must be very efficient, and we are seeing this in the computing and telecommunications industries now. Moreover, markets must expand to increase production volume and the information superhighway is coming along at just the right time.

So where is the upside? What are the technology opportunities in expanding areas? This is the watershed just described. The bottleneck will cease being access to information, it will become organizing information and making it more immediately useful. There are a number of technological implications of this.

Mobility and Ubiquity. The challenge has much to do with integrating a number of different information processing products into a smaller number and making them portable if possible. Now, we are dealing with telephones, FAX machines, computers, television sets, radios, printed media, and pages for accessing contemporaneous information. Consumers cannot tolerate more devices or access methods—there are already too many. The cellular telephone has demonstrated that mobility and ubiquity are a major concern.

Simplicity. Information technology must be simple to become widely deployed. The best example is the telephone. The dial interface works around the world and is immediately usable. It is simple but functional. The operation of FAX machines is similarly relatively uniform. The ultimate network interconnection scheme that implements future information systems

must be similarly transparent to the user and seamless with respect to applications.

Complexity. We are about to deploy a brave new world of integrated communications systems involving a wide variety of transport technologies and network protocols. Gateways between networks must minimize latencies due to buffering for flow control when there are data rate mismatches between networks. Control, fault tolerance, billing, transport technology, and application requirements will be extremely heterogeneous and decentralized. Tools for managing such complexity must emerge in the coming years if the information networks of the future are to be stable, ubiquitous, reliable, affordable, and capable of supporting diverse services.

New Applications. High speed communications networks, if extended to subscriber loops, will require an enormous investment in infrastructure. That investment must be recouped by offering novel services for which billing generates significant income to the service and infrastructure providers. Video-on-demand was previously mentioned as a significant new income source and home shopping is another that has surprised many. On-line personal services (travel, finance, want ads), information services (newspapers, libraries, ticker tapes, wire services), video conferencing, interactive education, and telemedicine are all candidate application areas.

Bandwidth. The existing information infrastructure represents an enormous investment which remains to be depreciated in the coming years. New signal processing and communications techniques are continually required to squeeze more bandwidth from the network transport layer as it exists at any time.

8.6 Conclusions

New information technologies have remarkable potential but their development and deployment must proceed hand in hand with new applications and services that will pay for the technology and drive the demand for it. Our present vantage point on future services is based on our current and past experiences, and is limited by the boundaries created by those past technologies. There are numerous past examples in which predictions of future directions based on past extrapolations have led to blind alleys and wasted resources. The development of appropriate novel services and applications will require creativity and capital commensurate with the resources devoted to building new physical infrastructure.

Index