WIRELESS COMMUNICATION USING NON-DIRECTED INFRARED RADIATION

by

John Robert Barry

Memorandum No. UCB/ERL M92/133

2 December 1992

WIRELESS COMMUNICATION USING NON-DIRECTED INFRARED RADIATION

by

John Robert Barry

Memorandum No. UCB/ERL M92/133

2 December 1992

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Wireless Communication Using Non-Directed Infrared Radiation

by

John Robert Barry

Abstract

Wireless Communication Using Non-Directed Infrared Radiation

by

John Robert Barry

Doctor of Philosophy in Engineering—Electrical Engineering and Computer Science
University of California at Berkeley
Professor Edward A. Lee, Co-Chair
Professor David G. Messerschmitt, Co-Chair

This dissertation investigates the feasibility of high-speed wireless communication using non-directed infrared radiation, and concludes that data rates near 100 Mb/s are practical. We identify the key impediments to high-speed communication, namely, a weak received optical signal and multipath dispersion, and propose design strategies to counter them. We present a detailed link analysis that accounts for path loss, antenna gain, reflection loss and filter loss. This analysis provides analytical expressions for the signal-to-noise ratio as a function of receiver position. By maximizing the worst-case signal-to-noise ratio within the coverage area of the transmitter, we optimize the transmitter radiation pattern and receiver optics. We present a procedure for designing a wideband preamplifier with sufficiently low noise that the shot noise inherent in background light is the limiting source of noise. We characterize multipath optical propagation for diffuse-reflector environments by presenting a theoretical model and comparing its predictions with experimental measurements. We evaluate the performance of various modulation schemes on the intensity-modulation channel, and show how it differs from a conventional linear Gaussian-noise channel. The emphasis throughout is on the physical-layer problem of establishing a high-speed and robust point-to-point communication link.

Edward A. Lee

David G. Messerschmitt

ACKNOWLEDGMENTS

Funding for this project was provided by a number of organizations. I am particularly grateful to the T. J. Watson Research Center of IBM for their generous financial support during the first two years of this project. Other sources include grant number MIP-86-57523 from the National Science Foundation, Sony Corporation, Dolby Laboratories, and the California State MICRO Program.

I am grateful to my first adviser, Professor Edward Lee, for providing me with extraordinary research facilities, and also for his constant encouragement and inspiration. During my last three years at Berkeley I had the good fortune of two advisers. Special thanks go to my second adviser, Professor David Messerschmitt; first, for instigating this project, and second, for showing me that beyond each accomplishment lies another challenge. This project was strongly influenced by the guidance and wisdom of Professor Joseph Kahn, to whom I am greatly indebted. Professor David Brillinger provided useful comments on an early draft of this manuscript, as did fellow graduate students Malik Audeh and Gene Marsh. I am grateful to Peter Hortensius of IBM for providing early guidance during the summer of 1990, and also for sharing his ideas on multipath dispersion that led to the theory of chapter 4. Bill Krause was instrumental in acquiring the experimental multipath measurements of chapter 4. Conversations with the following colleagues were also helpful: Andy Burstein, Jeff Carruthers, Shih-Fu Chang, Andrea Goldsmith, Paul Haskell, Chris Hull, Phil Lapsley, Chung-Sheng Li, Horng-Dar Lin, Vijay Madisetti, Kris Pister, Annis Porter, Rajiv Ramaswami, Gil Sih, Ravi Subramanian, and Greg Uehara.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The trend in telecommunications is towards wireless access, as is evident by the proliferation of cordless and cellular telephones, wireless PBX, and paging services, as well as the growing interest in personal communication networks [1][2][3]. The trend in computer technology, on the other hand, is towards smaller and faster machines, which has led to the current popularity of portable and pen-based computers. As the telephone network integrates digital services with voice services, the traditional distinction between the telecommunications and computer communications industries is blurring. It is likely, therefore, that these two trends will merge, and that portable computers will someday access high-speed network services through wireless links.

Data communication requires a much higher bit rate than voice communication. There are a few wireless radio products on the market that offer speeds of up to 2 Mb/s to portable computers, which is close to the 10-Mb/s data rate common in today's wired (Ethernet) networks. Future applications for wireless data communications will require much higher data rates. For example, consider the requirements of a portable high-quality digital display. To reduce its size, weight, battery-power consumption, and cost, it may be advantageous to make it as "dumb" as possible, relegating intensive signal-processing tasks such as video decompression to the transmitter platform (assuming it is not portable itself). For similar reasons, future portable computers may be nothing more than portable interfaces, with very little on-board computational power. To accomplish this, however, will require short-range wireless communication links with extremely high capacity.[1] In the extreme case, for example, an uncompressed high-definition television image can require a data rate of 1 Gb/s or more.[2] More realistically, data rates near 100 Mb/s may be adequate for practical applications. Multiplexed traffic for multiple users can drive the bit rate even higher. For a number of technical and regulatory reasons, current wireless technology for portable terminals cannot support such data rates; the fastest radio link available today for portable computers is a 2-Mb/s product by NCR Corp. called Wavelan [4], although speeds have been projected to approach the 10-20 Mb/s range in a few years [5].

Infrared radiation, particularly near-infrared radiation with wavelengths in the 700-1500 nm range, enjoys a number of advantages over radio as a medium for short-range wireless communication. The primary advantage is an abundance of unregulated bandwidth. In addition, infrared systems are immune to radio interference. Infrared radiation, like visible light, will not penetrate walls and other opaque materials, so that an infrared signal is confined to the room in which it originates. This makes infrared a secure medium, preventing casual eavesdropping. More importantly, it allows neighboring rooms to use independent infrared links without interference. This attribute is especially advantageous when high-speed access is required throughout a large building with many stories,

1. Shifting the computational burden from the terminal to the network infrastructure in this way does more than increase the required bit rate, it also constrains the way in which this bit rate can be achieved, preventing the receiver from using extensive coding and other signal-processing techniques that are themselves a computational burden.

2. Assuming 1400×1000 pixels per frame, 24 bits/pixel, 30 frames/sec.

because it bypasses the frequency re-use problem that plagues cellular radio networks. The fact that infrared networks will not be interference limited is a key feature distinguishing them from cellular radio networks.

Infrared has disadvantages as well. As discussed in later chapters, line-of-sight infrared links are susceptible to shadowing caused by objects or people positioned between the transmitter and receiver. Non-directed infrared links also have limited range, in the tens of meters, because too much path loss will attenuate the signal below the strong ambient-light noises that exist in typical office environments.

On balance, the virtually unlimited, unregulated bandwidth of infrared dwarfs these disadvantages, making infrared a promising medium for short-range wireless links. In fact, it may the only option for cost-effective high-speed communication at rates near 100 Mb/s. Infrared networks will not replace radio networks, nor will they replace wired networks. Instead, the role played by infrared in future networks will likely be complementary, replacing the last few meters of cable from the network to the user, allowing mobility and portability without sacrificing service.

1.1 WIRELESS OPTICAL COMMUNICATION

A wireless optical communication link can take a number of forms; we adopt the following notation. A *directed transmitter* has a narrow-beam radiation pattern, and a *directed receiver* has a narrow field of view. Likewise, a *non-directed transmitter* has a broad-beam radiation pattern, nearly omnidirectional, and a *non-directed receiver* has a wide field of view. A *directed link* consists of a directed transmitter and a directed receiver. Similarly, a *non-directed link* consists of a non-directed transmitter and a non-directed receiver. A *hybrid link* consists of either a directed transmitter and non-directed receiver or a non-directed transmitter and a directed receiver. Finally, a *line-of-sight (LOS) link* is one in which there exists an unobstructed line-of-sight path between the transmitter and receiver. The six possible link configurations (directed LOS, directed non-LOS, hybrid LOS, hybrid non-LOS, non-directed LOS, and non-directed non-LOS) are illustrated in Fig. 1-1. We will examine each of the six configurations and point out related works by other researchers, restricting our attention to indoor links conveying digital data.

Consider first the directed LOS configuration shown in Fig. 1-1-a. It makes efficient use of optical power because the signal energy is concentrated into narrow optical beams, and also because most of the ambient background light is rejected by the narrow field of view of the receiver. Furthermore, a directed LOS link does not suffer from multipath dispersion, both because reflectors are not illuminated by the transmitter and because reflectors are not in the receiver field of view. These advantages prompted a number of researchers to adopt the directed LOS configuration in experimental wireless links. For example, Yun and Crawford reported a wireless network in 1985 that used 1-Mb/s directed LOS links to connect multiple terminals to a centrally located base station [6]. The base station transmitted 165 mW into a circularly symmetric planar beam with a 3° vertical beam width, while the terminals transmitted 5 mW into a narrow 2° beam. The range was 50 m. To provide full duplex transmission without interference, different wavelengths were used for the up link (terminal to base station) and down link (base station to terminal).

A similar system was proposed by Chu and Gans [7]. By replacing the omnidirectional planar transmitter radiation pattern of the base station with a number of discrete 1° pencil beams carrying 1 mW of power each, they were able to achieve a data rate of 50 Mb/s at a range of 30 m. A product developed by BICC Communications is currently on the market that uses the directed LOS configuration; it is reported to have a throughput of 4 Mb/s and a range of 24 m [8].

A major drawback of the directed LOS systems is their inability to cope with the one-to-many (broadcast) and many-to-one communication modes. This problem is eliminated by using the directed non-LOS configuration, as shown in Fig. 1-1-b. This configuration was used by Photonics Corp. in 1985 for its Photolink product, which operates at

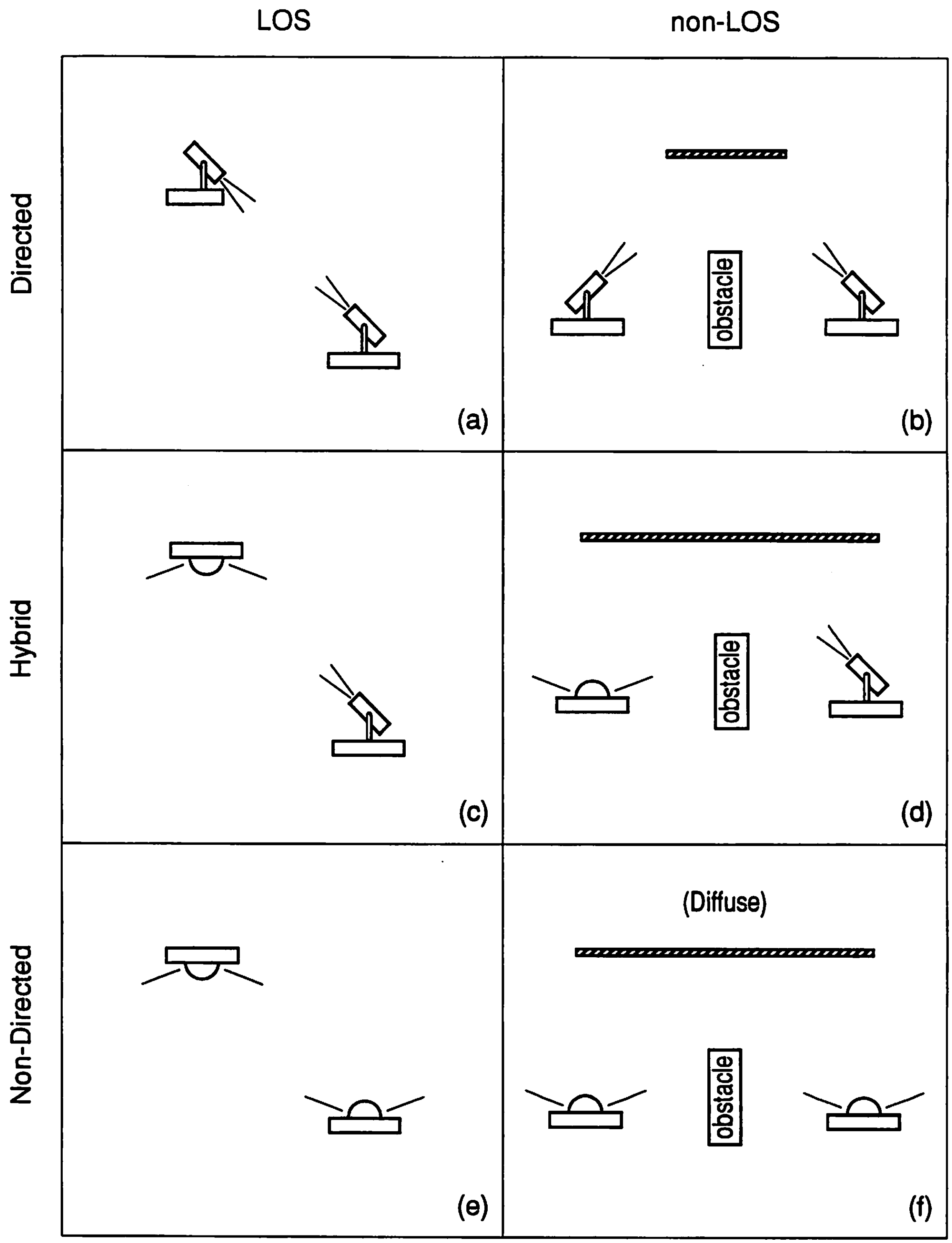


Fig. 1-1. Configurations for wireless optical links.

230 kb/s with a range of about 22 m [9][10]. In this system, each node in the network directs its transmitter and receiver towards a common "target" point on a diffusive-reflecting surface in the room, usually the center of the ceiling. (The width of each beam is estimated to be about 5° [10].) Many of the surfaces found in typical offices, such as painted plaster and ceiling tile, are nearly ideal Lambertian reflectors, which means that the reflected energy per unit solid angle is proportional to $\cos(\theta)$, where θ is the angle from the surface normal, regardless of the angle of incidence. This is because the surface variation on these surfaces is large relative to the wavelength of the infrared signal, and hence they act as nearly perfect diffuse reflectors. This has been verified experimentally [11]. With this configuration, energy from each terminal is re-radiated from the target reflector in every direction, some of which makes its way to the narrow-field-of-view receivers on each terminal. A similar system, proposed recently by Yun and Kavehrad [12], uses multiple directed non-LOS beams and multiple target reflectors to provide diversity against inadvertent shadowing.

By replacing the passive reflector in the Photolink system with an active repeater or base station, we arrive at the hybrid LOS configuration of Fig. 1-1-c. Like the Photolink system, this system can also support the one-to-many and many-to-one modes. There have been a number of systems based on the hybrid LOS configuration [13][14][15][16]. For example, Minami *et al.* [13] report a 19.2 kb/s system in which the base station transmits 135 mW into a 120° wide beam and the receivers transmit 75 mW into a moderately directed 60° wide beam. The range was 10 m with LOS intact and 5 m with LOS obstructed. With the LOS obstructed, the configuration falls under the hybrid non-LOS category, as shown in Fig. 1-1-d. Takashi and Touge extended Minami's work by increasing the transmitter power and changing the modulation scheme from subcarrier frequency-shift keying to subcarrier phase-shift keying [14], achieving 48 kb/s over the same range.

In a similar hybrid LOS system, reported by Nakata *et al.* [15], the base station transmitted 300 mW into a broad beam with angle 120°, while the terminal transmitter used a narrow 10° beam. The range was 5 m. Fuji and Kikkawa [16] reported a similar system, achieving 19.2 kb/s at a range of 10 m using a 150° down-link beam and a 10° up-link beam.

Unfortunately, none of the directed or hybrid configurations shown in Fig. 1-1-a through Fig. 1-1-d are appropriate for portable computers and other mobile platforms, because they all require alignment between transmitter and receiver. Wireless links for portable terminals must be non-directed, as illustrated in the bottom row of Fig. 1-1. The focus of this thesis is on non-directed links, both LOS and non-LOS.

Consider next the non-directed non-LOS configuration of Fig. 1-1-f, which was first proposed by Gfeller *et al.* in 1978 [11][17]. The transmitter emits infrared energy into a broad optical beam, and the receiver has a wide field of view. A link of this sort is often referred to as simply a *diffuse link*, because it relies on diffusive reflections to provide an optical path between a transmitter and receiver for which the LOS may be blocked. Gfeller measured the reflectivities of typical office materials such as painted surfaces, wood, carpets, plaster walls, and found that the fraction of infrared power reflected from their surface falls between 40% and 90%, with 80% being typical for plaster walls [11]. Thus, the optical signal in a diffuse link can undergo many reflections and still have appreciable energy. Furthermore, these surfaces are well-approximated by an ideal Lambertian reflector, so that incident infrared energy will re-radiate in all directions. This provides multiple redundant paths between the transmitter and receiver that makes the diffuse channel difficult to interrupt by shadowing [18]. (Multipath propagation is discussed in detail in chapter 4.)

Gfeller's 1978 paper was the first to propose the use of infrared for a wireless LAN [17]. (It was, in fact, the first wireless LAN proposal using any medium, radio included [5].) The base station illuminated the ceiling with a broad optical beam of

800 mW, so that the ceiling acted as a distributed secondary source [11][19]. The bit rate was 125 kb/s and the range was 10-20 m, depending on the severity of the ambient light noise. More recently, Photonics Corp. has developed a small 1-Mb/s transceiver for integration into portable computers [9][20] that uses the non-directed non-LOS (diffuse) configuration. Spectrix Corp. also uses the diffuse configuration for its 2-Mb/s wireless link [21]. Clearly, the diffuse configuration is the most convenient from the user's standpoint, because the user does not have to worry about alignment or maintaining a LOS path. Unfortunately, however, because of severe signal attenuation and multipath dispersion, the diffuse link is also the most challenging from a design standpoint.

As a compromise, the non-directed LOS configuration of Fig. 1-1-e makes better use of signal power than the diffuse link, but it requires that the LOS path be unobstructed. This configuration may find use in very-high-speed applications and long-distance applications, for which the energy transfer in the diffuse link is inadequate, and for which a LOS path can be maintained without undue inconvenience. For example, a recent multichannel public access telephone system was reported by Poulin *et al.* that is loosely based on the non-directed LOS configuration [22][23]. In this system, the base station consists of an array of moderately narrow directed beams, each pointing in a different direction, so that the combined effect is a wide optical beam. Similarly, the receiver at the base station consists of an array of narrow field-of-view detectors, each looking in a different direction, so that the net effect is a wide field-of-view receiver. The range was about 20 m and the data rate was 230 kb/s. The LOS approach is necessary for environments in which the diffuse approach is not viable; for example, large rooms with high ceilings, or outdoors.

1.2 AN APPLICATION: A HIGH-SPEED WIRELESS LAN

A potential application of non-directed links is a high-speed wireless LAN; see Fig. 1-2 [24]. The network consists of a backbone of base stations, connected to each other and to an information server by a cable, most likely fiber optic. The portable computers communicate with the base stations via a non-directed link. Associated with each base station is a cell or a coverage area, which is the region in the room in which a portable terminal can maintain a dependable link with that base station. In a non-directed LOS configuration, the base stations are fixed on the ceiling and flood the room with infrared radiation, whereas in a diffuse configuration, the base stations are below the ceiling and illuminate it with a broad optical beam. A single base station is sufficient to provide coverage for a small single- or double-occupant office, and office walls prevent interference between neighboring rooms. Due to a tight power budget, as discussed in chapter 2, the cell radius is small, about 5 m. Thus, in large open offices or factory floors, multiple base stations will be required, and some means for avoiding adjacent-cell interference are necessary. For example, each cell could be assigned a different wavelength or subcarrier frequency; see chapter 6.

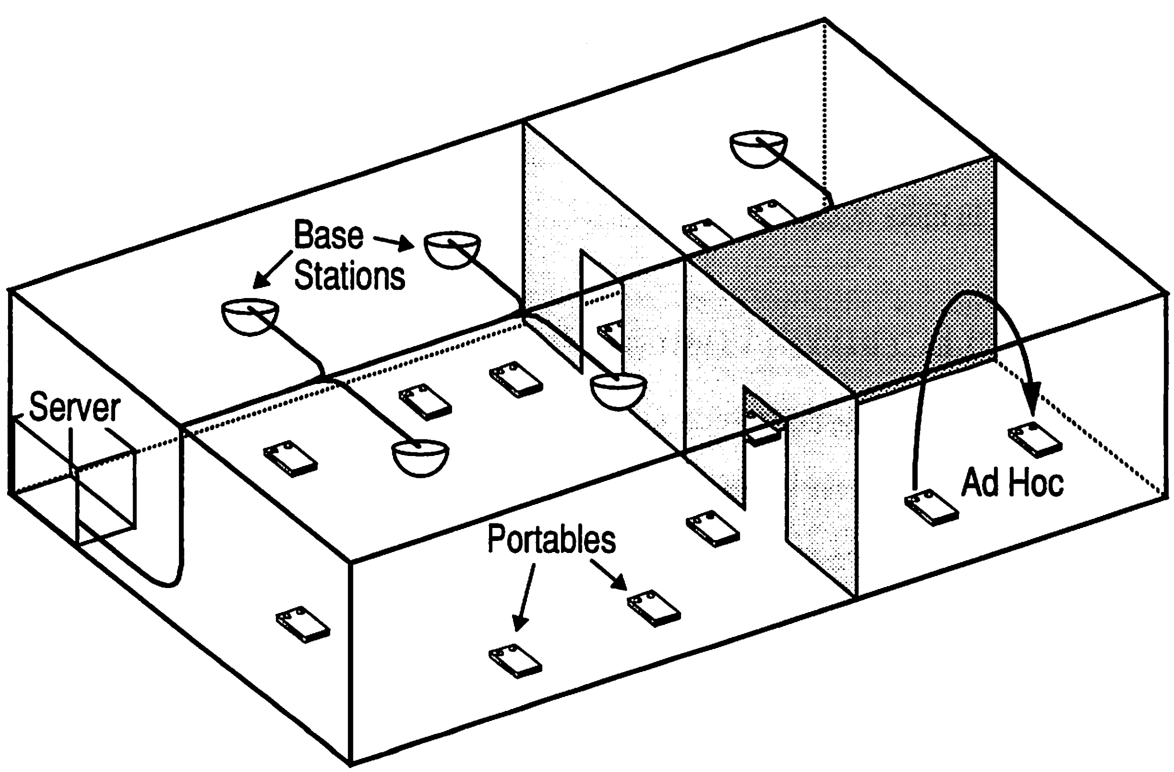


Fig. 1-2. An infrared local-area network.

Undoubtedly, some rooms will not have a base station, as illustrated in Fig. 1-2. In such cases two portable units should still be able to communication with one another, a concept referred to as *ad hoc networking*. This link would be of the non-directed non-LOS variety, with the ceiling being the prominent diffusive reflector.

Different wavelengths can be used for the up and down links to provide a full-duplex link without interference. Unfortunately, however, this approach is not conducive to ad hoc networking, because each portable would need the capability to detect two distinct wavelengths. An alternative to wavelength duplex is subcarrier duplex, in which the up and down links use different subcarrier frequencies. It is likely that the bit-rate requirements for the up and down links will be highly asymmetric; the down link bit rate can be quite high, since it would involve downloading large executable files, graphics, and video images, whereas the up link data rate will likely carry only key strokes, pen strokes, and voice. This observation eases the multiplexing task because it allows us to allocate a small fraction of the bandwidth to the up-link, say a few MHz of the subcarrier frequency band above a few hundred MHz, and leave the majority of the bandwidth for use by the down link. Subcarrier duplex may thus be preferable to wavelength duplex. Multiplexing issues are discussed in more detail in chapter 6.

1.3 OPTOELECTRONIC COMPONENTS

In choosing the transmitter for a high-speed non-directed link, the laser diode has a number of advantages over the light-emitting diode (LED). For example, laser diodes can be modulated faster, they convert electrical power to optical power more efficiently, and they can emit more optical power. Furthermore, the narrow linewidth of a laser diode relative to an LED allows the receiver to use a narrower optical filter and thus reject more background light. Therefore, when necessary, this dissertation assumes that laser diodes are used.

The two practical options for a photodetector are the p-intrinsic-n (PIN) diode and the avalanche photodiode (APD). The PIN diode is preferable to the APD because it is cheaper, easier to bias, and does not add noise of its own. As shown in chapter 3, it is always possible to approach shot-noise limited operation by careful receiver design, and hence the noisy gain of the APD is detrimental. For this reason we assume a PIN diode is used.

The wavelength band near 800 nm is attractive due to the availability of low-cost high-power GaAs laser diodes and large-area silicon PIN diodes. For specificity, we assume an operating wavelength of 810 nm when necessary.

1.4 OVERVIEW

To the author's knowledge, the highest speed achieved to date by a non-directed optical link is 4 Mb/s [21]. In contrast, this dissertation investigates the feasibility of non-directed links with higher speeds near 100 Mb/s. We identify the major impediments to achieving high speed, and present design strategies to counter them. Although most of the ideas presented here have not yet been verified experimentally, they suggest that high speeds near 100 Mb/s are indeed practical.

It should be emphasized that, although the wireless LAN discussed in section 1.2 was the motivation for the work in this dissertation, a complete design is beyond its scope. We concentrate instead on the physical-layer problem of establishing a robust point-to-point link using non-directed infrared radiation. This choice was guided by the observation that a high-performance network requires a high-performance physical layer. The physical layer is thus the natural place to start when pushing the limits of the medium. A second benefit of narrowing the scope in this way is that it opens up other applications of non-directed communication besides wireless LANs, such as mobile robot links and high-definition television broadcast. There is a danger, of course, in focusing on the physical layer without regard to higher level issues. For example, a modulation scheme well-suited for a point-to-point link will be inappropriate for a wireless LAN if it is not amenable to multiplexing and multiple access protocols. Therefore, to alleviate such concerns, we briefly address the interplay between the physical layer and higher layers in chapter 6.

This chapter and chapter 6 are the only chapters that discuss the LAN application of section 1.2; all other chapters examine the more fundamental problem of establishing a high-speed point-to-point non-directed link. The down link faces more technical challenges than the up link, primarily because the data rate of the down link is higher, and also because the complexity of the portable receiver is constrained by economic factors. The up link faces challenges of its own, of course, because of contention between multiple users and because of a limited transmitter power, but these appear less formidable at present. Therefore, this dissertation concentrates on the down link.

The primary impediment to high speed communication using non-directed links is a weak received signal relative to the potentially intense background light. In essence, the non-directed indoor channel is power limited. Chapter 2 addresses this issue in detail, and proposes design strategies for the transmitter and receiver optics that roughly maximize the received signal power while minimizing the detected background light. To collect sufficient signal power, the photodetector area must be large. Unfortunately, large-area photodetectors have high capacitance, which exacerbates the problem of designing a low-noise wideband preamplifier. Preamplifier design is addressed in chapter 3.

A second impairment to high speed communication, perhaps less severe than the first but important nevertheless, is intersymbol interference caused by multipath optical propagation. Chapter 4 presents a theoretical model for multipath propagation, the accuracy of which has been verified experimentally. It also examines the effects this multipath dispersion has on system performance.

In a practical non-directional link, data is encoded onto the transmitted infrared lightwave using intensity modulation, and it is recovered at the receiver using direct detection.

Chapter 5 evaluates the performance of various modulation schemes on this intensity-modulation channel, and shows how the intensity-modulation channel differs from a conventional radio or wire-based channel.

Neither of the two primary impediments to high speed, a weak signal power and multipath dispersion, is insurmountable. The key challenge to the system designer, however, is to overcome these impediments while meeting the stringent size, weight, and power-consumption requirements imposed by a portable computer.

CHAPTER 2

LINK ANALYSIS AND OPTICS DESIGN

In this chapter we present a link budget analysis for non-directed LOS optical communication, and present design procedures for optimizing the transmitter and receiver optics. We examine the signal-to-noise ratio (SNR) improvement of a hemispherical lens as an optical antenna. We propose a hemispherical thin-film optical filter, and compare its performance to that of traditional planar thin-film filters. For both filter types, we jointly optimize the transmitter radiation pattern, the filter orientation, and the filter bandwidth. The results of this chapter indicate that a 1-W transmitter and a 1-cm^2 photodetector are sufficient to achieve a 100-Mb/s non-directed link over a range of about 5 m, even in the high-noise case of brightly sky-lit environments.

2.1 INTRODUCTION

Because we intend to communicate using intensity modulation with direct detection, it is useful to think of the optical channel from the modulating signal at the transmitter to the photodetector output at the receiver as an equivalent baseband channel. This abstraction becomes quite useful in later chapters when we analyze optical multipath and consider modulation schemes and equalization schemes. For now, however, this abstraction is perhaps dangerous, because it makes us think of this equivalent channel as a black box, over which we have no control. In fact, however, a lot can happen between the laser output and the photodetector input, and the designer has a lot to say about what goes in the black box. This chapter focuses on the optical components in the system, with the goal of designing the optics so as to maximize the electrical signal-to-noise ratio at the output of the black box.

The results of this chapter can be applied — with little or no modifications — to all LOS systems, regardless of bit rate, modulation scheme, and cell size. For example, all LOS systems will benefit from the use of a hemispherical lens and a hemispherical filter, and all systems can use our procedure for jointly optimizing the transmitter radiation pattern and receiver optical filter. In keeping with our design goal, however, we present numerical results for a baseband on-off-keying system operating at a bit rate of 100 Mb/s.

The design procedure presented in this chapter can be summarized as follows:

1. Specify room size and maximum angle ψ_{max}.
2. Specify lens radius and refractive index (section 2.2).
3. Compute density functions $\{f_\psi^{(1)}(\theta)\}$ and $\{f_\psi^{(2)}(\theta)\}$ (section 2.4.2).
4. Optimize lens and detector antireflection coatings (section 2.4.3).
5. Jointly optimize transmitter radiation pattern with filter bandwidth $\Delta\lambda$ and orientation Θ (section 2.4.4, section 2.4.5).

The steps must be followed in this order. The sections in which these steps are discussed in detail are indicated in parenthesis. As discussed in section 2.4.2, the two families of density functions $\{f_\psi^{(1)}(\theta)\}$ and $\{f_\psi^{(2)}(\theta)\}$, parameterized by $\psi \in [0, \psi_{max}]$, characterize the angular distribution of light as it enters the lens through the curved surface ($f_\psi^{(1)}$) and as it exits the lens through the planar surface ($f_\psi^{(2)}$). Other unfamiliar terms are included here for reference; they will be defined later.

In the next section we examine the performance of a hemispherical lens as an optical antenna. Then, in section 2.3, we review the theory of thin-film optical filters. In section 2.4 we present the main results, including the link equation, optimization of the antireflection coatings, optimization of the transmitter, and optimization of the optical filter. We present results for both planar and hemispherical thin-film optical filters.

2.2 OPTICAL ANTENNA

The wide optical beam emitted by a non-directed transmitter results in a highly diluted signal irradiance (power per unit area) arriving at the receiver. To collect sufficient signal power, therefore, the receiver must use a large-area photodetector. We will see later that, at the output of the photodetector, the shot-noise power is directly proportional to the detector area, while the electrical signal power is proportional to the *square* of the detector area. The ratio of the two yields a shot-noise-limited electrical SNR that is directly proportional to detector area.

Unfortunately, the largest Si-PIN photodetectors available today have an area of only 1 cm^2. Fabricating larger photodiodes is undesirable not only because of the added expense, but also because of the bandwidth limitations caused by the resulting larger detector capacitance (note that capacitance is proportional to area).

An obvious alternative to fabricating larger photodiodes is to use a wide-field-of-view optical antenna to increase the photodetector's effective area. One way to achieve optical gain over a wide field of view is to use an array of narrow-field-of-view non-imaging concentrators, each pointing in a different direction; this approach was adopted by MPR Teltech for the base stations in their multichannel telephone prototype [22].

A second alternative, better-suited for a low-cost portable receiver, is a hemispherical lens. Its benefit in the context of non-directed communication was first noted by Kotzin and Marhic [25][26]. As we shall see, optical gains of more than 4 dB are practical. In this section we calculate the gain provided by a hemispherical lens for arbitrary angles of incidence. We will show that, neglecting reflection losses, the gain of a hemispherical lens of index n is approximately n^2, regardless of the angle of incidence.

2.2.1 Hemispherical Lens

Consider a hemispherical lens with refractive index n and radius R placed directly upon a circular photodiode with area $A_{det} = \pi r^2$ as shown Fig. 2-1 [25][26][27]. This configuration is commonly used in solar cell applications [28]. We will assume that the far-field radiation of the transmitter is a wide collimated beam with uniform irradiance (power per unit area), and that the beam makes an angle ψ with respect to the normal of the photodetector surface (see Fig. 2-1-b).

With the aid of Fig. 2-1-b, the calculation of the gain proceeds as follows. Define $A_1 = A_{det}\cos(\psi)$ so that, without the lens, light passing through (at normal incidence) a planar region with area A_1 will eventually hit the detector. Similarly, define A_2 so that, with the lens, light passing through (at normal incidence) a planar region with area A_2 will

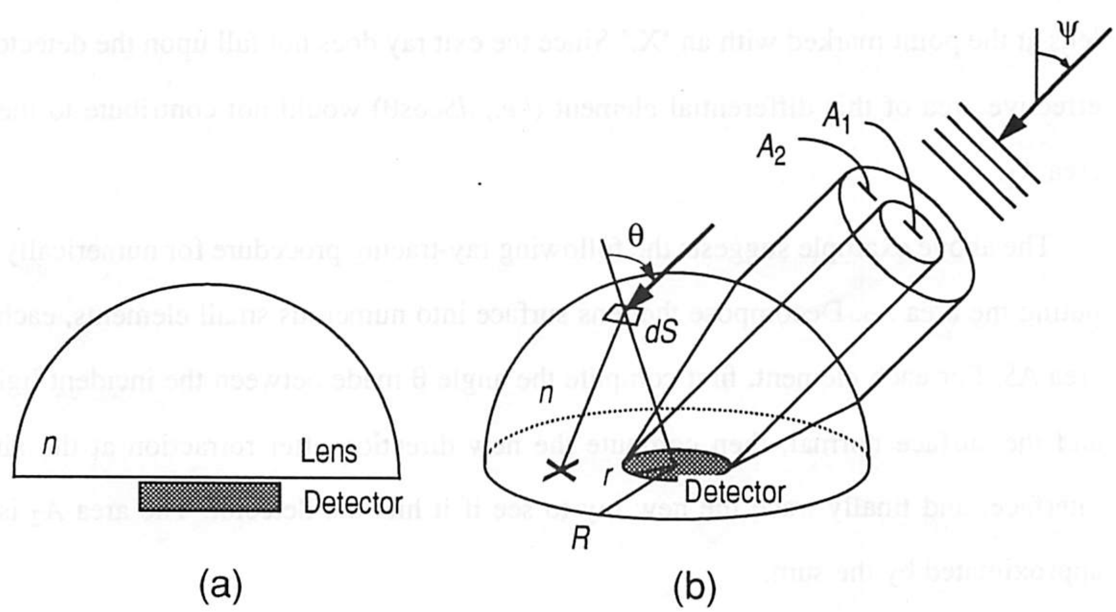


Fig. 2-1. A hemispherical lens; (a) cross-sectional view; (b) ray-tracing model.

eventually hit the detector. Due to refraction, A_2 will be greater than A_1, and the optical gain is simply the ratio:

$$G_\psi = \frac{A_2}{A_1} . \tag{2-1}$$

The area A_2 is given by:

$$A_2 = \int_{S_0} \hat{i} \cdot d\hat{S} , \tag{2-2}$$

where $\hat{i}$ is the direction of the incident light rays, $d\hat{S}$ is a vector-valued differential element of the lens surface with direction equal to the surface normal, and S_0 is the region of the lens surface for which light passing through will eventually hit the detector. For now, we are neglecting the loss due to reflections at the air-lens interface. We will account for reflection losses in section 2.4.3.

As an example, consider the differential surface element dS shown in Fig. 2-1-b. Light incident on this element can be modeled as a single ray making an angle θ with the lens surface normal. After refraction at the air-lens interface, this ray is redirected and exits the lens at the point marked with an 'X.' Since the exit ray does not fall upon the detector, the effective area of this differential element (*i.e.*, $dS\cos\theta$) would not contribute to the total area A_2.

The above example suggests the following ray-tracing procedure for numerically computing the area A_2. Decompose the lens surface into numerous small elements, each with area ΔS. For each element, first compute the angle θ made between the incident light ray and the surface normal, then compute the new direction after refraction at the air-lens interface, and finally trace the new ray to see if it hits the detector. The area A_2 is then approximated by the sum:

$$A_2 \approx \sum_{S_0} \cos(\theta)\, \Delta S . \tag{2-3}$$

2.2.2 Gain at Normal Incidence

For the special case of normal incidence ($\psi = 0$), the above ray-tracing algorithm can be avoided. Consider Fig. 2-1-b with $\psi = 0$; in this case, both region A_1 and region A_2 are circular. In fact, A_1 is just the detector area $A_{det} = \pi r^2$. Noting that the radius of A_2 is in this case $r\sqrt{G_0}$, where G_0 is the optical gain at normal incidence, then a simple geometric argument leads to the following relationship:

$$r = r\sqrt{G_0} - \sqrt{R^2 - r^2 G_0}\tan\left(\operatorname{asin}\frac{r\sqrt{G_0}}{R} - \operatorname{asin}\frac{r\sqrt{G_0}}{nR}\right) . \tag{2-4}$$

This transcendental equation can be inverted numerically to find G_0 as a function of n, r, and R. In Fig. 2-2 we plot the gain at normal incidence versus lens radius for indices

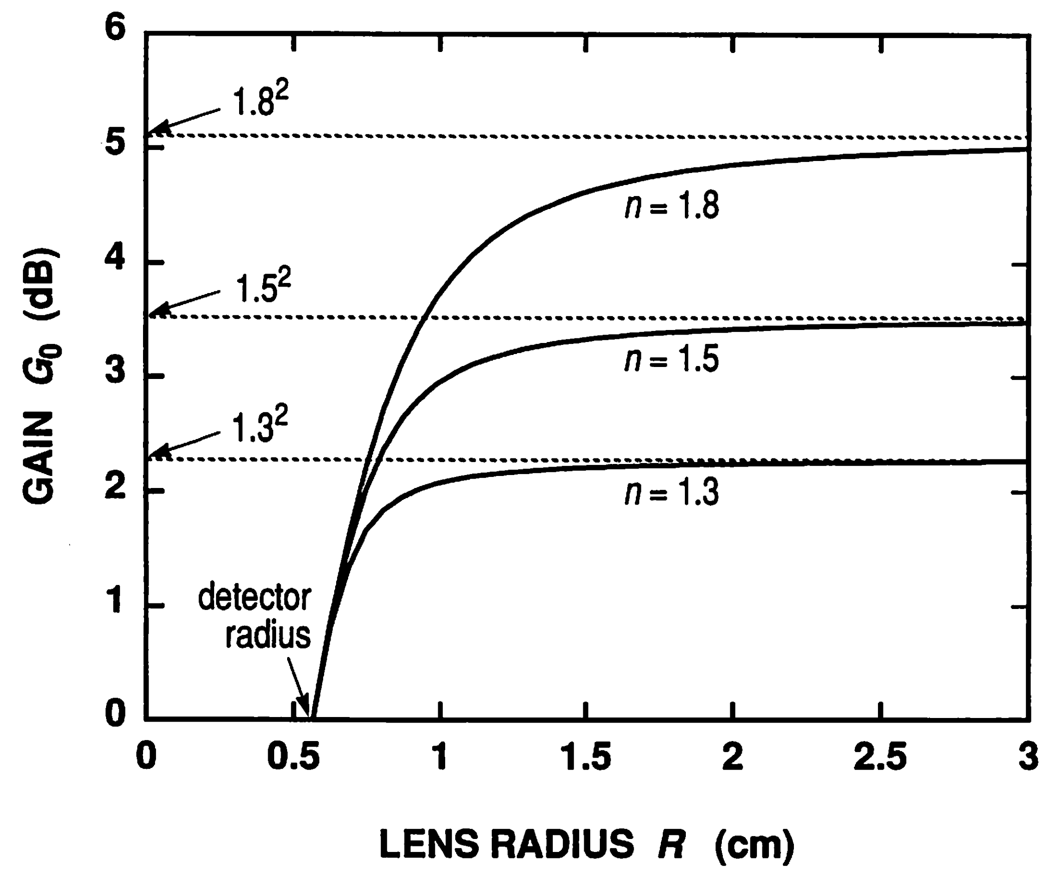


Fig. 2-2. Dependence of gain at normal incidence on lens radius, assuming $A_{det} = 1$ cm^2, $\psi = 0$, and $n \in \{1.3, 1.5, 1.8\}$.

$n \in \{1.3, 1.5, 1.8\}$, assuming a 1-cm^2 detector. We see that the gain is a monotonically increasing function of lens radius. Furthermore, for large radii, the gain approaches an asymptote of n^2, the thermodynamic limit for passive concentrators [25][28]. Inspection of Fig. 2-2 reveals the following rule of thumb: most of the asymptotic gain is achieved when the lens radius R roughly satisfies $R > n^2 r$.

2.2.3 Gain at Non-Normal Incidence

The symmetry that led to (2-4) breaks down when ψ is nonzero, in which case we must resort to the numerical method described in section 2.2.1 to compute the gain. To illustrate the effect of the angle of incidence on gain, we plot G_ψ versus ψ in Fig. 2-3 for the following typical parameters: lens index $n = 1.8$, detector area $A_{det} = 1$ cm^2, and lens radius $R = 2$ cm^2. The dotted line is n^2, the normal-incidence gain assuming infinite lens

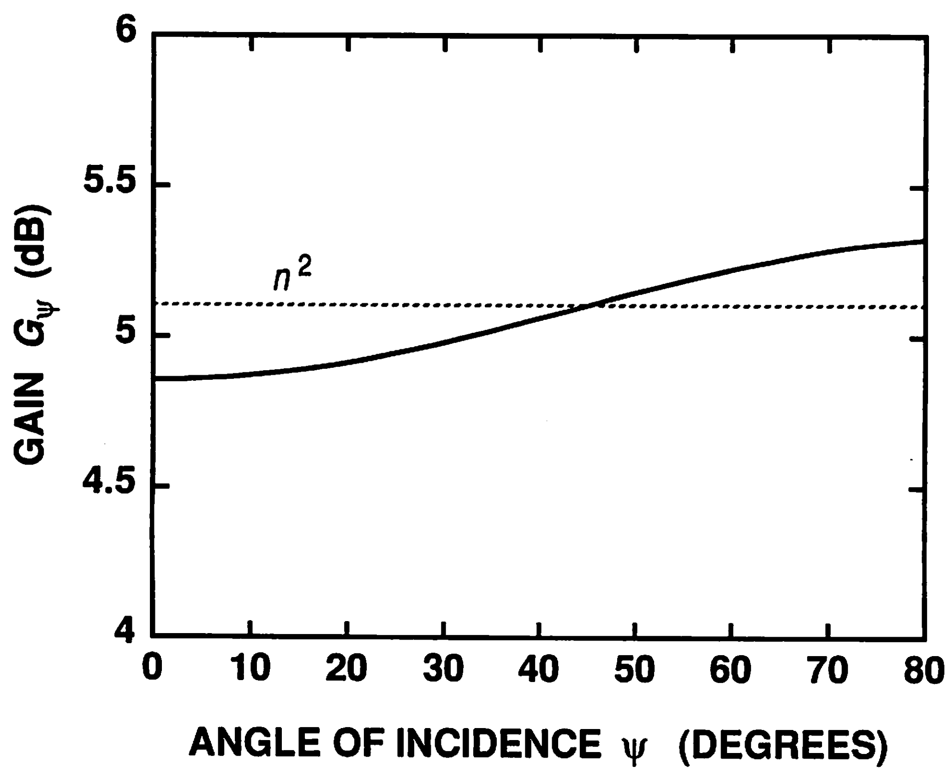


Fig. 2-3. Dependence of gain on angle of incidence, assuming $A_{det} = 1$ cm^2, $R = 2$ cm, and $n = 1.8$.

radius. The gain is seen to vary only slightly with angle, always staying within 0.3 dB of n^2.

It is interesting to consider the behavior of the gain as the angle of incidence ψ approaches $\pi/2$. Depending on the size of the lens relative to the detector, the gain can approach either a finite value or an infinite value. Consider Fig. 2-1-b with $\psi = \pi/2$. The light rays incident near the very top of the lens will refract at the air-lens interface and exit the bottom of the lens at a distance r_{min} from the center of the detector, where from Snell's law:

$$r_{min} = \frac{R}{\sqrt{n^2-1}}. \tag{2-5}$$

It can be shown that this is as close as any ray will get to the center of the detector. Thus, we conclude that there will be detected light (i.e., $A_2 > 0$) only when $r_{min} < r$, or equivalently, when the lens radius satisfies $R < \sqrt{n^2-1}\,r$. Furthermore, because $A_1 = \cos(\psi) = 0$ when $\psi = \pi/2$, the gain in this case will be infinite. In Fig. 2-3, for example, the lens radius exceeds $\sqrt{n^2-1}\,r$, and hence the gain is finite at $\psi = \pi/2$.

2.2.4 Reflections

For simplicity, we have ignored reflections in the previous sections. In practice, however, losses due to reflections will be significant, and hence cannot be ignored. To illustrate the effects of reflections, consider a hemispherical lens of index n placed above a photodetector, as shown in Fig. 2-1-a. Since the lens cannot be fabricated directly upon the detector, there will likely be a small gap between the two. If this gap were filled with air, then light trying to exit the bottom of the lens would experience total internal reflection (100% reflectivity) at angles of incidence greater than the critical angle $\theta_c = \text{asin}(1/n)$, which for a 1.8-index lens is about 34°. This results in a narrow field of view that is unacceptable for our non-directed application. The effects of reflections can be reduced significantly by the careful placement of anti-reflection coatings and index-matching gel within the system. We defer further discussion of these issues until section 2.4.3, because in the next section we introduce another optical component (the optical band-pass filter) that will affect our design strategies.

2.3 THIN-FILM OPTICAL FILTERS

In Fig. 2-4 we plot the power spectral densities for the three most common illumination sources: fluorescent light, sunlight, and incandescent light [11][29]. We see that a significant fraction of the powers of both sunlight and incandescent light are in the infrared region near 810 nm, the wavelength at which we propose to operate. Silicon photodiodes are sensitive to light over a broad band of wavelengths, spanning roughly from 600 nm to over 1000 nm. Without an optical filter to reject out-of-band ambient light, therefore, the receiver would be swamped by shot noise.

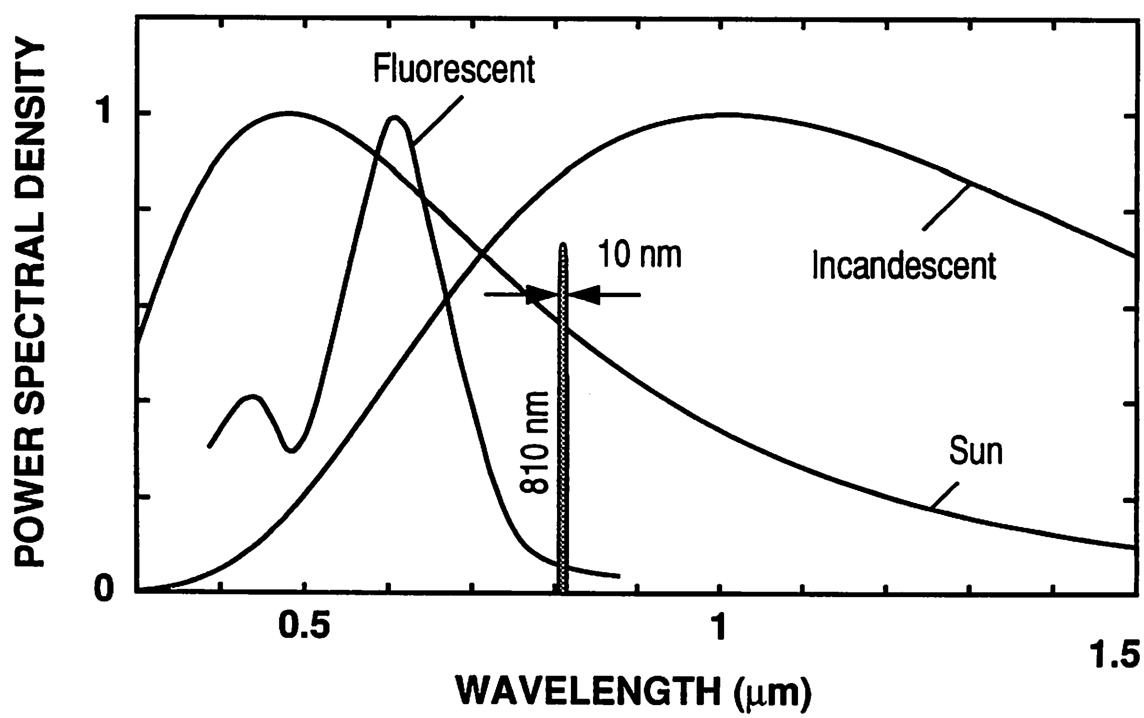


Fig. 2-4. Power spectral densities for common light sources [11][29].

Most narrow-band optical filters are thin-film devices. They are desirable because they are inexpensive and also because they can be deposited onto other components within the system, keeping the total component count low. Unfortunately, however, because their operation is based upon the principle of optical interference, their filter characteristics change with the input's angle of incidence. This angle dependence is critical in our wide field-of-view application. The purpose of this section, therefore, is to explore the angle-dependent performance of thin-film optical filters. We first present a brief synopsis of the theory of thin-film filters, and then present a simple four-parameter model for later use.

2.3.1 Theory

A thin-film optical filter consists of a stack of thin dielectric slabs with varying indices of refraction. In Fig. 2-5 we illustrate a light ray passing through a stack of $K-2$ dielectrics. The index of the input medium is n_1, the index of the output medium is n_K, and the indices of the dielectric slabs, from top to bottom, are n_2 through n_{K-1}. In this subsection we will review the theory of lightwave reflections at dielectric boundaries for glancing angles of incidence, and show how the dielectric stack of Fig. 2-5 can be made to perform as a narrow-band optical filter. The theory presented here is not new; it is based on the transmission-line analogy for reflections of electromagnetic waves, as popularized in [30]. We include it here because its usefulness in the context of thin-film filters has apparently been overlooked in texts on the subject, such as [31][32].

As shown in Fig. 2-5, the angle θ_k is defined as the angle between the light ray and the normal to the dielectric boundary in the k-th medium. The angles θ_k, $k \in \{2,\ldots,K\}$, can be computed recursively from θ_1 using Snell's law:

$$\theta_k = \sin^{-1}\left(\frac{n_{k-1}}{n_k}\sin\left(\theta_{k-1}\right)\right) . \tag{2-6}$$

For a lightwave incident to a dielectric boundary at non-normal incidence, define an "effective" index of refraction N_k for the k-th medium by [30][31][32]:

$$N_k = \begin{cases} n_k / \cos(\theta_k) & \text{for TE} \\ n_k \cos(\theta_k) & \text{for TM} \end{cases} \tag{2-7}$$

where TE and TM, which signify *transverse electric* and *transverse magnetic* plane waves, respectively, are the two orthogonal states of polarization for the input lightwave. Using the transmission-line analogy, define a complex-valued "load" index of refraction $N_{L,k}$ for the k-th medium as the effective index "seen" by the lightwave as it enters medium k, for $k \in \{2,3,\ldots,K-1\}$, given by [30]:

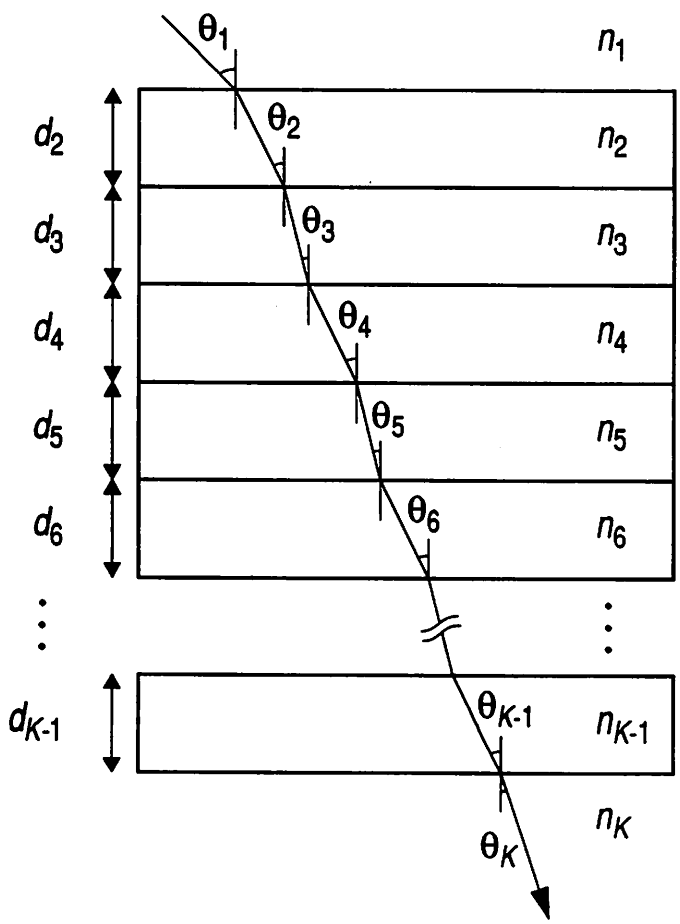


Fig. 2-5. Stack of thin-film dielectrics.

$$N_{L,k} = N_k \frac{N_{L,k+1}\cos\beta_k + jN_k\sin\beta_k}{N_k\cos\beta_k + jN_{L,k+1}\sin\beta_k}, \tag{2-8}$$

where:

$$\beta_k = 2\pi\frac{n_k d_k}{\lambda}\cos(\theta_k), \tag{2-9}$$

d_k is the slab thickness for the k-th dielectric, and λ is the wavelength of the light (in a vacuum). Starting with $N_{L,K} = N_K$, we can apply (2-8) recursively to arrive at $N_{L,2}$, the effective load index seen by the input lightwave as it enters the first slab of the dielectric stack. The reflection coefficient at the filter input is then simply [30]:

$$\rho = \frac{N_1 - N_{L,2}}{N_1 + N_{L,2}}. \tag{2-10}$$

The initialization of the N_k as either TE or TM, as specified in (2-7), determines whether the above procedure yields ρ_{TE} or ρ_{TM}. For the case when equal power is contained in the TE and TM polarization modes, the total fraction of power reflected at the input to the filter is given by:

$$R_{TOT} = \frac{1}{2}(\,|\rho_{TE}|^2 + |\rho_{TM}|^2\,). \tag{2-11}$$

Assuming perfect (lossless) dielectrics, the filter transmission T is related to the total reflectivity by $T = 1 - R_{TOT}$.

Once the filter is specified (via $\{n_k, d_k\}$ for $k \in \{2,3,\ldots,K-1\}$), the above procedure yields the total filter transmission T as a function of the input angle of incidence (θ_1) and the wavelength (λ). Despite the simplicity of this procedure, the equations become quite unwieldy for stacks with more than a few dielectric slabs. Fortunately, however, the procedure is easily implemented on a computer.

2.3.2 Filter Design

A full treatment of the design of thin-film filters is beyond the scope of this thesis. Instead, we examine the performance of a typical thin-film filter.

The basic building block of a thin-film passband filter is the Fabry-Perot filter, which is in turn based on the Fabry-Perot interferometer. A Fabry-Perot interferometer consists of two highly reflective mirrors that are facing each other and are spaced a distance d apart [31][33]. It acts as a comb filter with a series of passbands at wavelengths equal to $2d/l$ for all integers l. The thin-film all-dielectric Fabry-Perot filter can be visualized as a Fabry-Perot interferometer with the mirrors replaced by distributed reflectors consisting of a stack of alternating high- and low-index dielectrics, each a quarter-wave thick. One such reflector, a three-layer stack consisting of a low index quarter-wave slab sandwiched between two high-index quarter-wave slabs, will be denoted HLH. Extending the notation in the obvious way, the first-order ($l = 1$) Fabry-Perot filter can be written as $HLHL^2HLH$, where L^2 denotes a low-index "spacer" layer that is one half-wavelength (two quarter-wavelengths) thick.

To achieve narrow optical bandwidths below 100 nm, two or more Fabry-Perot filters are typically coupled together [31][32][34]. In Fig. 2-8 we show a typical three-cavity bandpass filter given by $(LH)^2L^2(HL)^4(LH)^4L^2(HL)^2$ [32]. The shaded slabs represent the high-index layers. The spacer layers are one half-wavelength thick and all other layers are one quarter-wavelength thick. In other words, if n_L, n_H, d_L, and d_H are the refractive indices and thicknesses of the low-index and high-index layers, respectively, then $n_L d_L = n_H d_H = \lambda_{normal}/4$ for some wavelength λ_{normal}, so-named because it is the wavelength of peak transmission for light at normal incidence. As shown in the figure, the composite filter can be viewed as three Fabry-Perot filters separated by low-index coupling layers.

The characteristics of this filter can be calculated using the theory of section 2.3.1. In Fig. 2-7 we plot filter transmission versus wavelength for a number of different angles of incidence, assuming $n_L = 2$, $n_H = 3.5$, $\lambda_{normal} = 850$ nm, and the refractive indices of the input and output media (n_1 and n_K of Fig. 2-5) are unity. The curves are labeled with the angle of incidence θ (or θ_1 in Fig. 2-5). At normal incidence, the filter is well-approximated by a first- or second-order Butterworth filter with center wavelength $\lambda_{normal} = 850$ nm and bandwidth $\Delta\lambda = 31$ nm. As the angle of incidence increases to 30°, the spectral shape and bandwidth remain unchanged, while the center wavelength shifts to

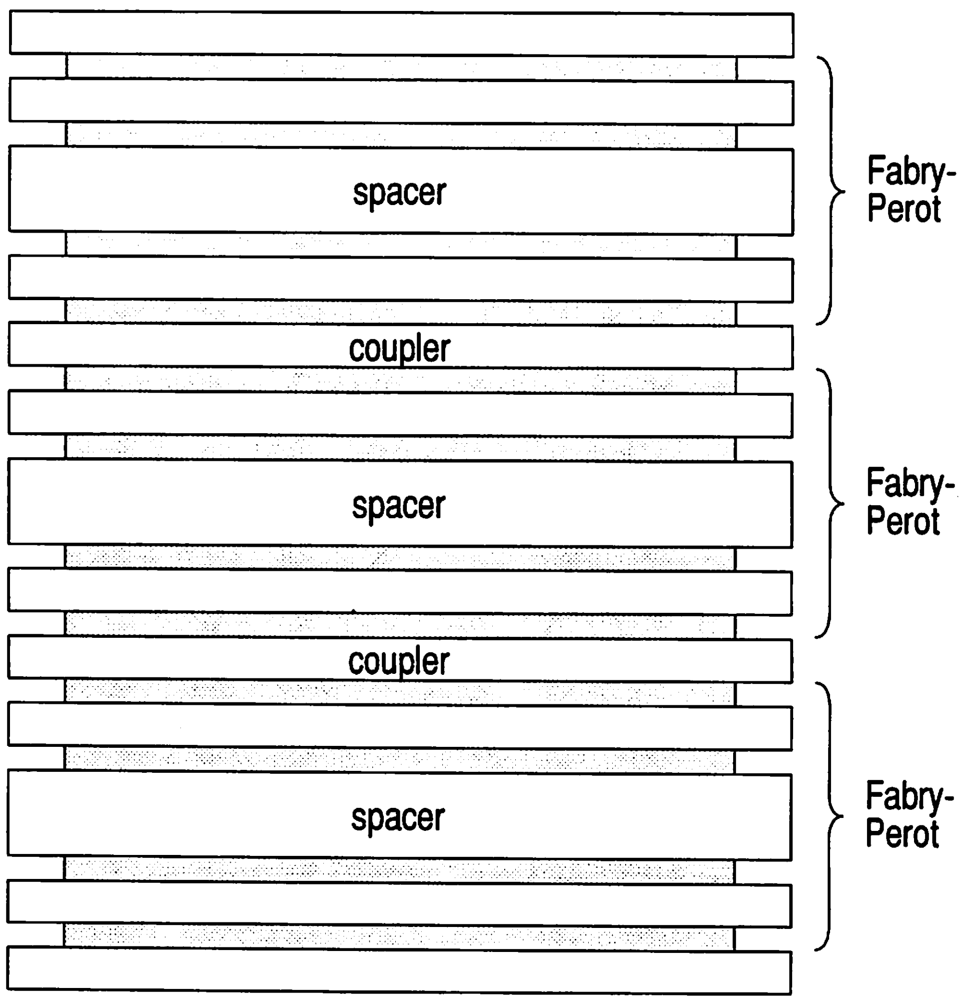


Fig. 2-6. Schematic diagram of a 25-layer, three-cavity thin-film filter [32], with $n_L = 2$, $n_H = 3.5$, $\lambda_{normal} = 850$ nm, and $n_1 = n_K = 1$.

lower wavelengths. For angles of incidence greater than about 30°, however, the spectral shape is seen to change considerably.

Define $\lambda_{pk}(\theta)$ as the wavelength of peak transmission for angle of incidence θ. It satisfies $\lambda_{pk}(0) = \lambda_{normal}$. In Fig. 2-8 we plot the fractional shift $\lambda_{pk}(\theta)/\lambda_{normal}$ versus θ for the filter of Fig. 2-7. The dashed curve in the figure represents the following analytic approximation [34]:

$$\lambda_{pk}(\theta) = \lambda_{normal}\sqrt{1 - (n_1/n^*)^2 \sin^2(\theta)}, \qquad (2\text{-}12)$$

where n_1 is the index of the input medium (see Fig. 2-5), and n^* is an effective index for the spacer layer; it can be found empirically by fitting the approximate curve to the actual one in Fig. 2-8 [31][34], with the result being $n^* = 2.276$. The figure shows that this analytic approximation is quite accurate.

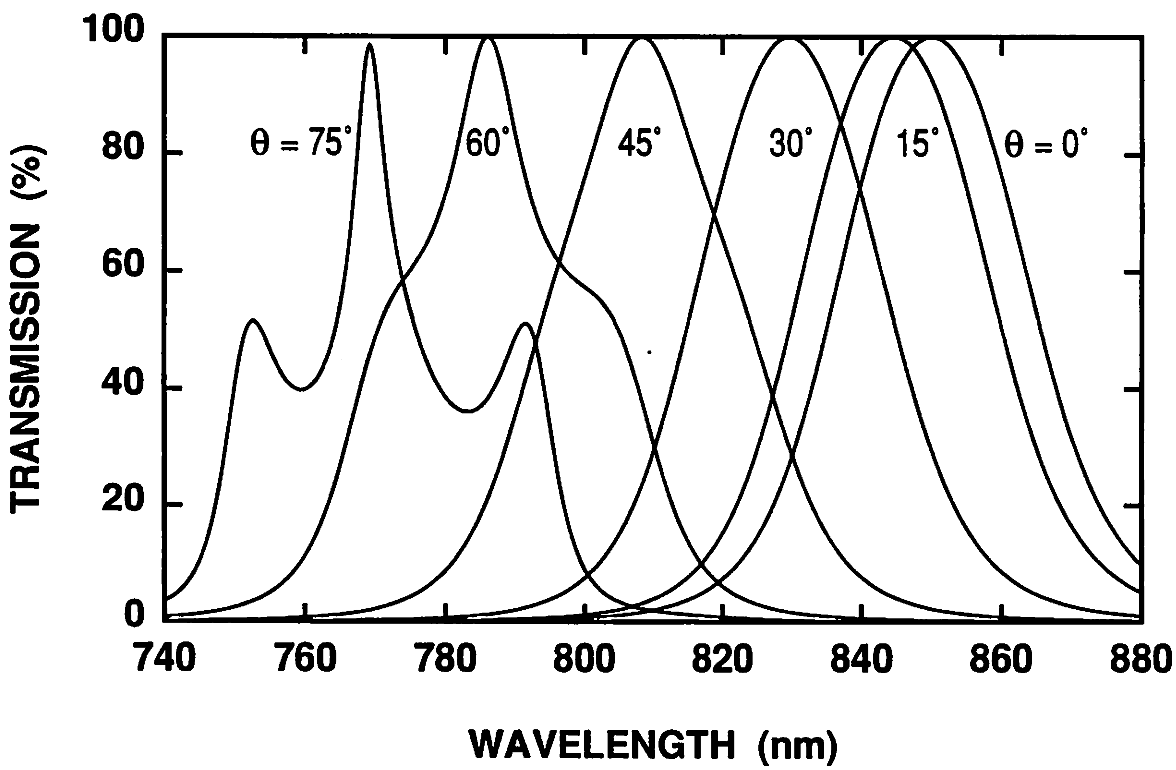


Fig. 2-7. Filter transmission for filter of Fig. 2-6.

The tendency for the peak wavelength to shift to shorter wavelengths at non-normal incidences makes it difficult to obtain an optical filter that has both a narrow passband and a wide field of view. As we explore in section 2.4.5, there is an optimal bandwidth that trades off the opposing goals of minimizing the admitted noise and maximizing the field of view.

2.3.3 Five-Parameter Model

In this subsection we seek a simplified model for thin-film filters which extracts only those features that are important to system design. We assume that the spectral shape of the filter has an m-th order Butterworth characteristic, and that this spectral shape remains

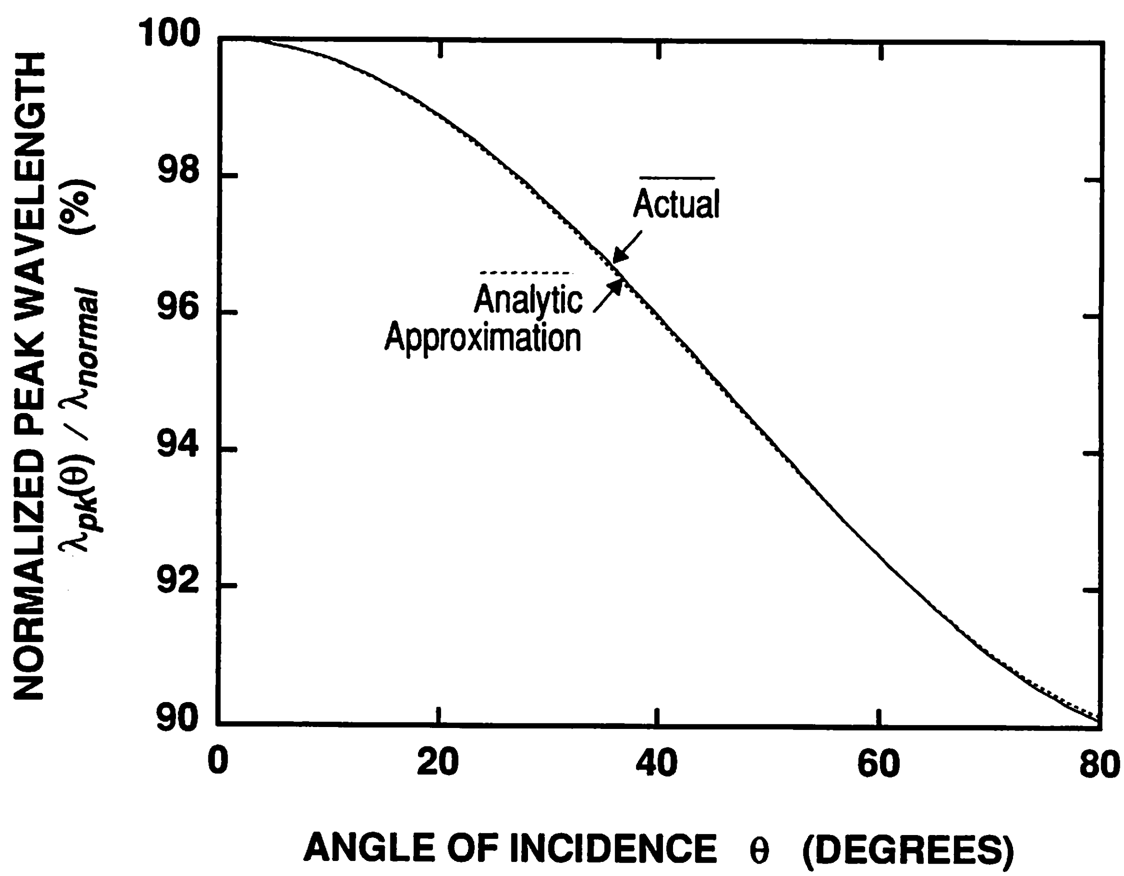


Fig. 2-8. Dependence of peak wavelength on angle of incidence for the filter of Fig. 2-7. Also shown is the analytic approximation of (2-12) with $n_1 = 1$ and $n^* = 2.276$.

the same for all angles of incidence. (The validity of this assumption is examined in section 2.3.4.) We assume that the peak wavelength shift is given by (2-12). Finally, we assume that the peak transmission is T_0. Although $T_0 = 1$ in the previous sections, practical filters use metal-dielectric "blocking" filters to reject the transmission peaks at wavelengths larger than the primary peak at λ_{normal}, and these blocking filters have an inherent loss [31][34]. Imperfect dielectrics, especially in the spacer layers, can also reduce the peak transmission [32]. Together, these effects result in T_0 ranging from 0.4 to 0.9 [35].

With these assumptions, the performance of a thin-film optical filter can be specified by the following five parameters:

$$\begin{aligned} \Delta\lambda &= \text{Full-width half-maximum bandwidth} \\ \lambda_{normal} &= \text{Peak wavelength at normal incidence} \\ m &= \text{Butterworth order} \\ n^* &= \text{Effective index of spacer layer} \\ T_0 &= \text{Peak transmission.} \end{aligned} \tag{2-13}$$

For light with wavelength λ incident at an angle θ, therefore, the filter transmission is given by:

$$T(\lambda,\theta) = \frac{T_0}{1 + \left(\frac{\lambda - \lambda_{pk}(\theta)}{\Delta\lambda/2}\right)^{2m}}. \tag{2-14}$$

In our application, the signal wavelength will always be λ_0. Define Θ as the "orientation" of the filter, such that $\lambda_{pk}(\Theta) = \lambda_0$. In other words, Θ is the angle at which the filter transmission is maximum for incident light with wavelength λ_0. With the analytic approximation of (2-12), Θ is related to λ_{normal} by:

$$\Theta = \lambda_{pk}^{-1}(\lambda_0) = \sin^{-1}\left(\frac{n^*}{n_1}\sqrt{1 - (\lambda_0/\lambda_{normal})^2}\right) \quad . \tag{2-15}$$

The parameters Θ and λ_{normal} can thus be used interchangeably. In practice, however, Θ is the more useful quantity. This is because its optimal value Θ_{opt}, as defined in section 2.4, is nearly independent of n^* and $\Delta\lambda$, whereas the corresponding optimum value for λ_{normal} is a strong function of both n^* and $\Delta\lambda$. For this reason we will usually use Θ in the sequel.

The filter transmission for wavelength λ_0 thus reduces to a single function of θ, parameterized by $\Delta\lambda$ and Θ:

$$T_{\Theta}^{\Delta\lambda}(\theta) = \frac{T_0}{1 + \left(\frac{\lambda_0 - \lambda_{pk}(\theta)}{\Delta\lambda/2}\right)^{2m}}. \tag{2-16}$$

The design of the optical filter thus boils down to specifying the two parameters $\Delta\lambda$ and Θ. The other three parameters (n^*, m, and T_0) are generally fixed by the technology and should be chosen as large as possible while meeting cost constraints. In section 2.4 we present a design procedure for optimizing $\Delta\lambda$ and Θ.

2.3.4 Polarization Effects

In Fig. 2-9 we compare the actual filter transmission with that predicted by the five-parameter model of (2-14). We see that the spectral shape of the passband begins to change for angles of incidence above about 30°. At $\theta = 75°$, the shape has broadened considerably and exhibits severe passband ripple. The cause of this variation is the "polarization effect;" the rates of change of peak wavelength as a function of θ are different for the TE and TM polarization modes [35].

Fortunately, polarization effects are not critical in our application, because we are not concerned with the filter shape at all wavelengths. Rather, we need only concern ourselves with the filter transmission at the operating wavelength λ_0. For example, inspection of Fig. 2-9 reveals that λ_0 near 810 nm would minimize the worst-case loss for $\theta \in [0, 75°]$. In Fig. 2-10 we plot the filter transmission at λ_0 = 810 nm versus θ for the filter of

Fig. 2-6. The dashed curve shows the transmission predicted by the analytical model of (2-16). The agreement between the two is good, and so in later sections we will use the five-parameter model of (2-16). This will greatly simplify the filter optimization procedure of section 2.4, because it will allow us to alter the filter bandwidth and normal-incidence wavelength without redesigning a new filter from scratch.

2.4 OPTIMIZATION OF TRANSMITTER AND FILTER

Choosing the best characteristics of the optical filter requires knowledge of the transmitter radiation pattern. The reverse is true as well, in that the best transmitter radiation pattern depends on the characteristics of the receiver optical filter. The best approach, therefore, is to jointly optimize both the transmitter radiation pattern and the optical filter,

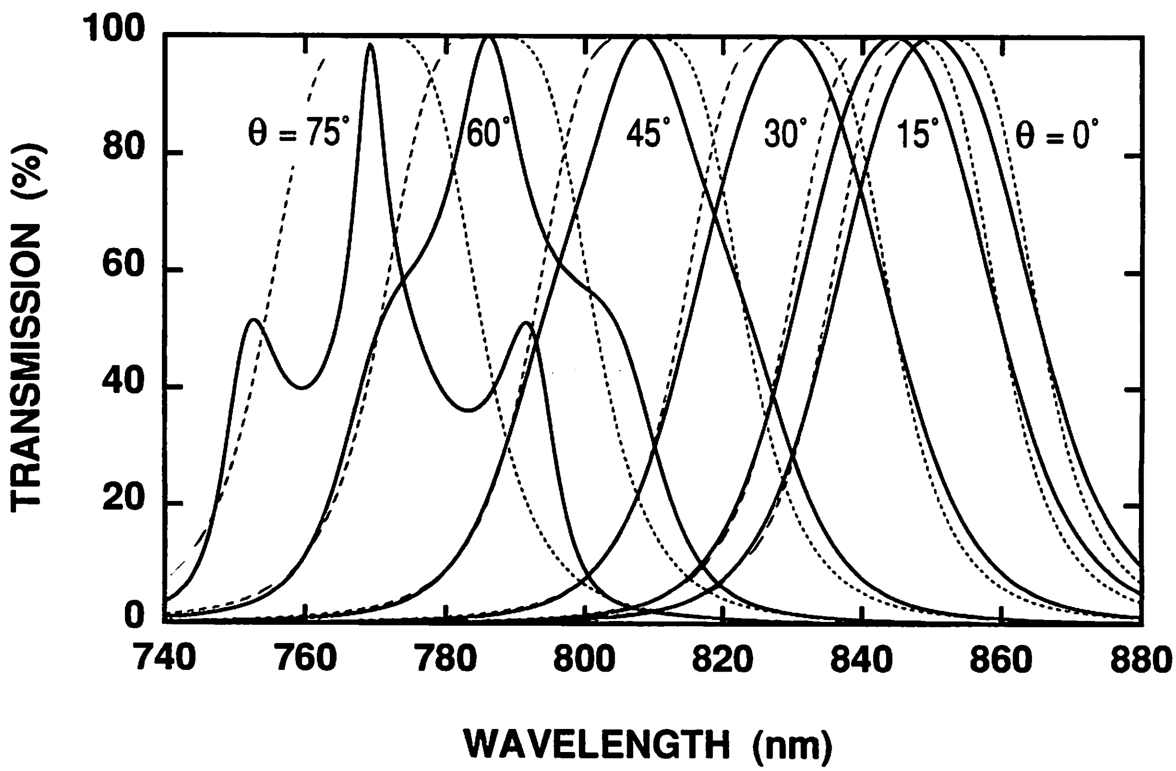


Fig. 2-9. Comparison between actual transmission of Fig. 2-6 and analytical model of (2-14) with $\Delta\lambda$ = 31 nm, λ_{normal} = 850 nm, n^* = 2.267, m = 2, and T_0 = 1.

which is the purpose of this section. We consider two alternatives for the optical filter: a planar filter, placed between the lens and the photodetector as shown in Fig. 2-11-a, and a hemispherical filter, deposited on the outer surface of the lens as shown in Fig. 2-11-b.

2.4.1 Figure of Merit

The key features of a cell are illustrated in Fig. 2-12, where we show a room cross-sectional view formed by a vertical plane passing through the center of the room and one corner of the room. The transmitter is assumed to be on the ceiling in the center of the cell. The shaded region represents the coverage area, and is specified by three dimensions:

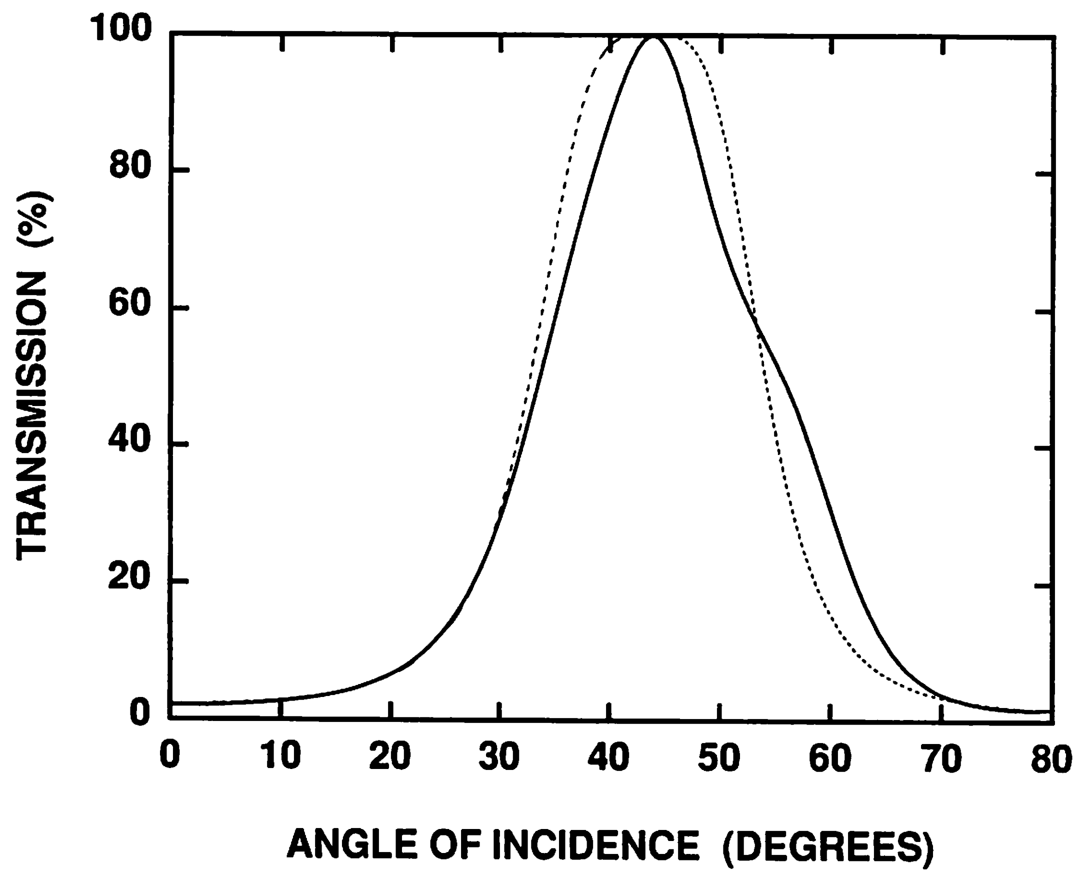


Fig. 2-10. Filter transmission at λ_0 = 810 nm as a function of angle of incidence for filter of Fig. 2-7: actual (solid curve) and analytical model (dashed curve: (2-16) with $\Delta\lambda = 31$ nm, $\lambda_{normal} = 850$ nm, $n^* = 2.267$, $m = 2$, and $T_0 = 1$).

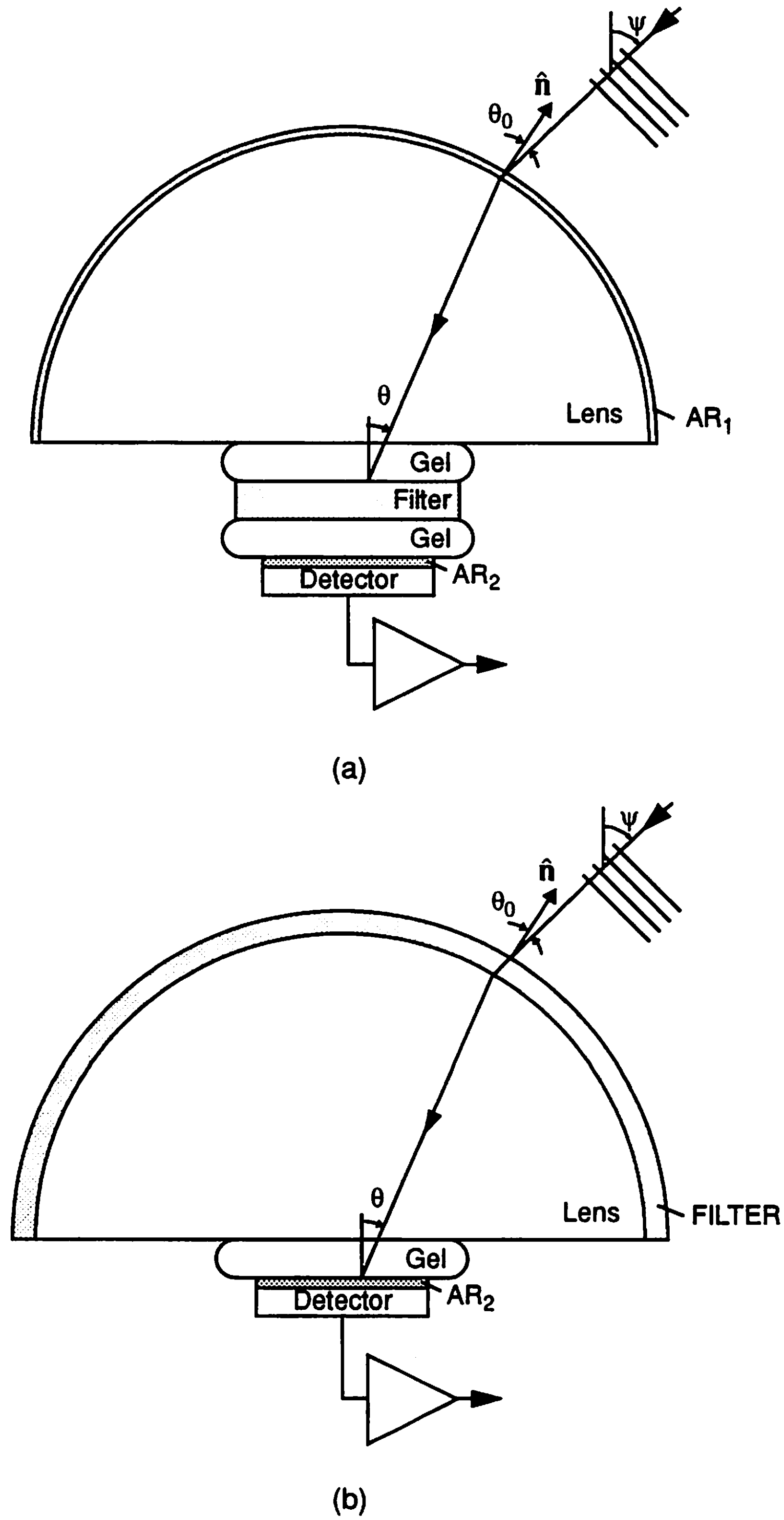


Fig. 2-11. Proposed receiver optics for: (a) planar filter; (b) hemispherical filter.

h_{min}, h_{max}, and d_{max}, where h_{min} and h_{max} are the minimum and maximum possible vertical distances between the transmitter and receiver, respectively, and d_{max} is the maximum horizontal distance between the transmitter and receiver. For illustration purposes, we will assume a 5 m × 5 m room with a 3-m ceiling, in which case these parameters are $h_{min} = 1.5$ m (roughly neck level), $h_{max} = 2.4$ m (roughly lap level), and $d_{max} = 2.5\sqrt{2}$ m ≈ 3.54 m (for the corner of the room).

In chapter 3 we will show that, for a well-designed receiver, shot noise from ambient light is the dominant source of noise. We therefore consider only shot noise in this section, and defer treatment of other noises such as electrical amplifier noise to chapter 3. Ambient light levels in typical offices are quite high, so the shot noise is well-modeled as a white Gaussian process. We can thus safely use the signal-to-noise ratio (SNR) as a performance metric. The peak SNR assuming shot-noise-limited operation is:

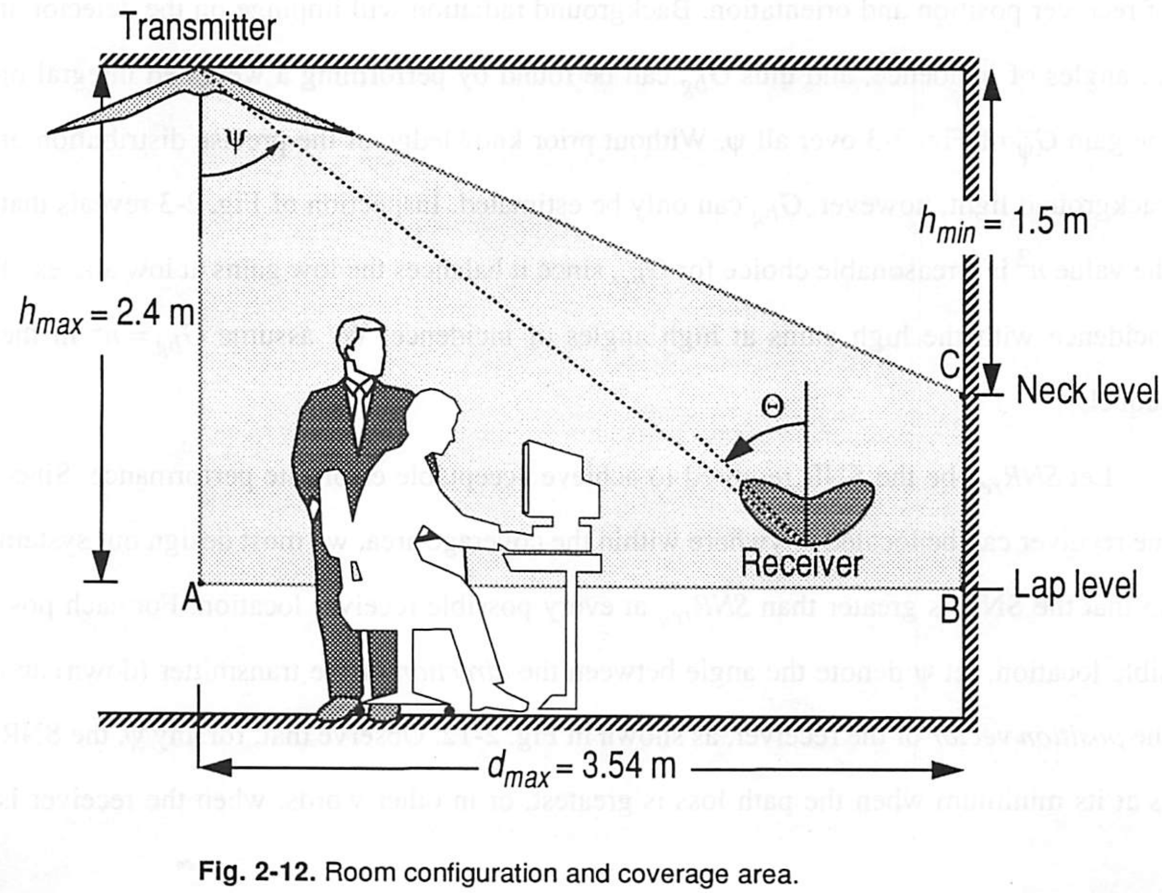


Fig. 2-12. Room configuration and coverage area.

$$SNR_{shot} = \frac{r^2 P_{sig}^2}{\sigma_{shot}^2}, \tag{2-17}$$

where r is the photodetector responsivity (A/W), P_{sig} is the peak received signal power, and σ_{shot}^2 is the power of the shot-noise component of the photodetector current [11]:

$$\sigma_{shot}^2 = I_2 B 2qrG_{bg} p_{bg} A_{det} \Delta\lambda , \tag{2-18}$$

where q is the charge of an electron (C), G_{bg} is the optical gain of the lens as seen by the background radiation, p_{bg} (W/(m^2·nm)) is the irradiance of the background light per unit bandwidth, $\Delta\lambda$ is the filter bandwidth, I_2 is a noise bandwidth factor, and B is the symbol rate [11]. For a given gain G_{bg}, area A_{det}, and filter bandwidth $\Delta\lambda$, the total detected background power P_{bg} is related to p_{bg} by $P_{bg} = G_{bg} p_{bg} A_{det} \Delta\lambda$. Because we have no prior knowledge of the distribution of background light, we are assuming that the background irradiance per unit filter bandwidth p_{bg} and the optical gain G_{bg} are constant, independent of receiver position and orientation. Background radiation will impinge on the detector at all angles of incidence, and thus G_{bg} can be found by performing a weighted integral of the gain G_ψ of Fig. 2-3 over all ψ. Without prior knowledge of the precise distribution of background light, however, G_{bg} can only be estimated. Inspection of Fig. 2-3 reveals that the value n^2 is a reasonable choice for G_{bg}, since it balances the low gains at low angles of incidence with the high gains at high angles of incidence. We assume $G_{bg} = n^2$ in the sequel.

Let SNR_{req} be the SNR required to achieve acceptable error-rate performance. Since the receiver can be located anywhere within the coverage area, we must design our system so that the SNR is greater than SNR_{req} at every possible receiver location. For each possible location, let ψ denote the angle between the *direction* of the transmitter (down) and the *position vector* of the receiver, as shown in Fig. 2-12. Observe that, for any ψ, the SNR is at its minimum when the path loss is greatest, or in other words, when the receiver is located at the boundary of the cell on the line labeled $\overline{AB}$ or the line labeled $\overline{BC}$ in Fig. 2-12. This leads to the following figure of merit

$$\text{Figure of Merit} = \min_{\overline{AB},\,\overline{BC}} SNR_{shot} \tag{2-19}$$

where the minimum is taken over all receiver locations on lines $\overline{AB}$ and $\overline{BC}$. We can guarantee a reliable link throughout the entire cell only if this figure of merit exceeds SNR_{req}. We will optimize the transmitter radiation pattern and optical filter by maximizing this figure of merit. (If the resulting maximum is less than SNR_{req}, then all is not lost, because the SNR can be increased in other ways such as by increasing the photodetector area; see chapter 3.) An equivalent figure of merit which isolates the parameters of interest is:

$$\text{Figure of Merit} = \min_{\overline{AB},\,\overline{BC}} \frac{P_{sig}}{\sqrt{\Delta\lambda}} \tag{2-20}$$

In the next four subsections we derive an equation for P_{sig}, optimize the antireflection coatings and transmitter radiation pattern assuming an arbitrary $\Delta\lambda$ and Θ, and then optimize $\Delta\lambda$ and Θ assuming an optimal radiation pattern.

2.4.2 Link Equation

To simplify analysis, we assume that the transmitter radiation pattern is axially symmetric, which would in fact be optimal only for circular rooms. To account for square rooms, we will circumscribe a circular room around the square room and proceed to analyze it. In this way, the two-dimensional problem of designing the system for operation at every position on the extremum of the two-dimension surface that comprises the cell boundary is reduced to a one-dimensional problem, since the axial symmetry obviates the second dimension. Extensions to rectangular rooms are straightforward, and in fact should be performed, especially for odd-shaped rooms. For our purposes, however, the extra dimension of such an analysis obscures the important results of our work, and so for simplicity we present results for the one-dimensional case only.

Define the transmitter radiation pattern $R(\psi)$ (W/sr) as the power per unit solid angle emitted from the source at an angle ψ from the orientation of the transmitter. A receiver with a bare detector of area A_{det} pointing straight up and located at an angle ψ and vertical distance h from the source will detect a total power of:

$$P_{sig,bare} = L_{REF,2}(\psi)R(\psi)d\Omega \tag{2-21}$$

where $d\Omega$ is the solid angle subtended by the detector, given by:

$$d\Omega = \cos^3\psi \, A_{det} / h^2 \tag{2-22}$$

and $L_{REF,2}(\psi)$ accounts for the reflection losses at the surface of the photodetector. With a hemispherical lens and optical filter placed on top of the detector, the received signal is increased by a factor $G_\psi L_\psi(\Delta\lambda,\Theta)L_{REF,1}(\psi)$, yielding:

$$P_{sig} = \frac{1}{h^2}R(\psi)\cos^3\psi \, A_{det}G_\psi L_\psi(\Delta\lambda,\Theta)L_{REF}(\psi), \tag{2-23}$$

where G_ψ is the gain of the lens as discussed in section 2.2.1 (see (2-1)), $L_{REF}(\psi) = L_{REF,1}(\psi)L_{REF,2}(\psi)$ accounts for the total loss due to reflections at all interfaces, $L_{REF,1}(\psi)$ accounts for reflections at the air-lens interface, and $L_\psi(\Delta\lambda,\Theta)$ accounts for the loss of the optical filter. In the next two subsections we derive expressions for $L_\psi(\Delta\lambda,\Theta)$ assuming a planar and hemispherical optical filter, respectively, and in section 2.4.3 we calculate $L_{REF}(\psi)$.

2.4.2.1 Planar Filter Loss

We first derive the total filter transmission $L_\psi(\Delta\lambda,\Theta)$ assuming a planar filter. (Note that the *total* filter transmission $L_\psi(\Delta\lambda,\Theta)$, which is averaged over all angles of incidence, is different from the raw filter transmission $T(\theta)$ of (2-16)). We refer to Fig. 2-11-a, which shows a collimated beam of uniform irradiance impinging on the lens. Consider a single ray from this beam, incident at an angle ψ from the detector normal; it will refract at the air-lens interface and will travel towards the bottom of the lens at a new angle θ. With proper index matching at the lens-filter interface, θ will also be the angle of incidence into the filter. The filter attenuates this ray according to (2-16). To determine the total loss due to the filter, we must sum (integrate) over all possible angles of incidence resulting from all possible rays in the original beam. We therefore introduce[1] $f_\psi^{(2)}(\theta)$ as the "power angle density" for light incident on the filter, defined such that $f_\psi^{(2)}(\theta)\Delta\theta$ is the fraction of the power destined to hit the detector that passes through the filter with an angle of incidence in the range $[\theta, \theta+\Delta\theta)$. We assume that the filter is at least as large as the detector, as shown in Fig. 2-11-a, so that light must pass through the filter to hit the photodetector. The density function satisfies:

$$\int_0^{\pi/2} f_\psi^{(2)}(\theta)\, d\theta = 1. \tag{2-24}$$

Using ray-tracing, as discussed in section 2.2, we computed the density functions for a 1-cm^2 detector with a 2-cm radius, 1.8-index lens. In Fig. 2-13 we show some sample results, where we plot $f_\psi^{(2)}(\theta)$ versus θ for $\psi \in \{0°,15°,30°,45°,60°,75°\}$.

The total transmission due to the planar filter at angle ψ is given by:

$$L_\psi(\Delta\lambda,\Theta) = \int_0^{\pi/2} f_\psi^{(2)}(\theta)\, T_\Theta^Q(\theta)\, d\theta. \tag{2-25}$$

2.4.2.2 Hemispherical Filter Loss

In Fig. 2-11-b we show a collimated beam impinging on a hemispherical filter. The angle of incidence with respect to the detector normal is ψ. A single ray in the beam is

1. We denote the curved surface of the lens (the lens input) as boundary 1, and the flat surface of the lens (the lens output) as boundary 2; hence the notation $L_{REF,1}$, $L_{REF,2}$, $f^{(1)}(\theta)$, and $f^{(2)}(\theta)$.

shown to make an angle θ_0 with the lens surface normal $\hat{n}$. Observe that light eventually detected by the photodetector will hit the hemispherical filter at nearly normal incidence.

Let $f_\psi^{(1)}(\theta)$ describe the angular distribution of light passing through the hemispherical filter, with $f_\psi^{(1)}(\theta)$ defined as the fraction of the power destined to hit the detector that passes through the filter with an angle of incidence in the range $[\theta, \theta+\Delta\theta)$.

As with $f_\psi^{(2)}(\theta)$, $f_\psi^{(1)}(\theta)$ is defined so that its integral is unity, and the total transmission due to the hemispherical filter is again given by (2-25), with $f_\psi^{(2)}$ replaced by $f_\psi^{(1)}$.

In Fig. 2-14 we show some sample density functions assuming the same parameters used in Fig. 2-13; a detector area of 1 cm^2 and a 2-cm radius lens with index n = 1.8. In contrast to the planar filter case, we see that the angles of incidence are confined to a narrow range of angles ranging from about 0° to 30°, regardless of ψ.

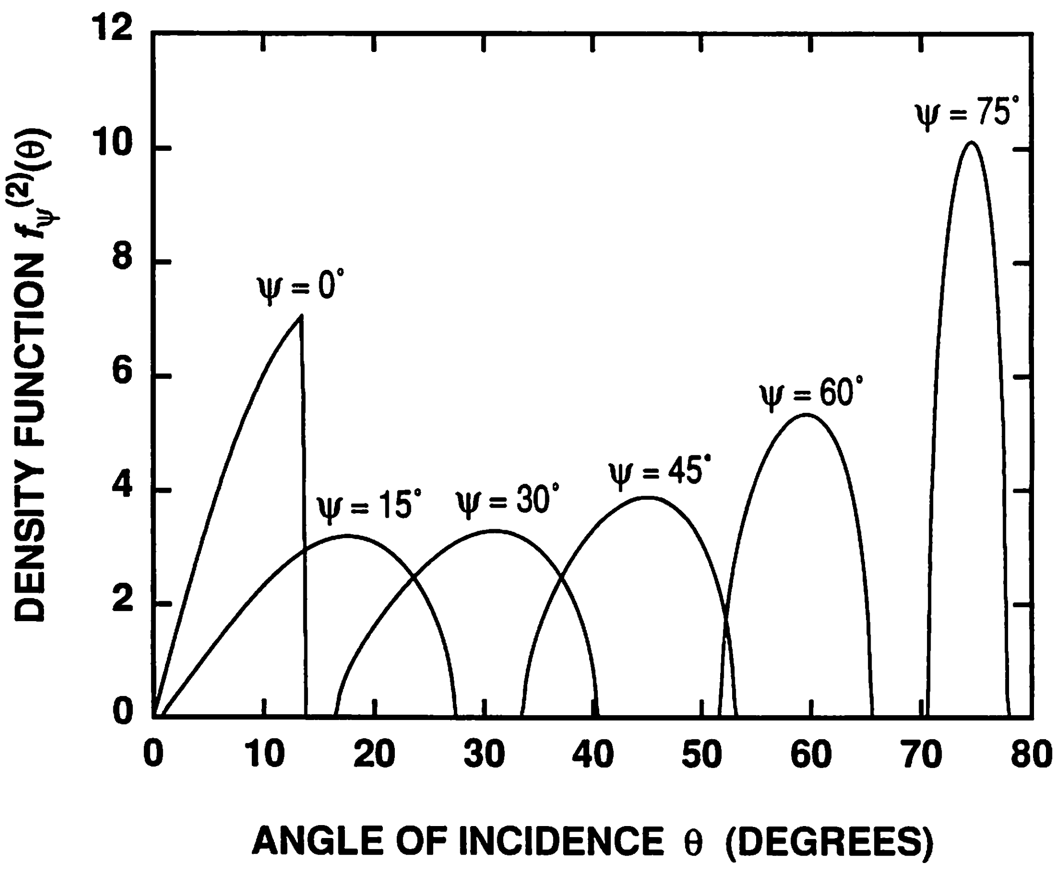


Fig. 2-13. Density functions for planar filter, assuming lens radius R = 2 cm and index n = 1.8.

Incidentally, it can be shown that:

$$f_0^{(1)}(\theta) = \begin{cases} k\sin(2\theta) & \text{for } \theta \in [0, \theta_{max}] \\ 0 & \text{elsewhere} \end{cases}, \qquad (2\text{-}26)$$

where $k = 1/\sin^2(\theta_{max}) = R^2/(r^2 G_0)$, and where R is the lens radius, r is the detector radius, and G_0 is the normal-incidence gain of the lens. This fact can be used to verify the accuracy of the $\psi = 0$ curve of Fig. 2-14.

In Fig. 2-15 we illustrate the superiority of the hemispherical filter in achieving a wide field of view by plotting the filter transmission L_ψ versus ψ using a polar plot. These curves were obtained by substituting (2-16) into (2-25), using the following parameters in (2-16): $\Theta = 47.2°$ for the planar filters and $\Theta = 15.2°$ for the hemispherical filters, $m = 2$,

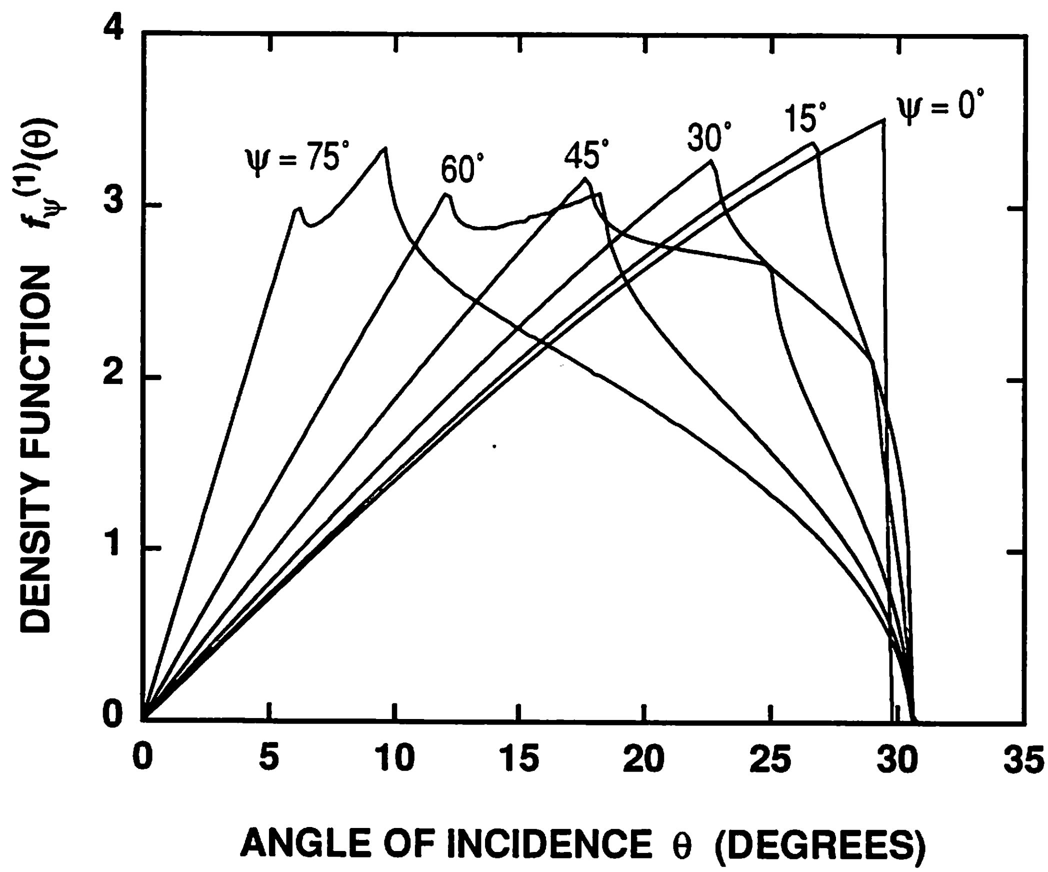


Fig. 2-14. Density functions for hemispherical filter, assuming lens radius $R = 2$ cm and index $n = 1.8$.

$n^*/n_1 = 2.276$, and $T_0 = 75\%$. Fig. 2-15-a assumes a wide bandwidth of $\Delta\lambda = 71.7$ nm while Fig. 2-15-b assumes a narrow bandwidth of $\Delta\lambda = 11.6$ nm. The planar filter is seen to be highly directional, even for the wide-bandwidth case. The hemispherical filter, on the other hand, is nearly omnidirectional, even for the narrow-bandwidth case. We can summarize the two advantages of the hemispherical filter over the planar filter as follows:

- For a given field of view, a hemispherical filter can have a narrower bandwidth, and thus reject more noise; this advantage is due to the narrow range of angles of incidence as shown in Fig. 2-14.
- Unlike the planar filter, which at some angles of incidence ψ can suffer significant loss, the hemispherical filter will always be near peak transmission; this advantage is due to similarity of the density curves shown in Fig. 2-14.

The extent of these advantages will become clear later when we make quantitative comparisons.

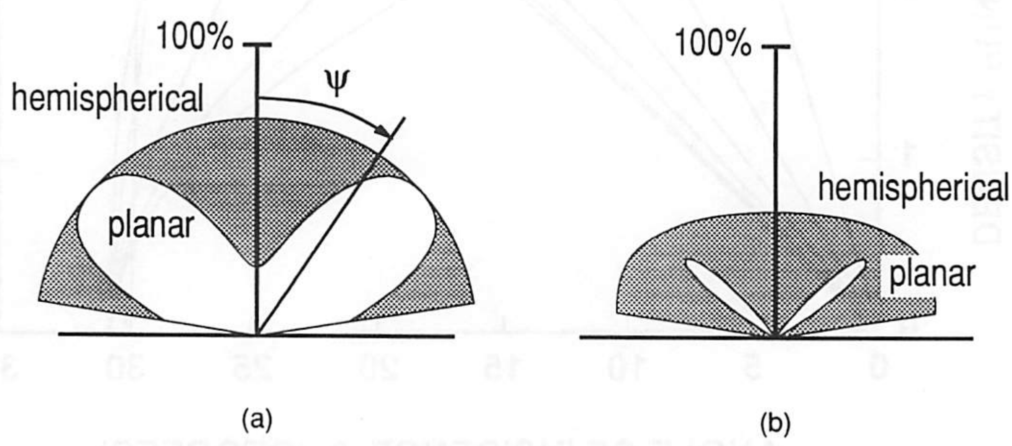


Fig. 2-15. Polar plots of total filter transmission $L_\psi(\Delta\lambda,\Theta)$ versus ψ for (a) $\Delta\lambda = 71.7$ nm, and (b) $\Delta\lambda = 11.6$ nm.

2.4.3 Optimization of Antireflection Coatings

In the next two subsection we derive expressions for the reflection transmissions $L_{REF,1}(\psi)$ and $L_{REF,2}(\psi)$, as defined in section 2.4.2.

2.4.3.1 Lens Coating

The *reflection transmission* factor $L_{REF,1}(\psi)$ as defined in section 2.4.2 accounts for reflection losses at the interface between the air and the lens. It comes into play only for the planar filter case, because, for the hemispherical filter case, reflections at the air-hemi-spherical-filter boundary are subsumed into the definition of the hemispherical filter transmission $L_{\psi}(\Delta\lambda,\Theta)$. In Fig. 2-11-a we show the proposed optics for the planar filter case. An antireflection coating AR_1 is shown on the lens surface. The purpose of this subsection is twofold; first, to optimize AR_1 so as to minimize the worst-case reflection loss, and second, to quantify the resulting transmission $L_{REF,1}(\psi)$.

An antireflection coating can be viewed as a special case of a thin-film optical filter. Therefore, the theory of thin-film optical filters presented in section 2.3.1 can be applied directly here. To wit, consider Fig. 2-6 with $K = 3$, $n_1 = 1$, $n_2 = n_{c,1}$, $d_2 = d_{c,1}$, and $n_3 = n$ (the lens index); this picture now describes the single-layer antireflection coating AR_1, where $n_{c,1}$ and $d_{c,1}$ are the refractive index and thickness of the coating material, respectively. Let $R_1(\theta;\, d_{c,1}, n_{c,1}, n)$ be the total reflection loss as specified by (2-11) for this case. Note that, since there are only $K = 3$ media, a closed-form expression for R_1 can be written out using only a couple of lines, but we omit it here for brevity [34]. In Fig. 2-16 we plot $(1 - R_1(\theta))$ versus θ under the assumptions listed in the caption.

The total reflection transmission $L_{REF,1}(\psi)$ must take into account the fact that, for a given angle ψ between transmitter orientation and receiver position, light that eventually hits the photodetector will pass through the filter at a distribution of angles described by the density functions of Fig. 2-14. The transmission is therefore given by:

$$L_{REF,1}(\psi) = 1 - \int_0^{\pi/2} f_\psi^{(1)}(\theta) R_1(\theta)\, d\theta. \tag{2-27}$$

The design of the AR coating AR_1 involves specifying two parameters: its thickness $d_{c,1}$ and its refractive index $n_{c,1}$. In light of our figure of merit (2-20), the optimum values will minimize the worst-case reflection loss, and will therefore satisfy:

$$(n_{c,1}, d_{c,1})_{opt} = \arg \min_{(n,d)} \left\{ \max_{\psi \in [0, \psi_{max}]} \int_0^{\pi/2} f_\psi^{(1)}(\theta) R_1(\theta)\, d\theta \right\}. \tag{2-28}$$

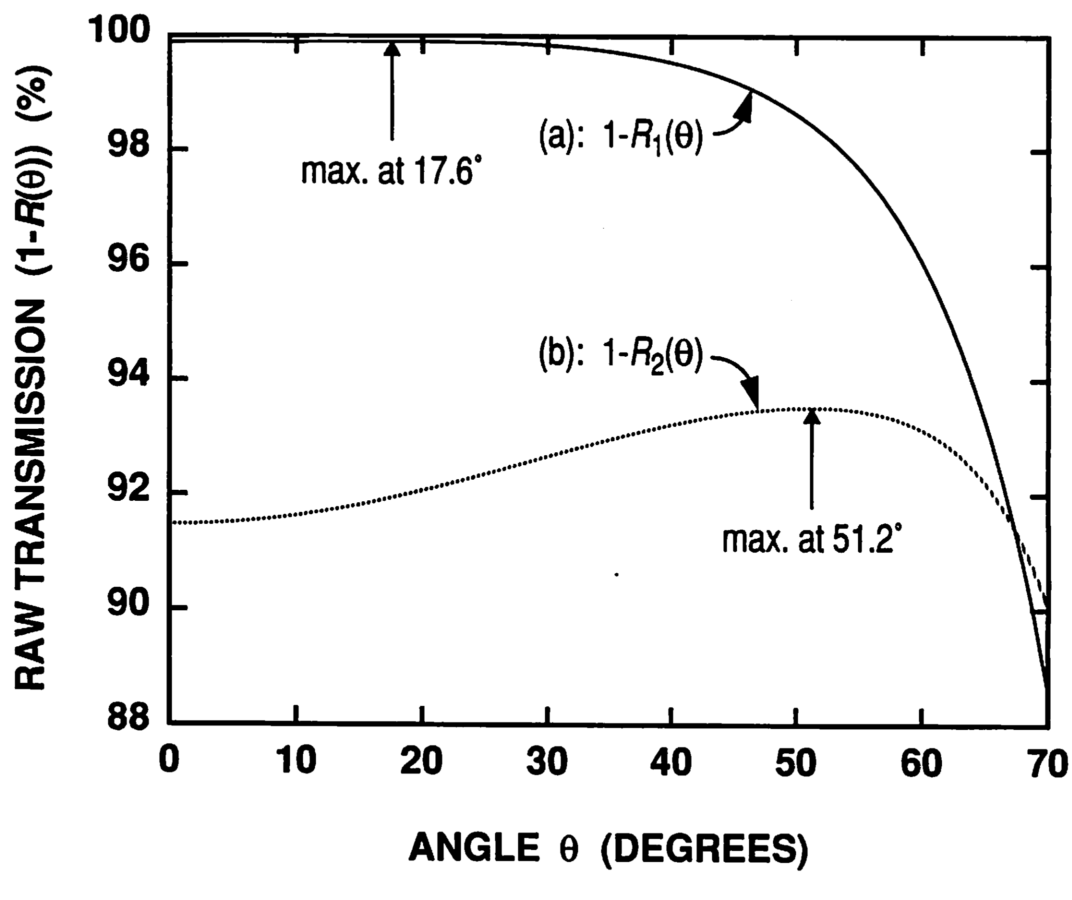


Fig. 2-16. Raw reflection transmissions for dielectric interfaces with optimal anti-reflection coatings: (a) $(1 - R_1(\theta))$ assuming $d_{c,1} = 1.0335$ QWOT, $n_{c,1} = 1.38$, and $n = 1.8$; (b) $(1 - R_2(\theta))$ assuming $n = 1.8$, $d_{c,2} = 1.4376$ QWOT, $n = 2.0$, and $n_d = 3.686$.

In practice, a continuum of values for $n_{c,1}$ is not available, but rather the designer must choose $n_{c,1}$ from a discrete set of practical values. The optimal coating index will be near $\sqrt{n}$, where n is the refractive index of the lens. For example, the common coating material magnesium fluoride (MgF_2) has a refractive index of $n_{c,1}$ = 1.38 [31], and thus makes a good choice for a lens with index n = 1.8.

Once $n_{c,1}$ is chosen, the optimal thickness can be found using a one-dimensional version of (2-28):

$$d_{c,1,opt} = \arg \min_{d_c} \left\{ \max_{\psi \in [0, \psi_{max}]} \int_0^{\pi/2} f_{\psi}^{(1)}(\theta) R_1(\theta) d\theta \right\}. \tag{2-29}$$

Using numerical techniques we found that, for $n_{c,1} = 1.38$, $n = 1.8$, $R = 2$ cm, and $\psi_{max} = 67°$, the optimal coating at the lens surface was $d_{c,1,opt}$ = 1.0335 QWOT. Here, QWOT is a unit representing one quarter-wave optical thickness, so that $d_{c,1}$ = 1 QWOT implies that $n_{c,1}d_{c,1} = \lambda_0/4$ for operating wavelength λ_0 [32]. In fact, $d_{c,1}$ = 1 QWOT is the coating that minimizes total reflection *at normal incidence* for light with wavelength λ_0. Increasing the thickness to $d_{c,1,opt}$ = 1.0335 QWOT shifts the angle of minimum reflection to about $\theta = 17.6°$. This is illustrated in curve (a) of Fig. 2-16, which achieves a maximum at $\theta = 17.6°$.

A summary of these and other numerical results is given in Table 2-1 near the end of the chapter (see page 52).

2.4.3.2 Detector Coating

The coating for the detector can be designed in a similar manner. First, consider the receiver optics for the planar filter case illustrated in Fig. 2-11-a. To minimize reflection loss between the lens and the planar filter, the intervening space should be filled with an index-matching gel. By choosing the index of the gel to equal the index n of the lens, reflections at the lens-gel interface can, in principle, be eliminated. As before, reflections at the input and output of the planar filter can be subsumed into the transmission characteristic of the filter itself and need not be considered separately. The only remaining issue is to minimize the reflections between the filter output and the photodetector input. Again, an index matching gel should be used between the filter and detector to minimize reflection loss. The optimal refractive index for this gel layer is not obvious; it depends on the output wave impedance of the optical filter, which is hard to predict a priori, and the input wave impedance of the photodetector, which will be AR coated. A reasonable choice seems to be to use the same gel below the filter as above. This has two benefits: first, it simplifies the packaging requirements, and second, less obviously at this point, it allows the same AR-coated detector to be used for both the hemispherical-filter case and the planar-filter case. Therefore, we assume that the index of both gel layers in Fig. 2-11-a is n, the index of the lens.

The reflection loss for the AR-coated detector can now be found using the theory of section 2.3.1, assuming the input medium has index n and the output medium has index $n_d = 3.686$, the refractive index of silicon at $\lambda_0 = 810$ nm [36]. Let $n_{c,2}$ and $d_{c,2}$ be the refractive index and thickness of the detector coating material, respectively and let $R_2(\theta;\, d_{c,2}, n, n_{c,2}, n_d)$ be the total reflection loss as specified by (2-11) for this case. Averaging over all possible angles of incidence yields the following expression for the total reflection transmission at the detector surface:

$$L_{REF,2}(\psi) = 1 - \int_0^{\pi/2} f_\psi^{(2)}(\theta)\, R_2(\theta)\, d\theta, \tag{2-30}$$

which is similar to (2-27) except that here we use the density functions $f_\psi^{(2)}$ of Fig. 2-13.

The AR coating of the detector can now be chosen to minimize the worst-case reflection loss. First, $n_{c,2}$ should be chosen near $\sqrt{nn_d}$. For a lens index of $n = 1.8$ and a detector index of $n_d = 3.686$, for example, $n_{c,2}$ should be chosen near 2.6. We will assume SiO with $n_{c,2} = 2.0$ in our numerical examples.

Once $n_{c,2}$ is chosen, the optimal thickness becomes:

$$d_{c,2,opt} = \arg \min_{d_c} \left\{ \max_{\psi \in [0, \psi_{max}]} \int_0^{\pi/2} f_\psi^{(2)}(\theta) R_2(\theta)\, d\theta \right\}. \tag{2-31}$$

With lens index $n = 1.8$, lens radius $R = 2$ cm, coating index $n_{c,2} = 2.0$, detector index $n_d = 3.686$, and maximum angle $\psi_{max} = 67°$, we used numerical techniques to find $d_{c,2,opt} = 1.4376$ QWOT (see Table 2-1). This corresponds to a reflectivity minimum at $\theta = 51.2°$. This is illustrated in curve (b) of Fig. 2-16, which achieves a maximum at $\theta = 51.2°$.

In Table 2-1 we present optimization results for 10 m × 10 m rooms as well as for 5 m × 5 m rooms. There, we see that the lens coating does not change for the larger room, whereas the detector coating does.

2.4.3.3 Combined Reflection Loss

For the planar filter case, the total reflection transmission at all interfaces, as first introduced in the link equation (2-23), can be written as:

$$L_{REF}(\psi) = L_{REF,1}(\psi) L_{REF,2}(\psi), \tag{2-32}$$

where $L_{REF,1}(\psi)$ and $L_{REF,2}(\psi)$ are given by (2-27) and (2-32), respectively. For the hemispherical filter case, on the other hand, the reflections at the air-lens interface are subsumed into the filter transmission characteristic, and therefore the total reflection transmission is simply $L_{REF}(\psi) = L_{REF,2}(\psi)$. In Fig. 2-17 we plot the individual and combined reflection transmissions when the AR coatings are optimized according to the parameters listed in the caption. Note that ψmax = 67° in this case, and that the optimization of (2-31) results in the interesting property that $L_{REF,2}(0) = L_{REF,2}(\psi_{max})$. The total loss ranges from about 7% to 9% for $\psi \in [0, \psi_{max}]$. This is considerably better performance than can be achieved without the index-matching gels and AR coatings. In fact, as discussed in section 2.2.4, without the gel there would be 100% loss due to total internal reflection for all angles greater than about 34°.

2.4.4 Optimization of Transmitter

We now optimize the transmitter radiation pattern $R(\psi)$. We assume $R(\psi)$ is zero for $\psi > \psi_{max}$, where $\psi_{max} \equiv \mathrm{atan}(d_{max}/h_{min})$ is the maximum angle for the cell defined in section 2.4.1; this truncation may be difficult to achieve in practice, but it is clearly optimal with respect to the figure of merit of (2-19). The following observation greatly simplifies the optimization of the transmitter radiation pattern.

> **Theorem 1.** The optimal radiation pattern $R(\psi)$, maximizing the figure of merit of (2-19), is that which makes the received signal power constant everywhere, independent of position, on the lines $\overline{AB}$ and $\overline{BC}$.

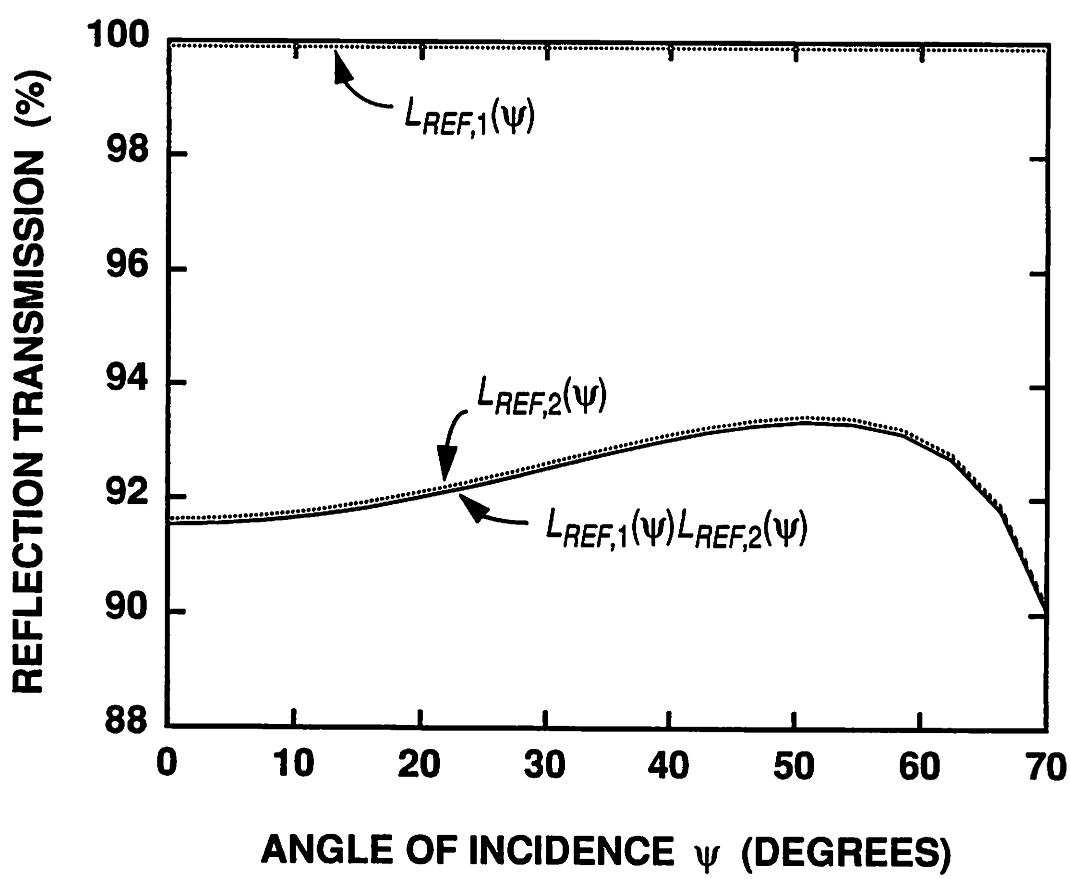


Fig. 2-17. Total reflection transmissions, assuming $d_{c,1}$ = 1.0335 QWOT, $n_{c,1}$ = 1.38, n = 1.8, $d_{c,2}$ = 1.4376 QWOT, $n_{c,2}$ = 2.0, and n_d = 3.686.

Proof. By contradiction. Let $R_0(\psi)$ make the received power equal to the same constant at every point on the lines $\overline{AB}$ and $\overline{BC}$, and suppose that:

$$R(\psi) = R_0(\psi) + \delta R(\psi) \tag{2-33}$$

yields a higher figure of merit as specified in (2-19), with $\delta R(\psi)$ non-zero.

We first show that there exists an angle ψ_- such that $\delta R(\psi_-) < 0$. Both $R_0(\psi)$ and $R(\psi)$ must satisfy the same power constraint:

$$P_T = 2\pi \int_0^{\psi_{max}} R_0(\psi) \sin(\psi)\, d\psi \tag{2-34}$$

and

$$P_T = 2\pi \int_0^{\psi_{max}} R(\psi) \sin(\psi)\, d\psi \tag{2-35}$$

$$= 2\pi \int_0^{\psi_{max}} R_0(\psi) \sin(\psi)\, d\psi + 2\pi \int_0^{\psi_{max}} \delta R(\psi) \sin(\psi)\, d\psi \tag{2-36}$$

$$= P_T + 2\pi \int_0^{\psi_{max}} \delta R(\psi) \sin(\psi)\, d\psi\ . \tag{2-37}$$

The last integral must therefore be zero. Since $\sin(\psi) > 0$ for all $\psi \in (0, \psi_{max})$, we thus conclude that there exists an angle $\psi_- \in (0, \psi_{max})$ such that $\delta R(\psi_-) < 0$.

Next, define P_{const} as the constant received power on lines $\overline{AB}$ and $\overline{BC}$ when the transmitter radiation pattern is $R_0(\psi)$; at any $\psi \in (0, \psi_{max})$ it is given by the link equation (2-23):

$$P_{const} = \frac{1}{h^2} R_0(\psi) \cos^3\psi\, A_{det} G_\psi L_\psi(\Delta\lambda, \Theta) L_{REF}(\psi). \tag{2-38}$$

Similarly, define $P(\psi)$ as the received power at angle ψ on lines $\overline{AB}$ and $\overline{BC}$ when the radiation pattern is $R(\psi)$ of (2-33):

$$P(\psi) = \frac{1}{h^2}\{R_0(\psi) + \delta R(\psi)\}\cos^3\psi\, A_{det}G_\psi L_\psi(\Delta\lambda,\Theta)L_{REF}(\psi) \tag{2-39}$$

$$= P_{const} + \frac{1}{h^2}\{\delta R(\psi)\}\cos^3\psi\, A_{det}G_\psi L_\psi(\Delta\lambda,\Theta)L_{REF}(\psi)\,. \tag{2-40}$$

In particular, at $\psi = \psi_-$ we have:

$$P(\psi_-) = P_{const} + \frac{1}{h^2}\{\delta R(\psi_-)\}\cos^3\psi_-\, A_{det}G_\psi L_{\psi\text{-}}(\Delta\lambda,\Theta)L_{REF}(\psi_-)$$

$$< P_{const} \tag{2-41}$$

where the last inequality follows because $\delta R(\psi_-) < 0$ and $\psi_- \in (0, \pi/2)$. Thus, we see that any deviation from $R_0(\psi)$ results in a reduced received signal power at some position ψ_-, which guarantees that $R_0(\psi)$ is optimal with respect to the figure of merit of (2-19). ♦

We can now derive the optimal $R(\psi)$ by setting the received signal power $P_{sig}(\psi)$ on lines $\overline{AB}$, $\overline{BC}$ equal to a constant $P_{const}(\Delta\lambda,\Theta)$, independent of ψ:

$$P_{sig}(\psi) = \frac{1}{h^2(\psi)}R(\psi)\cos^3\psi\, A_{det}G_\psi L_\psi(\Delta\lambda,\Theta)L_{REF}(\psi) = P_{const}, \tag{2-42}$$

where $h(\psi)$ is the vertical distance between the transmitter and the receiver located on lines $\overline{AB}$, $\overline{BC}$ at an angle ψ from the source normal:

$$h(\psi) = \begin{cases} h_{max} & \text{on } \overline{AB} \\ d_{max}/\tan\psi & \text{on } \overline{BC} \end{cases}. \tag{2-43}$$

Solving (2-42) for $R(\psi)$ yields:

$$R_{opt}(\psi) = \frac{h^2(\psi)\,P_{const}(\Delta\lambda,\Theta)}{A_{det}\cos^3\psi G_\psi L_\psi(\Delta\lambda,\Theta)\,L_{REF}(\psi)} \qquad \psi \in [0, \psi_{max}]. \tag{2-44}$$

Examples of $R_{opt}(\psi)$ are plotted in Fig. 2-18 and Fig. 2-19 in the next subsection.

Define P_T as the total power emitted from the transmitter, in which case $R_{opt}(\psi)$ must satisfy:

$$P_T = 2\pi \int_0^{\psi_{max}} R_{opt}(\psi) \sin(\psi)\, d\psi \,. \tag{2-45}$$

Substituting (2-44) into (2-45) and solving for $P_{const}(\Delta\lambda,\Theta)$ yields:

$$P_{const}(\Delta\lambda,\Theta) = \frac{A_{det} P_T}{2\pi \displaystyle\int_0^{\psi_{max}} \frac{h^2(\psi)\sin\psi}{\cos^3(\psi)\, G_\psi L_\psi(\Delta\lambda,\Theta)\, L_{REF}(\psi)}\, d\psi} \,. \tag{2-46}$$

In summary, for a given $\Delta\lambda$ and Θ, the optimal transmitter radiation pattern is given by (2-44). This radiation pattern causes the detected signal power to be the same at every point on lines $\overline{AB}$ and $\overline{BC}$, namely $P_{const}(\Delta\lambda,\Theta)$ as specified in (2-46). In the next section we optimize the optical filter parameters $\Delta\lambda$ and Θ.

2.4.5 Optimization of Filter

For a given $\Delta\lambda$ and Θ, when the radiation pattern is chosen optimally according to (2-44), the figure of merit of (2-20) reduces to:

$$\text{Figure of Merit} = \frac{P_{const}(\Delta\lambda,\Theta)}{\sqrt{\Delta\lambda}} \,. \tag{2-47}$$

Note that, since the detected signal power is identical everywhere on the cell extremum, the minimization over lines $\overline{AB}$ and $\overline{BC}$ is no longer necessary. The optimal filter parameters $\Delta\lambda_{opt}$ and Θ_{opt} thus satisfy:

$$(\Delta\lambda_{opt}, \Theta_{opt}) = \arg \max_{\Delta\lambda,\Theta} \frac{P_{const}(\Delta\lambda,\Theta)}{\sqrt{\Delta\lambda}} \tag{2-48}$$

and can be found numerically by performing, for example, a grid search.

In Table 2-1 we present sample optimization results assuming a square room with a 3-m ceiling and a peak transmitter power of 1 W. As illustrated in Fig. 2-12, we assume a maximum and minimum vertical distance of $h_{max} = 2.4$ m and $h_{min} = 1.5$ m, respectively. The maximum angles $\psi_{max} = \tan^{-1}(d_{max}/h_{min})$ for the 5 m × 5 m and 10 m × 10 m rooms are 67° and 78°, respectively. The first row of the table is $d_{c,1,opt}$, the optimal thickness of the lens antireflection coating AR_1. The second row is $d_{c,2,opt}$, the optimal thickness of the photodetector antireflection coating AR_2. The third row of the table is $\Delta\lambda_{opt}$, the optimal filter bandwidth, and the fourth is Θ_{opt}, the optimal filter orientation. The fifth row is $\lambda_{normal,opt}$, the optimal normal-incidence wavelength; it is related to Θ_{opt} by (2-15). The sixth row is $p_{sig} = P_{const}(\Delta\lambda,\Theta)/A_{det}$, the minimum received signal *irradiance* anywhere in the cell. The seventh row is SNR_{min}, which we define as the shot-noise-limited SNR from (2-17) and (2-18) with a worst-case received power of $P_{sig} = A_{det}p_{sig}$, assuming the following: a photodetector area of $A_{det} = 1$ cm^2, a photodetector responsivity of 0.53 A/W, a background irradiance per unit bandwidth of $p_{bg} = 5.8$ μW/(cm^2·nm), a noise bandwidth factor of $I_2 = 0.562$, and a bit rate of $B = 100$ MHz. The last row of the table is $A_{det,SNL}$,

TABLE 2-1: Sample Optimization Results.

	5 m × 5 m		10 m × 10 m			
parameter	planar	hemisp.	planar	hemisp.	units	assumptions
$d_{c,1}$	1.0335	NA	1.0335	NA	QWOT	$n_{c,1} = 1.38$
$d_{c,2}$	1.4376	1.4376	1.7711	1.7711	QWOT	$n_{c,2} = 2.0$
$\Delta\lambda_{opt}$	71.7	11.6	73.5	8.3	nm	Butter. $m = 2$
Θ_{opt}	47.2°	15.2°	57.6°	12.2°	degrees	$n^*/n_1 = 2.276$,
$\lambda_{normal,opt}$	855.7	815.4	872.3	813.5	nm	$\lambda_0 = 810$ nm
p_{sig}	2.6	2.3	0.66	0.52	μW/cm^2	
SNR_{min}	21.7	28.5	9.7	17.0	dB	$A_{det} = 1$cm^2
$A_{det,SNL}$	0.97	0.20	15.4	2.9	cm^2	for SNR = 144

which we define as the detector area required to achieve an SNR of $SNR_{req} = 21.6$ dB, assuming shot-noise-limited operation. ($SNR_{req} = 21.6$ dB was chosen because it is the required peak SNR for a baseband on-off-keyed system with additive Gaussian noise to achieve a bit-error rate of 10^{-9}). When we consider other noises besides shot noise in chapter 3 we will see that the required detector area will be somewhat larger than $A_{det,SNL}$.

The first column in Table 2-1 assumes a 5 m × 5 m room and a planar filter, where we see that $\Delta\lambda_{opt} = 71.7$ nm, and the optimal orientation angle is $\Theta_{opt} = 47.2°$. The optimal transmitter radiation pattern in this case, as specified by (2-44), is shown in Fig. 2-18 using a polar plot. The worst-case signal irradiance is $p_{sig} = 2.6\ \mu\text{W/cm}^2$, corresponding to an SNR of $SNR_{min} = 21.7$ dB. An area of 0.97 cm^2 is required to achieve an SNR of $SNR_{req} = 21.6$ dB.

The second column is also for a 5 m × 5 m room, but this time assuming a hemispherical filter. The difference in results is striking. The optimal bandwidth in this case is

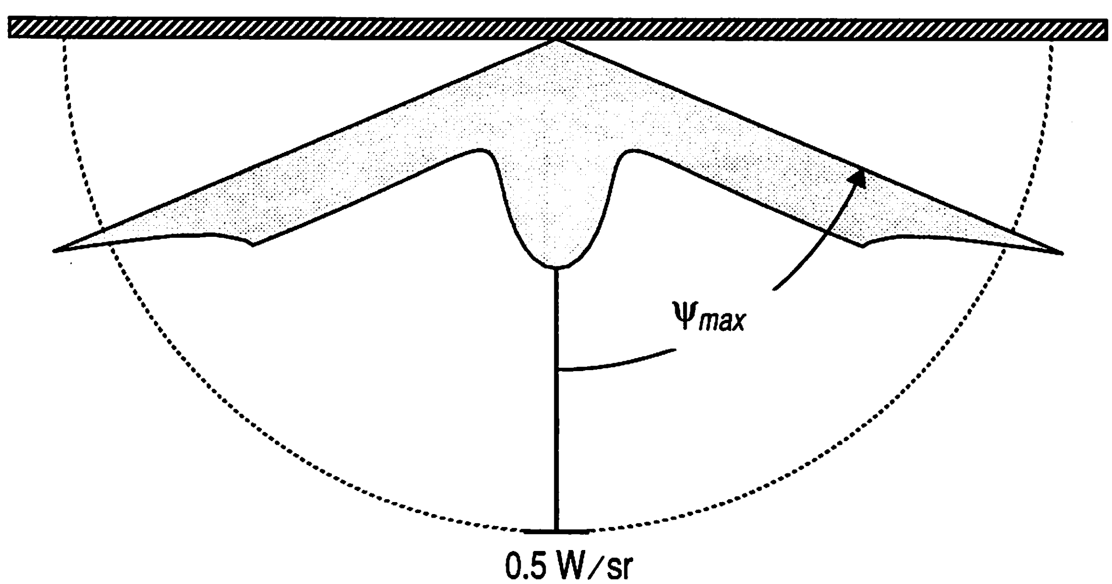


Fig. 2-18. Optimal transmitter radiation pattern for 5 m × 5 m room with planar filter ($\Delta\lambda = 71.7$ nm, $\Theta = 47.2°$).

$\Delta\lambda_{opt} = 11.6$ nm; the hemispherical filter thus rejects $71.7/11.6 = 7.9$ dB more noise power than does the optimal planar filter. Furthermore, the price paid for this superior noise rejection in terms of filter loss is not great; the minimum signal irradiance is $p_{sig} = 2.3\ \mu W/cm^2$, only 0.53 dB optical (1.06 dB electrical) less than that for the planar filter. The net result is that the hemispherical filter achieves a 6.8 dB improvement in SNR when compared to the optimal planar filter. The optimal radiation pattern for the hemispherical filter case is shown in Fig. 2-19.

The results in the first two columns of Table 2-1 are encouraging because they show that practical detector areas (0.97 cm^2 for a planar filter and 0.2 cm^2 for a hemispherical filter) are sufficient for a 5 m × 5 m room. This leads us to consider larger rooms. We repeated the optimization procedure for a 10 m × 10 m room. The only difference in this case is that d_{max} is now $5\sqrt{2}$ m ≈ 7.07 m and ψ_{max} = 78°. The results are shown in the third and fourth columns of Table 2-1, where we see that the planar filter bandwidth

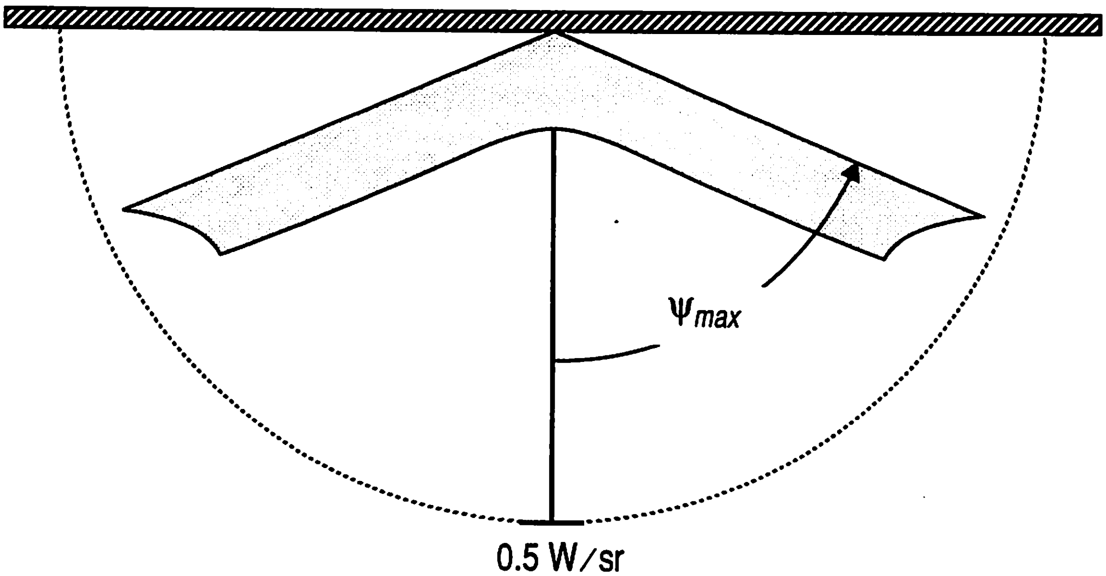


Fig. 2-19. Optimal transmitter radiation pattern for 5 m × 5 m room with hemispherical filter ($\Delta\lambda$ = 11.6 nm, Θ = 15.2°).

increased from 71.7 nm to 73.5 nm for the larger room, whereas the hemispherical filter bandwidth decreased from 11.6 nm to 8.3 nm. This discrepancy is due to the increased field-of-view requirement for the larger room: to accommodate the worst-case field-of-view of $\psi_{max} = \operatorname{atan}(d_{max}/h_{min})$, the planar filter has to decrease its bandwidth. The larger ψ_{max} has less impact on the hemispherical filter, because the density functions $f_{\psi}^{(1)}(\theta)$ tend to concentrate more of their mass near normal incidence as ψ gets large; see Fig. 2-14. In contrast, the planar filter density functions $f_{\psi}^{(1)}(\theta)$ tend to diverge as ψ gets large; see Fig. 2-13.

The SNR improvement provided by the hemispherical filter increases from 6.8 dB for the 5 m $\times$ 5 m room to 7.3 dB for the 10 m $\times$ 10 m room. Unfortunately, however, the required detector areas for the large room are excessive: 15.4 cm^2 for the planar filter case and 2.9 cm^2 for the hemispherical filter case. We thus conclude that, with a 1-W transmitter, a 10 m $\times$ 10 m room is grossly impractical for a planar-filter system, while just barely out of reach for a hemispherical-filter system.

2.5 SUMMARY AND CONCLUSIONS

This chapter has presented a procedure, summarized in section 2.1, for designing the optical components in a wireless link using the non-directed LOS configuration (see Fig. 1-1-e). We have quantified the optical gain provided by a hemispherical lens as a function of lens radius and angle of incidence. We have defined two families of density functions $\{f_{\psi}^{(1)}(\theta)\}$ and $\{f_{\psi}^{(2)}(\theta)\}$, parameterized by $\psi \in [0, \psi_{max}]$, that characterize the angular distribution of light as it enters the lens through the curved surface ($f_{\psi}^{(1)}$) and as it exits the lens through the planar surface ($f_{\psi}^{(2)}$). We showed how these density functions could be calculated using a ray-tracing algorithm. We presented a five-parameter model for approximating the performance of thin-film optical bandpass filters. Using these results, we showed how the antireflection coatings, transmitter radiation pattern, and optical filter could be optimized. The optimal transmitter radiation pattern is that which makes the detected power equal at every position on the extremum of the cell.

The conclusions to draw from this chapter are obvious. First, a hemispherical lens is advantageous and should be used, because it provides a gain of approximately n^2 at all angles of incidence, where n is its refractive index. Likewise, a hemispherical thin-film filter should also be used, since, unlike the planar filter, it can provide both a narrow bandwidth and a wide field-of-view. Finally, to maximize the worst-case SNR within the cell, the optical gain of the transmitter and receiver should be designed jointly, not separately.

CHAPTER 3

RECEIVER DESIGN

In this chapter we examine the problem of designing the receiver front end, which consists of one or more photodetectors followed by a low-noise, wide-band preamplifier. The design of the preamplifier is challenging due to the high capacitance of large-area photodiodes and a potentially intense background light. The error performance of the entire system depends to a large degree on the ability of the preamplifier to amplify the detected signal over the desired bandwidth without adding undue noise. We show that it is always possible to make the preamplifier noise negligible in comparison to the background light shot noise. We show that, in practical cases, a PIN photodiode is preferable to an avalanche photodiode. We propose a transimpedance amplifier using a current-feedback pair, and present a detailed bandwidth and noise analysis. We present design procedures for two typical design scenarios. Numerical examples show that, using wide-gate FETs, low-noise operation over bandwidths near 100 MHz can be achieved for moderate detector areas near 1 cm^2.

3.1 INTRODUCTION

There are a number of similarities between preamplifier design for wireless infrared links and preamplifier design for fiber-optic links, and thus much can be learned from the extensive body of literature on fiber-optic receiver design [37]—[44]. There are two fundamental differences, however, which deserve special attention. First, a wireless infrared receiver must contend with an enormous amount of background radiation from sun, incandescent, and fluorescent lights. No such background light exists within the confines of an optical fiber. The in-band background power is typically 25 dB greater than the received signal power. If not properly accounted for, the shot noise from this background radiation may overwhelm the received signal. The background light may also saturate an amplifier that is not a.c. coupled.

Second, unlike the highly concentrated beam of light that emanates from an optical fiber, the received optical signal in a non-directed wireless system will be spread over a wide region, and so a large-area photodetector will be required to collect sufficient signal energy. The high capacitance of large-area photodetectors is the primary impediment to achieving a wide bandwidth.

For these reasons, the receiver design strategies of fiber-optic receivers cannot be adopted to nondirected wireless receivers. Therefore, this chapter presents a careful evaluation of all impediments and presents new strategies for low-noise wideband preamplifier design.

An optical receiver front end consists of two components: a photodiode and a wideband preamplifier. We will show in section 3.5.2 that, for our application, a PIN photodiode is preferable to an avalanche photodiode, and so we assume a PIN photodiode throughout this chapter.

Preamplifiers for optical receivers can be classified in three categories: low-impedance, high-impedance, and transimpedance. A simplified diagram for a low-impedance and high-impedance front end is shown in Fig. 3-1-a. Current from a reverse-biased photodiode produces a voltage across a load resistor R_F, and this voltage is fed into wideband amplifier. For a low-impedance receiver, R_F is chosen small (typically 50 Ω), so that the receiver bandwidth $1/(2\pi R_F C_T)$ is sufficient for the signal bandwidth. Here, C_T represents the total input capacitance, including the photodetector capacitance and the input capacitance of the following amplifier. The drawback of this approach is its reduced sensitivity; the thermal noise associated with R_F can be quite large for small R_F. A high-impedance receiver uses the same configuration but with R_F large, thus diminishing the effects of its thermal noise. However, the receiver bandwidth $1/(2\pi R_F C_T)$ is then usually smaller than the signal bandwidth, requiring an equalization stage immediately following the preamplifier. The equalizer cancels the pole at $1/(2\pi R_F C_T)$ with a zero. This can be tricky, because the load resistance R_F and input capacitance C_T will change with age and temperature. Also, the equalizer reduces the overall dynamic range of the receiver [37][41][42]. The third category of receiver, the transimpedance front end, solves these problems by using a large feedback resistance R_F and an inverting amplifier as shown in Fig. 3-1-b. This boosts the bandwidth with respect to a high-impedance amplifier roughly by a factor of the amplifier gain [41][42], without thermal noise and dynamic range problems.

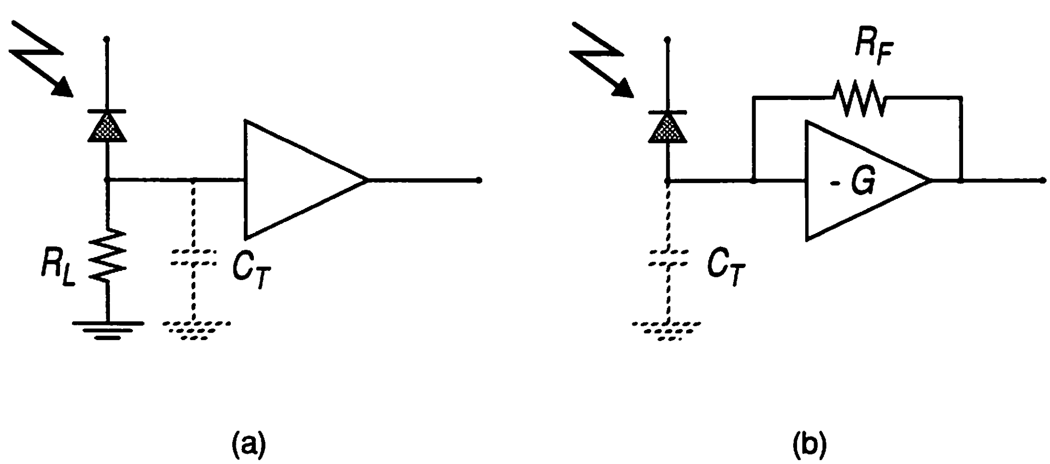


Fig. 3-1. Preamplifier types: (a) high-impedance and low-impedance configuration; (b) transimpedance configuration.

Personick showed that, of the three front-end types, the high-impedance front end provides the highest sensitivity for fiber-optic systems [38]. In our application, however, the dominant source of noise is due to background radiation, not thermal and circuit noise. This makes the sensitivity of the transimpedance front end identical to that of the high-impedance front end. Therefore, because of it larger dynamic range, the transimpedance front end is preferable in our case. We will therefore concentrate on a transimpedance preamplifier in this chapter.

3.2 ANALYSIS OF CURRENT-FEEDBACK PAIR

In Fig. 3-2-a we show a simplified schematic diagram for the proposed transimpedance amplifier. It is a current-feedback pair, a popular choice for wideband optical preamplifiers [37][43][45][46][47]. The current from the reverse-biased PIN photodiode is fed into the gate of a FET in a common-source configuration. The second stage is a source-follower. Current from the second stage is fed back to the first via a feedback resistor R_F, and for this reason this amplifier is referred to as a current-feedback pair. The output of the second stage is passed through a buffer stage to the equalization and detection portion of the receiver. We assume that this buffer amplifier has infinite input impedance, no noise, and infinite bandwidth, so that the small-signal current i_o, as labeled in Fig. 3-2-a, can be viewed as the preamplifier output. This is a valid assumption in practice, because the bandwidth limitations of the photodetector and the noise of the background radiation are generally dominant. The small-signal output voltage v_o, which is the input to the third-stage buffer amplifier, is proportional to i_o:

$$v_o = (R_S \parallel R_F) i_o \, , \qquad (3\text{-}1)$$

where $a \parallel b \equiv ab/(a+b)$.

In Fig. 3-2-b we show a simplified equivalent small-signal model, which was arrived at using standard feedback theory: the feedback from the second stage to the first stage was removed and replaced by an equivalent current source, taking into account the loading effects of the feedback circuit on the input and output stages [45]. The first current source provides $i_p - i_F$, the photodetector signal current minus the feedback current, plus a white Gaussian noise with (two-sided) power spectral density $N_{0,shot} + N_{0,res}$, where $N_{0,shot}$ is due to the background-light shot noise, given by:

$$N_{0,shot} = qrP_{bg} \tag{3-2}$$

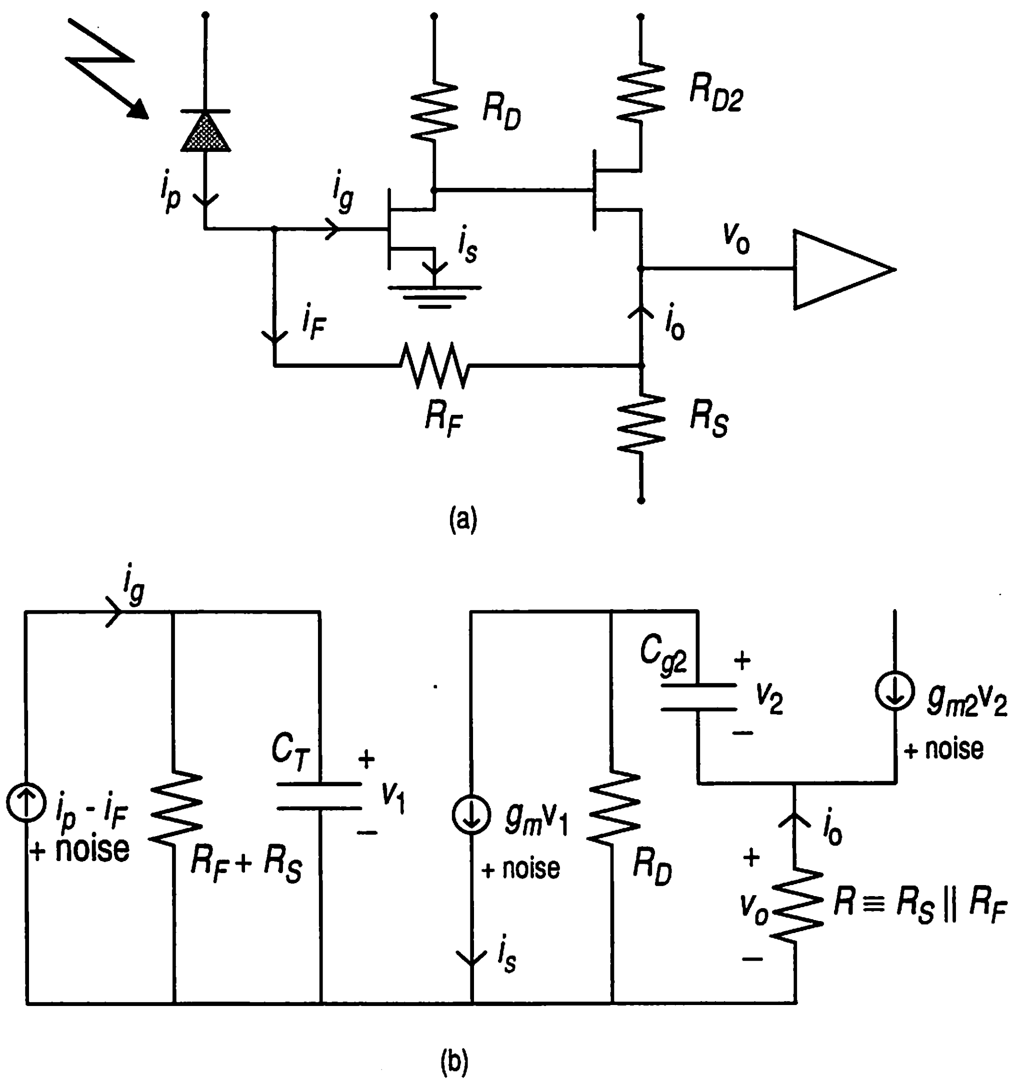


Fig. 3-2. Current-feedback pair transimpedance amplifier:
(a) simplified schematic diagram; (b) small-signal model.

where q is the charge of an electron, r is the photodiode responsivity, and P_{bg} is the total power of the detected background light, and $N_{0,res}$ is due to the thermal noise in the feedback circuit:

$$N_{0,res} = \frac{2kT}{R_F + R_S} \qquad (3\text{-}3)$$

where k is Boltzmann's constant and T is temperature. Recall that P_{bg} is related to p_{bg} of (2-18) by:

$$P_{bg} = G_{bg} p_{bg} A_{det} \Delta\lambda \, . \qquad (3\text{-}4)$$

The first-stage FET has gate capacitance C_g and transconductance g_m, while the second-stage FET has gate capacitance C_{g2} and transconductance g_{m2}. The total input capacitance is denoted C_T, and is given by:

$$C_T = C_{det} + C_{stray} + C_g \qquad (3\text{-}5)$$

where C_{det} is the detector capacitance and C_{stray} is stray capacitance due to detector mounting and other parasitics.

The noise in the second current source of Fig. 3-2-b has power spectral density $N_{0,FET} + N_{0,D}$, where $N_{0,FET}$ is due to the thermal noise in the channel of the first FET, given by:

$$N_{0,FET} = 2kT\Gamma g_m \, , \qquad (3\text{-}6)$$

where Γ is the channel noise factor of the first FET, and $N_{0,D}$ is due to the thermal noise in the drain resistor R_D:

$$N_{0,D} = \frac{2kT}{R_D} \, . \qquad (3\text{-}7)$$

The noise in the third current source of Fig. 3-2-b has power spectral density $N_{0,FET2} + N_{0,D2}$, where $N_{0,FET2}$ is due to the thermal noise in the channel of the second FET, given by:

$$N_{0,FET2} = 2kT\Gamma_2 g_{m2}, \tag{3-8}$$

where Γ_2 is the channel noise factor of the second FET, and where $N_{0,D2}$ is due to thermal noise in the second-stage drain resistor R_{D2}:

$$N_{0,D2} = \frac{2kT}{R_{D2}}. \tag{3-9}$$

The small-signal circuit of Fig. 3-2-b can be redrawn in a traditional block diagram form, as shown in Fig. 3-3-a. Define $G_1(s)$ as the transfer function between the small-signal currents i_g and i_S, as labeled in Fig. 3-2; it is given by:

$$G_1(s) = \frac{G_{10}}{1 + s/p_1}, \tag{3-10}$$

where G_{10} is the low-frequency current gain of the first stage:

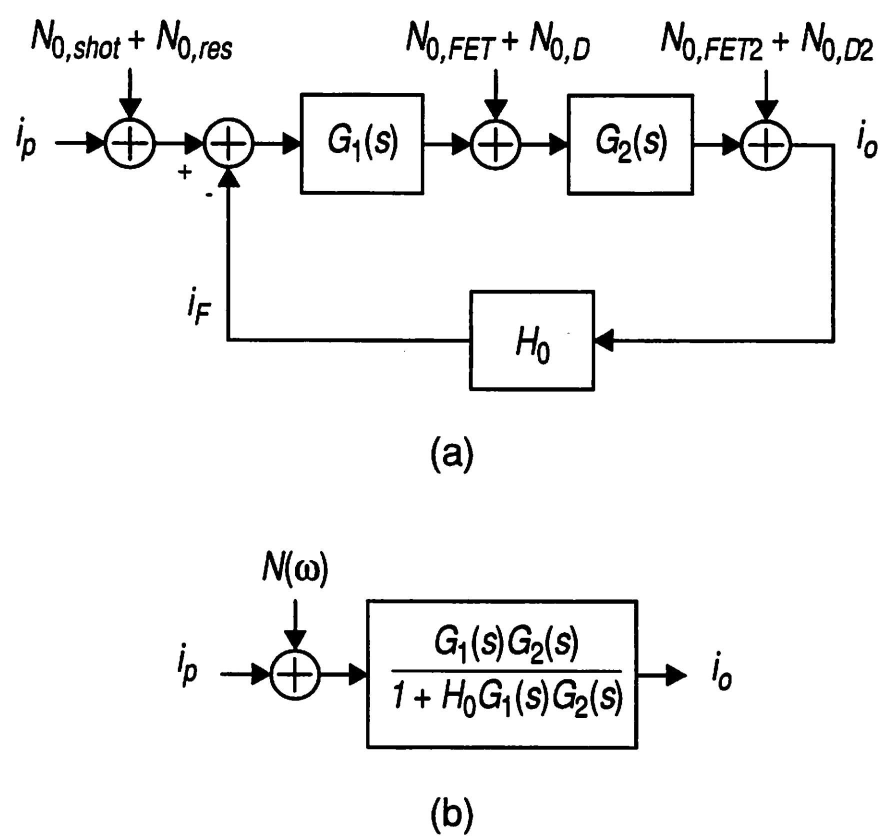


Fig. 3-3. Block diagram of preamplifier.

$$G_{10} = g_m(R_F + R_S), \tag{3-11}$$

and p_1 is the dominant pole of the input stage:

$$p_1 = \frac{1}{(R_F + R_S)\,C_T}. \tag{3-12}$$

Similarly, define $G_2(s)$ as the transfer function between i_S and i_o, given by:

$$G_2(s) = G_{20}\frac{1 + s/z_2}{1 + s/p_2} \tag{3-13}$$

where G_{20} is the low-frequency current-gain of the second stage:

$$G_{20} = \frac{g_{m2}R_D}{1 + g_{m2}R}\ , \tag{3-14}$$

where $R \equiv R_S \parallel R_F$, and:

$$z_2 = g_{m2}/C_{g2} \tag{3-15}$$

$$p_2 = \frac{1 + g_{m2}R}{(R_D + R)\,C_{g2}}\ . \tag{3-16}$$

The feedback gain H_0 is that of a current divider:

$$H_0 = \frac{R_S}{R_S + R_F}\ . \tag{3-17}$$

The closed-loop system of Fig. 3-3-a is equivalent to that of Fig. 3-3-b, where the power spectral density $N(\omega)$ of the input-referred noise is given by:

$$N(\omega) = N_{0,shot} + N_{0,res} + \frac{N_{0,FET} + N_{0,D}}{|G_1(s)|^2} + \frac{N_{0,FET2} + N_{0,D2}}{|G_1(s)\,G_2(s)|^2}$$

$$\approx N_{0,shot} + N_{0,res} + \frac{N_{0,FET} + N_{0,FET2}/G_{20}^2 + N_{0,D} + N_{0,D2}/G_{20}^2}{|G_1(s)|^2}\ , \tag{3-18}$$

the second approximation being valid when p_2 and z_2 exceed the system bandwidth, a desirable condition that will be met in practice. The power spectral density can be further simplified when $G_{20}^2 \gg 1$, in which case the second-stage FET channel noise and drain resistor thermal noise can be neglected. The equivalent noise power spectral density can then be written as:

$$N(\omega) = qrP_{bg} + \frac{2kT}{R_F + R_S}\left(1 + \frac{\Gamma_{eq}}{g_m(R_F + R_S)}\right) + \frac{2kT\Gamma_{eq}}{g_m}\,\omega^2 C_T{}^2 \tag{3-19}$$

where $\Gamma_{eq} = \Gamma + 1/(g_m R_D)$. Unlike the first two terms, which are constant, the third term is quadratic in frequency, and for this reason is often referred to as f^2 noise. The total variance of the input-referred noise, accounting for pulse shaping after the preamplifier, is:

$$\sigma^2 = \underbrace{I_2 B 2qrP_{bg}}_{\sigma^2_{shot}} + \underbrace{I_2 B\left[\frac{4kT}{R_F + R_S}\left(1 + \frac{\Gamma_{eq}}{g_m(R_F + R_S)}\right)\right]}_{\sigma^2_{res}} + \underbrace{I_3 B^3\left[\frac{4kT\Gamma_{eq}}{g_m}(2\pi C_T)^2\right]}_{\sigma^2_{FET}}, \tag{3-20}$$

where B is the bit rate and I_2 and I_3 are noise bandwidth factors; they are functions of the transmitter pulse shape and equalized pulse shape only, and are independent of bit rate [38]. For example, $I_2 = 0.562$ and $I_3 = 0.0868$ for a rectangular transmitter pulse shape and a full raised-cosine equalized pulse shape. To account for $1/f$ noise in the FET channel, the I_3B^3 factor in (3-20) should be replaced by $(I_3B^3 + I_f f_c B^2)$, where I_f is another noise bandwidth factor and f_c is the $1/f$-noise corner frequency [41][42].

The expression for noise power given in (3-20) is similar to that presented elsewhere in the fiber-optic literature [37][41], with the primary difference being the absence of shot-noise terms due to gate leakage current and dark current, which in our application are negligible compared to the background shot noise. In addition, although the expressions look similar, the total capacitance C_T will much larger in our case than for a fiber-optic receiver.

3.3 CHOOSING THE RIGHT TRANSISTOR

The FET transconductance g_m is related to the gate capacitance C_g through the short-circuit common-source gain-bandwidth product $\omega_T = 2\pi f_T$:

$$\omega_T = \frac{g_m}{C_g} \ . \tag{3-21}$$

It is inversely proportional to gate length, with the constant of proportionality determined by the transistor technology (such as GaAs HEMT, GaAs MESFET, Si MOSFET, or Si JFET). Both the transconductance and the gate capacitance are proportional to the gate width, so ω_T can be viewed as the ratio of the two constants of proportionality. The two dominant characteristics of the FET for our purposes are ω_T and g_m. For a given technology, specifying ω_T and g_m is equivalent to specifying the gate length and width. Note that the ratio of the gate width to gate length cannot be made arbitrarily large, which precludes the possibility of a short-length (high ω_T) wide-gate (high g_m) device.

The first-stage FET should be chosen to minimize σ^2_{FET}, defined as the third term in (3-20). Using (3-21), this variance can be rewritten as:

$$\sigma^2_{FET} = 16\pi^2 (I_3 B^3 + I_f f_c B^2) kT\Gamma_{eq} \frac{\left(C_{det} + C_{stray} + \dfrac{g_m}{\omega_T}\right)^2}{g_m} \ . \tag{3-22}$$

It is easily shown that, given ω_T, the optimal g_m, minimizing σ^2_{FET}, is approximately:

$$g_{m,opt} \approx \omega_T (C_{det} + C_{stray}). \tag{3-23}$$

In other words, the gate width should be chosen so that the gate capacitance C_g is equal to $C_{det} + C_{stray}$. (The true optimal value will be slightly larger, due to the dependence of Γ_{eq} on g_m, but this effect is negligible.)

To a first order approximation, the power consumed by the FET is proportional to the gate width and hence proportional to g_m. It is likely, therefore, that the optimal transconductance would lead to excessive power consumption, in which case a transconductance smaller that $g_{m,opt}$ should be used. In this case, the transconductance can be thought of as fixed at some maximum value g_m, and the designer must choose the device ω_T. Transistors with high values of ω_T are expensive, and so ω_T should be kept small. On the other hand, C_g is inversely proportional to ω_T, so ω_T cannot be too small. The best value for ω_T should be chosen to balance these two tradeoffs; a reasonable value is that which makes C_g about ten times less than $C_{det} + C_{stray}$: $\omega_{T,opt} = 10 \times g_m/(C_{det} + C_{stray})$.

The ω_T of the second-stage FET does not affect noise, and can thus be chosen based on bandwidth considerations only. Therefore, ω_{T2} need not be chosen any larger than the value which makes the pole p_2 and zero z_2 of $G_2(s)$ negligible.

3.4 DESIGN PROCEDURES

The purpose of this section is two-fold. First, in subsection 3.4.1, we identify a few heuristic principles for receiver design. Second, in the remaining subsections, we present specific design procedures for typical design scenarios.

3.4.1 Basic Philosophy

The two goals of the design are wide bandwidth and low noise. We first address bandwidth issues, then noise issues.

3.4.1.1 Bandwidth

The closed-loop transfer function for the preamplifier is:

$$A(s) = \frac{G_1(s)\,G_2(s)}{1 + H_0 G_1(s)\,G_2(s)}. \tag{3-24}$$

We will often assume that the zero z_2 and pole p_2 of $G_2(s)$ are negligible, in which case the closed-loop transfer function reduces to:

$$A(s) \approx \frac{G_0}{s/p_1 + 1 + G_0 H_0}, \tag{3-25}$$

where $G_0 \equiv G_{10}G_{20}$ is the open-loop low-frequency gain. In this case, the closed-loop low-frequency current gain is:

$$\text{gain} = \frac{G_0}{1 + G_0 H_0} \tag{3-26}$$

and the closed-loop bandwidth (rad/s) is:

$$\text{bandwidth} = (1 + G_0 H_0) p_1. \tag{3-27}$$

The "gain-bandwidth product" is therefore fixed at $G_0 p_1$ for all feedback gains H_0:

$$\text{gain-bandwidth product} = G_0 p_1 = \frac{g_m}{C_T} \cdot \frac{g_{m2} R_D}{1 + g_{m2} R_F R_S / (R_F + R_S)} \ . \tag{3-28}$$

(Note that the factor g_m / C_T can be viewed as an effective ω_T for the first FET.) This illustrates that, under the single-pole assumption, extremely high bandwidths can be obtained by increasing the feedback gain H_0, at the expense of a decreased low-frequency gain. A more desirable way of achieving higher bandwidth, however, is by increasing the open-loop gain G_0, or more generally, by increasing the gain-bandwidth product $G_0 p_1$.

Our basic design philosophy, therefore, will be to choose the circuit parameters so as to maximize $G_0 p_1$. This will allow the closed-loop system to achieve the required bandwidth with a minimal decrease in gain. As we shall see, this approach will lead to acceptable results with respect to both bandwidth and noise considerations.

Consider first R_S, the source resistor of the second-stage FET of Fig. 3-2-a. Inspection of (3-28) reveals that $G_0 p_1$ is maximum when R_S is zero. The output small-signal voltage, however, is given by $v_o = i_o(R_S \parallel R_F)$. Recall our assumption that the noise of the third-stage amplifier is negligible; in order for this to be true, v_o cannot be arbitrarily small. Let $v_{o,min}$ be the minimum voltage level required to make the third-stage noise negligible; in order for v_o to exceed $v_{o,min}$ when the output current has a minimum value of $i_{o,min}$, R_S must exceed $R_{S,min}$, where $R_{S,min}$ satisfies $v_{o,min} = i_{o,min}(R_{S,min} \parallel R_F)$. A Bode plot of $G_0 p_1$ versus R_S shows that $G_0 p_1$ starts rolling off appreciably beyond a "cutoff" value of:

$$R_S = \frac{R_F}{1 + g_{m2}R_F} \approx \frac{1}{g_{m2}}. \tag{3-29}$$

The last approximation is valid when $g_{m2}R_F \gg 1$, which is typically the case. A good choice for R_S, therefore, is:

$$R_S = \max\{\frac{1}{g_{m2}}, R_{S,min}\}. \tag{3-30}$$

This provides a good balance between maximizing the gain-bandwidth product (small R_S), and providing appreciable output voltage (large R_S).

A Bode plot of $G_0 p_1$ versus R_F shows that, for all $R_F > R_S$, the gain-bandwidth product is essentially independent of R_F. The condition $R_F > R_S$ will always be satisfied in practice, because R_F must be large to reduce the effects of thermal noise. Thus, unlike R_S, there is no a priori optimal value for R_F, and so it can be chosen based on noise considerations.

Inspection of (3-28) reveals that the gain-bandwidth product is proportional to R_D, the drain resistor of the first-stage FET. Therefore, from the viewpoint of both bandwidth and noise, R_D should be chosen as large as possible. The single-pole assumption breaks down for very large R_D, however, and so R_D should not be chosen so large that the second pole p_2 becomes significant. Also, the small-signal output resistance r_o of the first FET, neglected in Fig. 3-2-b, becomes significant when R_D is too large.

3.4.1.2 Noise

Define γ_R as the ratio of thermal-noise variance to shot-noise variance:

$$\gamma_R = \frac{\sigma^2_{res}}{\sigma^2_{shot}} \tag{3-31}$$

where σ^2_{shot} and σ^2_{res} are defined by the first and second terms in (3-20), respectively. Similarly, define γ_F as the ratio of f^2-noise variance to shot-noise variance:

$$\gamma_F = \frac{\sigma^2_{FET}}{\sigma^2_{shot}}, \tag{3-32}$$

where σ^2_{FET} is defined by the third term in (3-20).

Recall that the signal-to-shot-noise ratio SNR_{shot}, given by (2-17), is proportional to detector area A_{det}. Define $A_{det,shot}$ as the detector area required for SNR_{shot} to achieve a value of SNR_{req}, given by:

$$A_{det,shot} = \frac{2qI_2BG_{bg}p_{bg}\Delta\lambda}{rp^2_{sig}}SNR_{req} \tag{3-33}$$

where p_{sig} is the received signal irradiance and p_{bg} is the detected background irradiance per unit filter bandwidth; see (2-17) and (2-18). The total SNR, including the thermal noise and f^2 noise as well as the shot noise, can be written as:

$$SNR = \frac{SNR_{shot}}{1+\gamma_R+\gamma_F} = \frac{SNR_{req}}{1+\gamma_R+\gamma_F}\left(\frac{A_{det}}{A_{det,shot}}\right). \tag{3-34}$$

We can thus view $(1 + \gamma_R + \gamma_F)$ as the total SNR penalty due to electrical noise, with γ_R as the penalty due to thermal noise and γ_F as the penalty due to f^2 noise. Larger penalties will require larger detector areas to achieve a given SNR_{req}.

In the presence of intense background noise, the thermal-noise and FET-noise penalties will be small in a well-designed receiver, in which case the shot-noise-limited assumption is approximately correct. This observation, which was first noted by Kahn *et al.* [48], simplifies the design process considerably.

3.4.2 Scenario One

In the first design scenario, assume that the received signal irradiance p_{sig} is given, and the design problem is to specify the parameter set $\{A_{det}, g_m, g_{m2}, \omega_T, \omega_{T2}, R_F, R_S, R_D\}$ to meet an SNR requirement of SNR_{req} and a bandwidth requirement of B.

If R_F and g_m can take on any value, then the designer can make γ_R and γ_F as small as desired. Let $\gamma_{R,des}$ and $\gamma_{F,des}$ be desirable values for γ_R and γ_F, respectively. For example, $\gamma_{R,des} = 0.3$ and $\gamma_{F,des} = 0.3$ results in a total SNR penalty of 2 dB. The designer can then choose R_F such that $\gamma_R = \gamma_{R,des}$ and g_m such that $\gamma_F = \gamma_{F,des}$. From (3-34), the detector area required to achieve an SNR of SNR_{req} will then be:

$$A_{det,req} = (1 + \gamma_{R,des} + \gamma_{F,des})A_{det,shot}\,. \tag{3-35}$$

The drain resistor R_D should be chosen large enough to achieve the bandwidth requirement. Implementing these ideas is complicated by the interdependence of the parameters involved. In Fig. 3-4 we present a design procedure that solves this interdependency problem by iterating a sequence of computations until the parameters converge and the constraints are satisfied.

The iteration of steps 5 through 11 is necessary because both g_m as calculated in step 5 and R_F as calculated in step 9 are functions of R_D, which itself is not determined until step 10. They are only weakly dependent on R_D, however, with complete independence as $R_D \to \infty$. Therefore, by initializing R_D to a large value in step 4, few if any iterations will be necessary.

Building on the link analysis of chapter 2, Table 3-1 presents numerical results of this design procedure. Two different room sizes with both planar and hemispherical optical filters are considered. In each case, the received signal irradiance p_{sig} and the optimal filter bandwidth $\Delta\lambda$ are taken from Table 2-1. Other assumptions, such as the background gain G_{bg} and the background irradiance per unit bandwidth p_{bg}, are listed in Table 3-2. Note that $G_{bg} = 3.24$ is the ideal n^2 gain of a lens with refractive index $n = 1.8$, and that $p_{bg} = 5.8\ \mu W/(cm^2{\cdot}nm)$ is the worst-case background noise measured in a daylight environment near a window but not exposed to direct sunlight [11]. The desired noise penalties $\gamma_{R,des}$ and $\gamma_{F,des}$ are both set to 0.3.

1. Set noise penalties $\gamma_{R,des}$ and $\gamma_{F,des}$ to desired values.
2. Calculate $A_{det,shot}$ using (3-33).
3. Calculate A_{det} using (3-35).
4. Initialize R_D to any large value.
5. Set g_m so that $\gamma_F = \gamma_{F,des}$. The exact expression for g_m is complicated and omitted for brevity; it is the positive root of the second-order polynomial in g_m formed by substituting ω_T from step 6 into (3-22), and then substituting the result into (3-32).
6. Set ω_T so that the gate capacitance C_g is negligible:

$$\omega_T = 10 \times g_m / (C_{det} + C_{stray}).$$

7. Set $g_{m2} = g_m$.
8. Set $R_S = 1/g_{m2}$.
9. Set R_F so that $\gamma_R = \gamma_{R,des}$. The exact expression for R_F is complicated and omitted for brevity; it is the positive root of the second-order polynomial in R_F obtained by substituting σ^2_{res} from (3-20) into (3-31).
10. Set R_D to meet the bandwidth requirement:

$$R_D = (2\pi B(R_F + R_S)C_T - 1)(1 + g_{m2}R)/(g_m g_{m2} R_S) .$$

This equation is based on the single-pole assumption of (3-25), so the true bandwidth may differ slightly from B.

11. Set ω_{T2} so that $p_2 = 4 \times 2\pi B$, ensuring that the secondary pole p_2 of (3-16) is negligible, even after the loop is closed:

$$\omega_{T2} = 4 \times 2\pi B g_{m2}(R_D + R)/(1 + g_{m2}R) .$$

12. Iterate steps 5 through 11, if necessary, until $|\gamma_F - \gamma_{F,des}|$ and $|\gamma_R - \gamma_{R,des}|$ are acceptably small.

Fig. 3-4. Design procedure for scenario one.

The first column of Table 3-1 presents results for a 5 m × 5 m × 3 m room with a planar optical filter. From Table 2-1, the optimal filter bandwidth in this case is $\Delta\lambda = 71.7$ nm,

TABLE 3-1: Receiver Design Results for Scenario One.

Parameter	5 m × 5 m / planar	5 m × 5 m / hemispherical	10 m × 10 m / planar	10 m × 10 m / hemispherical
A_{det}	1.56 cm^2	0.32 cm^2	24.8 cm^2	4.5 cm^2
g_m and g_{m2}	40.3 mS	62.6 mS	585 mS	870 mS
$\omega_T/2\pi$	1.1 GHz	7.5 GHz	1.1 GHz	8.7 GHz
$\omega_{T2}/2\pi$	2.5 GHz	16.9 GHz	2.4 GHz	20.3 GHz
R_F	150 Ω	4.6 kΩ	9.2 Ω	463 Ω
R_S	24.8 Ω	16 Ω	1.7 Ω	1.1 Ω
R_D	269.3 Ω	1.3 kΩ	17.4 Ω	115 Ω

TABLE 3-2: Basic Parameter Assumptions.

Parameter	Value	Description
G_{bg}	3.24	optical gain seen by background light
p_{bg}	5.8 μW/(cm^2·nm)	background irradiance per unit bandwidth
r	0.53 A/W	photodetector responsivity
c_{det}	35 pF/cm^2	photodetector capacitance per unit area
G	0.82	FET channel noise factor
I_2	0.562	noise bandwidth factor for white noise
I_3	0.0868	noise bandwidth factor for f^2 noise
I_f	0.184	noise bandwidth factor for $1/f$ noise
f_c	10 MHz	cutoff frequency for $1/f$ noise
B	100 MHz	desired bandwidth
$\gamma_{R,des}$	0.3	desired penalty due to thermal noise
$\gamma_{F,des}$	0.3	desired penalty due to FET noise
SNR_{req}	144	desired peak SNR

and the received signal irradiance is $p_{sig} = 2.6\ \mu W/cm^2$. From Table 3-1 we see that the requirements are met with a moderately high value for g_m (40.3 mS) and with reasonable values for ω_T. The required detector area is $A_{det} = 1.56\ cm^2$. Because of the large filter bandwidth, the shot noise power is high, and so a small value for R_F (150 Ω) is sufficient to make the thermal noise penalty $\gamma_R = 0.3$.

The second column of Table 3-1 presents results for the same room, but with a hemispherical filter. Table 2-1 shows that the optimal filter bandwidth is $\Delta\lambda = 11.6$ nm, and the received signal irradiance is $p_{sig} = 2.3\ \mu W/cm^2$. The detector area required is only $A_{det} = 0.32\ cm^2$. The shot noise power is low due to the narrow optical filter bandwidth and small detector area. Therefore, to make γ_R equal 0.3, the hemispherical-filter receiver requires a larger feedback resistor R_F (4.6 kΩ) than the planar-filter receiver. This in turn moves the dominant open-loop pole p_1 closer to the origin, requiring a larger R_D of 1.3 kΩ to meet the bandwidth requirement. This larger R_D requires a larger ω_{T2} ($2\pi \times 17$ GHz) to prevent the secondary pole p_2 from becoming significant.

The third column of Table 3-1 presents results for a 10 m $\times$ 10 m $\times$ 3 m room with a planar filter. From Table 2-1, $\Delta\lambda = 73.5$ nm and $p_{sig} = 0.66\ \mu W/cm^2$. The required detector area in this case is $A_{det} = 25\ cm^2$; too large to be considered seriously.

The fourth column of Table 3-1 presents results again for the 10 m $\times$ 10 m room, but with a hemispherical filter. From Table 2-1, $\Delta\lambda = 8.3$ nm and $p_{sig} = 0.52\ \mu W/cm^2$. The required detector area of $A_{det} = 4.5\ cm^2$ is large but still feasible. The required g_m is large (870 mS), which implies high power consumption. The small value for R_S (1.1 Ω) is too small to provide appreciable voltage to the third stage, so the procedure of Fig. 3-4 should be modified by fixing R_S to be somewhat larger, say 15 Ω, in step 8.

3.4.3 Scenario Two

We next consider a more constrained design scenario, in which the optical filter bandwidth, photodetector, and transistors are specified, and the problem is to specify the circuit parameters $\{R_S, R_F, R_D\}$ such that the total power penalty $(1 + \gamma_R + \gamma_F)$ is fixed at a desired value of $(1 + \gamma_{TOT,des})$ with a bandwidth B. Here, γ_{TOT} is defined by the sum $\gamma_{TOT} = \gamma_R + \gamma_F$. Also of interest is the signal irradiance p_{sig} required to achieve an SNR of SNR_{req}.

The principles in this case are similar to those of the first scenario, so this discussion will be brief. First, initialize R_D to a large value, and choose $R_S = 1/g_{m2}$ for the reasons mentioned earlier; see (3-29). Second, calculate γ_F from (3-32). Third, using step 9 of the procedure of Fig. 3-4, calculate R_F such that $\gamma_R = \gamma_{R,des} \equiv \gamma_{TOT,des} - \gamma_F$. Fourth, using step 10 of the procedure of Fig. 3-4, compute the value for R_D required to achieve the bandwidth requirement. As before, iterations may be necessary to fine tune the resulting parameters. Finally, the required signal irradiance p_{sig} can be found by substituting (3-33) into (3-34) with $SNR = SNR_{req}$ and solving the result for p_{sig}.

As a numerical example, assume the chosen detector has area $A_{det} = 1\ \text{cm}^2$. Assume the two FETs are identical with $g_m = g_{m2} = 70$ mS and $\omega_T = \omega_{T2} = 2\pi \times 44$ GHz; these correspond to the GaAs HEMTs (OKI KGF1850) proposed for use in an experimental prototype [49]. Assume the filter bandwidth is $\Delta\lambda = 23$ nm, also proposed for use in [49]. Assume the desired value for γ_{TOT} is $\gamma_{TOT,des} = 0.5$. With these assumptions, along with those listed in Table 3-2, the procedure outlined in the previous paragraph yields the following results: $R_S = 14.3\ \Omega$, $R_D = 659\ \Omega$, and $R_F = 1.0\ \text{k}\Omega$. The resulting closed-loop low-frequency current gain is 37 dB, and the low-frequency transresistance gain is 48.5 dBΩ.

A large drain resistor R_D can be problematic because a large swing in photodetector current will result in large voltage swing V_{DS} across the first FET, risking device failure. Unfortunately, by setting $R_S = 1/g_{m2}$, the value of R_D required to achieve the desired bandwidth is rather large. The interplay between R_S and R_D can be understood as follows. From (3-27), the "loop gain" G_0H_0 must satisfy $G_0H_0 = 2\pi B/p_1 - 1$ under the single-pole assumption. Thus, for a given bandwidth B and pole p_1, this relation fixes the required loop gain, independent of R_S and R_D. From (3-11), (3-14), and (3-17), the loop gain can be written as:

$$G_0H_0 = \frac{g_m R_S g_{m2} R_D}{1 + g_{m2} R} , \tag{3-36}$$

where $R = R_S \parallel R_F$. This equation shows that, when G_0H_0 is held fixed at the value necessary to achieve the required bandwidth, smaller values of R_S require larger values of R_D. Thus, by making R_S significantly larger than $1/g_{m2}$, we can decrease the required value of R_D. The gain-bandwidth product decreases rapidly for $R_S > g_{m2}$, so the price paid for this smaller R_D is a smaller low-frequency current gain.

In the above numerical example, the low-frequency current gain is large (37 dB), and so there is room to decrease R_D by increasing R_S at the expense of decreased gain. For example, by initializing R_S to 100 Ω rather than 14.3 Ω, R_D is reduced from 659 Ω to 358 Ω, and R_F decreases slightly from 1.0 kΩ to 933 Ω. The closed-loop low-frequency current gain decreases from 37 dB to 20 dB, and the low-frequency transresistance gain decreases from 48.5 dBΩ to 39.5 dBΩ.

3.5 OPTIONAL DESIGNS

In this section we consider a number of design embellishments that can improve system performance at the cost of increased receiver complexity.

3.5.1 Feedback Zero Compensation

Our design strategy outlined in section 3.4 was to choose the transistors so that the single-pole assumption of (3-25) was always valid. This approach, although fruitful, may not be cost-effective in the long run because it may require high-performance transistors that are not conducive to monolithic integration. It is appropriate, therefore, to consider design strategies for the case in which both p_2 and p_1 in (3-24) are significant.

In Fig. 3-4 we show the root locus for the transimpedance amplifier of Fig. 3-2 [45][50]. The root locus for the ideal single-pole case of (3-25) is not shown; it is simply a straight line beginning at $s = -p_1$ and ending at $s = -\infty$. Curve (a) is the root locus for the ideal two-pole system of (3-24) with $z_2 = \infty$. When the feedback gain H_0 is zero, the poles are located at $-p_1$ and $-p_2$, as shown. As H_0 increases, the poles shift towards each other on the real axis, eventually colliding and then branching out vertically in both directions. In practice, higher order poles, neglected in our analysis, will cause the locus to bend back towards the imaginary axis, as shown by curve (b).

For the case of curve (a), the feedback can still provide a large bandwidth, but the resulting frequency response will exhibit peaking near the cutoff frequency. The more realistic case of curve (b), however, shows that higher-order effects can cause the amplifier to be unstable. Under these conditions it is advantageous to place a capacitor C_F in parallel with the feedback resistor R_F of Fig. 3-2-a, a technique referred to as feedback-zero com-

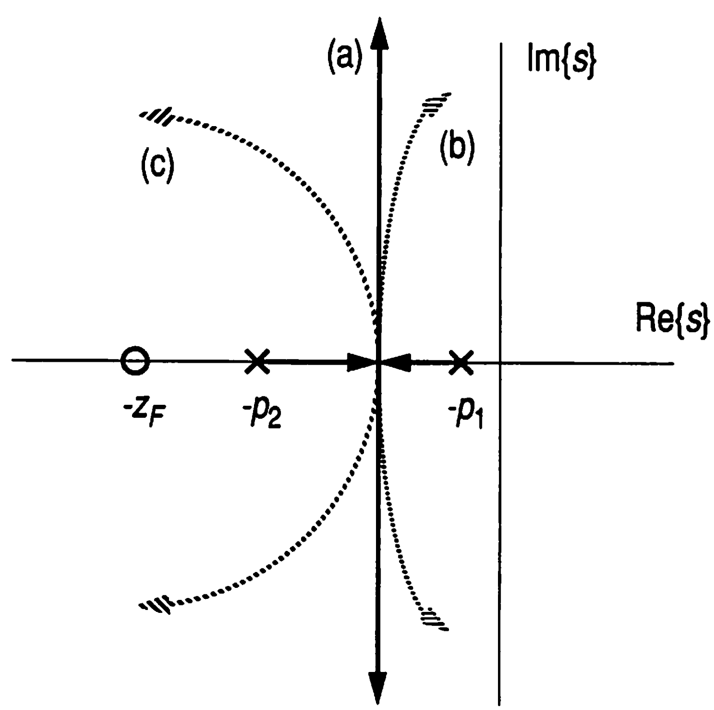


Fig. 3-4. Root locus for current-feedback pair: (a) ideal case; (b) practical case; (c) ideal case with feedback zero.

pensation [45]. Define $H(s)$ as the transfer function between the output current i_o and the feedback current i_F, as shown in Fig. 3-2; with the feedback capacitor, it is given by:

$$H(s) = H_0 \frac{1 + s/z_F}{1 + s/p_F} \approx H_0(1 + s/z_F)\,, \tag{3-37}$$

where $z_F \equiv 1/(R_F C_F)$ and $p_F \equiv z_F/H_0$; the approximation is valid because $1/H_0$, which is roughly the low-frequency closed loop gain, is large so that p_F is much larger z_F. In the ideal two-pole case, this type of compensation causes the root locus to lie on a circle centered at the feedback zero ($-z_F$), as shown in curve (c). With properly chosen feedback gain and feedback capacitor, the two poles can be arranged at $\pm$ 45° from the real axis to create a second-order Butterworth frequency response. As is evident from the figure, the feedback zero increases the total bandwidth as well as provides a margin for stability.

3.5.2 Avalanche Photodiode

Avalanche photodiodes (APDs) are advantageous over PIN photodiodes in applications where shot noise is negligible and the electrical noises of the preamplifier are dominant. They enjoy great success in fiber-optic systems, where the only sources of shot noise—photodetector dark current and the signal itself—are weak. In wireless systems, on the other hand, the shot noise from background power will usually dominate, even with a PIN diode, thus limiting the usefulness of APDs. The purpose of this section is to establish the conditions under which an APD would be useful in a wireless system.

The bias electric field of an APD is larger than that of a PIN diode, so that electron-hole pairs generated by the absorption of photons in the depletion layer will accelerate until their kinetic energy is sufficient to generate new carriers, a process known as avalanche multiplication [33]. The APD gain can be characterized in part by its mean M and second moment $M^2F(M)$, where $F(M) = kM + (1 - k)(2 - 1/M)$ is the excess noise factor for a mean gain M, and k is the ionization ratio [41][42]. The APD gain is not a Gaussian process, so its first and second means are not sufficient to characterize it completely; it is sufficient for our purposes, however, because we are interested only in second-order statistics.

A PIN diode can be thought of as an APD with $M = 1$ and $F(1) = 1$. Compared to a PIN-based receiver, the signal power with an APD receiver will be increased by a factor M^2. The shot noise power, on the other hand, will be increased by a factor $M^2F(M)$. Defining σ_{sig}^2 as the signal power with a PIN, and using the definitions of (3-20), the SNR with an APD is thus:

$$SNR_{APD} = \frac{M^2\sigma_{sig}^2}{M^2F(M)\sigma_{shot}^2 + \sigma_{res}^2 + \sigma_{FET}^2} = \frac{SNR_{shot}}{F(M) + \gamma_{TOT}/M^2}, \tag{3-38}$$

where $SNR_{shot} = \sigma_{sig}^2/\sigma_{shot}^2$ and $\gamma_{TOT} = \gamma_R + \gamma_F$. The optimal APD gain, maximizing the SNR, or equivalently maximizing the receiver sensitivity, can be found by solving the third-order polynomial formed by setting the derivative of (3-38) with respect to M to zero, with the result being:

$$M_{opt} = \max\left\{1, \left(\sqrt{\left(\frac{1-k}{3k}\right)^3 + \frac{\gamma_{TOT}^2}{k^2}} + \frac{\gamma_{TOT}}{k}\right)^{\frac{1}{3}} - \left(\sqrt{\left(\frac{1-k}{k}\right)^3 + \frac{\gamma_{TOT}^2}{k^2}} - \frac{\gamma_{TOT}}{k}\right)^{\frac{1}{3}}\right\}. \tag{3-39}$$

We see that the equation is dependent only on γ_{TOT} and k. In Fig. 3-5 we plot this optimal gain (left scale) versus the total electrical noise penalty γ_{TOT}, along with the corresponding SNR improvement (right scale), assuming $k = 0.02$. A well-designed receiver will generally have $\gamma_{TOT} < 0.5$, and from the figure we see that in this case the optimal APD gain is unity, i.e., a PIN diode is best. When $\gamma_{TOT} > 0.5$, there can be a significant SNR improvement when an APD with optimal gain is used. It is likely, however, that it would be more cost-effective to improve the preamplifier design (perhaps by increasing the transistor g_m and f_T) to reduce γ_{TOT} to an acceptable level rather than by switching to an APD.

3.5.3 Photodetector Array

This thesis has so far considered receivers using a single photodiode. There are a number of possible motivations for using multiple photodetectors, however. For example, multiple photodetectors can be separated in space to provide spatial diversity against shadowing, or they can be angled in different directions to provide angle diversity against receiver tilt. Multiple photodiodes will also be necessary when the required detector area exceeds the largest area available commercially.

One simple way to amplify the outputs of an array of photodiodes is to feed all of the photodetectors into one preamplifier. With regard to bandwidth and electrical noise, this configuration is approximately equivalent to a single photodetector with a total capaci-

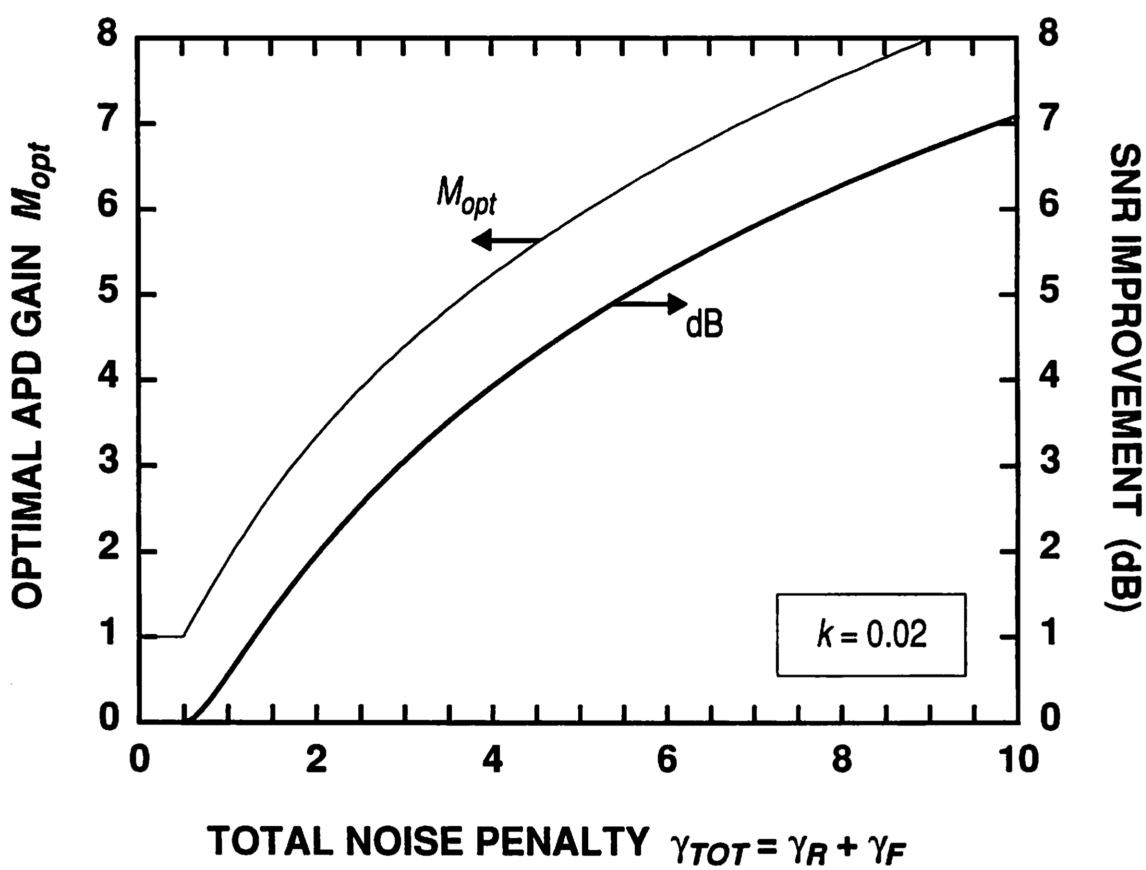


Fig. 3-5. Optimal APD gain (left scale) and the resulting SNR improvement (right scale) as a function of electrical noise penalty.

tance equal to the sum of the capacitances of the individual photodiodes. Whenever a receiver uses more than one photodetector, however, the designer has a second option: provide a separate preamplifier for each photodiode, as shown in Fig. 3-6. The purpose of this subsection is to examine the benefit of such a photodetector/preamplifier array.

Consider Fig. 3-6, which shows an array of N photodiodes, each with area A_{det}/N. Each photodiode has its own transimpedance preamplifier, such as the current-feedback pair of Fig. 3-2. The outputs of the N preamplifiers are added together to produce a total output signal i_{TOT}. Note that a single-detector receiver can be viewed as a special case with $N = 1$.

In this discussion we assume that the same signal irradiance and background irradiance is incident on each of the photodetectors. We also assume that the N preamplifiers are identical. From (3-20), the shot-noise variance in each branch is proportional to the area of one detector:

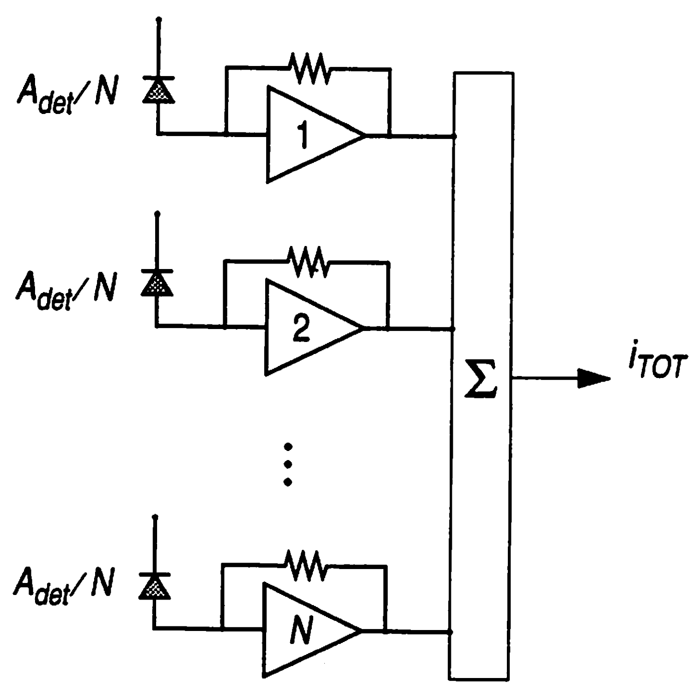


Fig. 3-6. Photodetector/preamplifier array.

$$\sigma_{shot}^2 \propto A_{det}/N\,, \tag{3-40}$$

and the thermal noise in each branch is independent of area. In addition, since the total input capacitance C_T of each preamplifier is dominated by the detector capacitance C_{det}, which itself is proportional to area A_{det}/N, the FET noise in each branch is approximately proportional to the square of the area of one detector:

$$\sigma_{FET}^2 \propto (A_{det}/N)^2\,. \tag{3-41}$$

The signal component in each branch combines coherently so that the power of the signal component after the summing node, σ_{sig}^2, is independent of N. The noise power after the summing node, on the other hand, is the sum of the noise powers in each branch. From (3-40) and (3-41), therefore, the SNR after the summing node can be written as:

$$SNR_N = \frac{\sigma_{sig}^2}{\sigma_{shot}^2 + N\sigma_{res}^2 + \sigma_{FET}^2/N} = \frac{SNR_{shot}}{1 + N\gamma_R + \gamma_F/N}\,, \tag{3-42}$$

where SNR_{shot}, γ_R, and γ_F were defined in (2-17), (3-31), and (3-32), respectively.

To characterize the benefit of using a photodetector array, define the SNR improvement as the increase in SNR over that of a single-detector receiver, assuming the preamplifiers are identical in both cases:

$$\text{SNR improvement} = \frac{SNR_N}{SNR_1} = \frac{1 + \gamma_R + \gamma_F}{1 + N\gamma_R + \gamma_F/N}\,. \tag{3-43}$$

If N is too large, the N-fold increase in thermal noise is dominant, whereas is N is too small, the N-fold decrease in FET noise is not sufficient. The optimal array size, balancing these tradeoffs and maximizing the improvement, depends only on the ratio γ_F/γ_R:

$$N_{opt} = \arg \min_N \{\, N + \frac{1}{N}\gamma_F/\gamma_R \,\}\,. \tag{3-44}$$

Although there is no closed-form expression for N_{opt}, it is one of the two integers nearest to $\sqrt{\gamma_F/\gamma_R}$, and can be calculated numerically.

In Fig. 3-7 we plot N_{opt} versus the ratio $\gamma_F/\gamma_R = \sigma^2_{FET}/\sigma^2_{res}$ along with the SNR improvement that results. The photodetector array is most beneficial when both γ_R and γ_F/γ_R are large. For example, when $\gamma_R = 1$ and $\gamma_F = 65$, we see that a receiver using an array of 8 photodiodes with 8 identical preamplifiers will have an SNR that is 6 dB larger than a receiver using the same total detector area but using only one of the same preamplifiers.

Earlier in this chapter we showed that, by choosing high-transconductance, high-f_T transistors, the SNR penalty relative to the shot-noise limit could be made small, a few dB

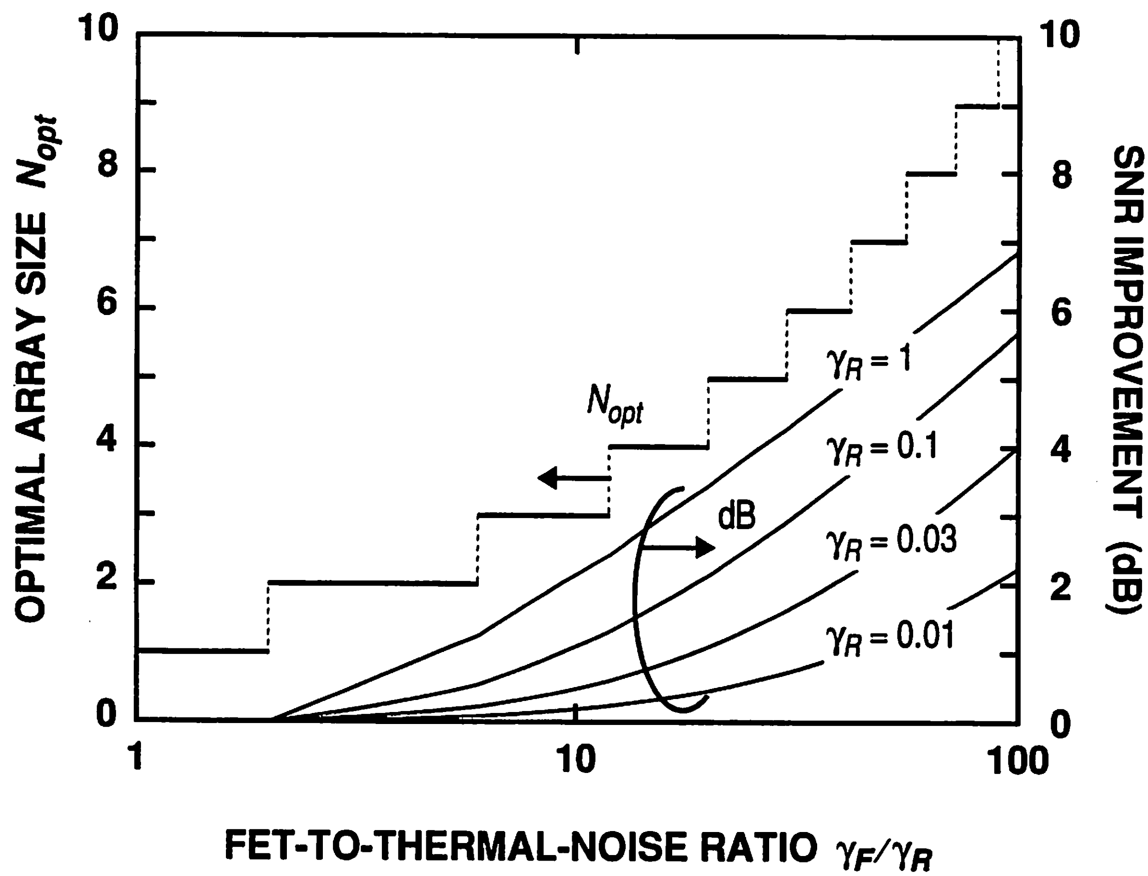


Fig. 3-7. Optimal number of photodetectors (left scale) and resulting SNR improvement (right scale) as a function of the ratio of FET noise power to thermal noise power $\gamma_F/\gamma_R = \sigma^2_{FET}/\sigma^2_{res}$.

or less. For example, the receiver designed in section 3.4.2 had small penalties; $\gamma_R = \gamma_F = 0.3$. In this case, as shown in Fig. 3-7, there is no SNR improvement gained through the use of a photodetector array; $N_{opt} = 1$.

We conclude that, when the designer has complete control over the transistors being used, a photodetector array will rarely offer any SNR improvement. There may be applications, however, in which the designer does not have such control. For example, an integrated front-end that includes the photodetector, preamplifier, and other signal processing (such as timing recovery and equalization) may require that the preamplifier transistors use the same substrate as the digital circuitry, likely silicon, which severely restricts the achievable transconductance and f_T. In such instances, a photodetector array may be beneficial.

In deriving the expression for the SNR improvement above, we made the assumption that the N preamplifiers used to calculate SNR_N in the numerator of (3-43) were *identical* to the preamplifier used to calculate SNR_1 in the denominator. This is not a fair comparison, because the difference in capacitance will cause a preamplifier designed for a photodetector with area A_{det} to differ significantly from a preamplifier designed for a photodetector with area A_{det}/N. This fact suggests that each preamplifier should be designed according to the procedure of Fig. 3-4, with the following modifications:

- Modify step 5: Set g_m so that $\gamma_F/N = \gamma_{F,des}$.
- Modify step 9: Set R_F so that $N \cdot \gamma_R = \gamma_{R,des}$.
- Modify step 12: Iterate until $\gamma_F/N = \gamma_{F,des}$ and $N \cdot \gamma_R = \gamma_{R,des}$.

The SNR improvement in this case will exceed that in (3-43), in essence because we have added an extra degree of freedom, which can only be beneficial.

3.5.4 Differential Detection and Common-Mode Rejection

The intensity of traditional fluorescent lights flickers at a rate equal to the line frequency, 60 Hz in North America. The electrical spectrum at the output of a detector in the presence of fluorescent lights consists of a train of spikes at multiples of 60 Hz, extending up to a few hundred kHz. Recently, higher efficiency fluorescent lights have been introduced that use radio-frequency ballast circuits operating at frequencies near 22 kHz, with harmonics extending to several hundred kHz. In either case, the interference caused by fluorescent lights can be detrimental to system performance.

One approach to countering fluorescent-light interference is to borrow the differential signal concept from circuit theory. Instead of transmitting the signal on a single channel and detecting this signal in an absolute sense, a second "reference" channel is used, and the receiver recovers the signal by measuring the difference between the two channels. Ideally, this reference channel is subject to the same interference as the primary channel, so that the difference is free of interference.

We propose the differential receiver front end illustrated in Fig. 3-8. Two photodiodes are reverse biased in piggyback fashion. In front of the two photodiodes are two narrowband optical filters. The filter for the primary channel, labeled filter 1, is chosen with a center wavelength near the signal wavelength, in the manner prescribed in section 2.4.5. The filter for the reference channel is chosen to have an identical bandwidth, but with a

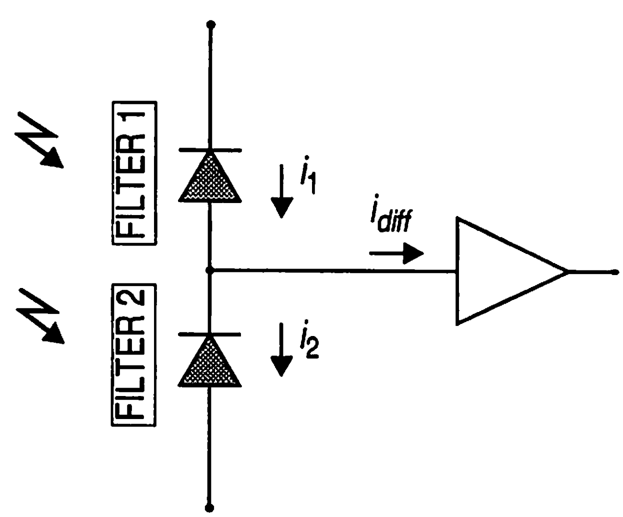


Fig. 3-8. A differential receiver front end utilizing common-mode rejection.

center wavelength different from the signal wavelength. Defining $i_1(t)$ and $i_2(t)$ as the photodetector currents for the primary and reference channels, respectively, they are given by:

$$i_1(t) = I_{bg} + i_{fluor}(t) + n_{shot,1}(t) + i_{sig}(t) \tag{3-45}$$

$$i_2(t) = I_{bg} + i_{fluor}(t) + n_{shot,2}(t) \tag{3-46}$$

where $I_{bg} = rG_{bg}p_{bg}\Delta\lambda A_{det}$ is the d.c. current offset due to background light, $i_{fluor}(t)$ is the fluorescent-light interference, $n_{shot,1}(t)$ and $n_{shot,2}(t)$ are the shot noises of the two photodiodes, and $i_{sig}(t)$ is the desired signal current.

The current $i_{diff}(t)$ that feeds into the preamplifier of Fig. 3-8 is simply the difference between $i_1(t)$ and $i_2(t)$:

$$i_{diff}(t) = i_{sig}(t) + n_{shot,1}(t) - n_{shot,2}(t). \tag{3-47}$$

The fluorescent light interference is thus eliminated.

The differential receiver of Fig. 3-8 has a second advantage; besides eliminating fluorescent-light interference, it also eliminates the d.c. current I_{bg}. In a single-detector receiver, such as the current-feedback pair of Fig. 3-2, large variations in I_{bg} can cause large voltage swings that can shift the biasing circuitry away from the desired operating point. For example, the receiver in the numerical example of section 3.4.3 has a detector area of 1 cm^2 and a filter bandwidth of 23 nm, so that, using the parameters of Table 3-2, the background current near a window is I_{bg} = 230 μA. If the window is closed, the background current in a well-lit room using fluorescent lights will drop to about I_{bg} = 1.5 μA [11]. With a feedback resistance of R_F = 1 kΩ, this 228.5-μA current swing will cause the gate-source voltage of the first FET to change by 228.5 mV, significantly altering the transistor g_m, and possibly endangering the device as well. Active bias control can be used to monitor current changes and adjust the bias voltages accordingly, at the expense of increased circuit complexity. The differential receiver of Fig. 3-8 completely bypasses this problem by eliminating the d.c. current before the preamplifier.

There are a few disadvantages of the differential front end. First, its operation relies on the fact that both channels are subject to the same interference. This requires that both channels be matched precisely, with identical optics, filter bandwidths, and detector responsivities. Furthermore, it assumes that both photodetectors are subject to identical background radiation, a condition that is difficult to achieve in practice because of unpredictable shadowing. Nevertheless, even with some channel mismatch, the differential configuration is still advantageous, because it still reduces I_{bg} and the fluorescent-light interference.

A second disadvantage of the differential front end is that it has twice the shot-noise power of a single-detector front end. This problem can be mitigated by low-pass filtering the current of the reference photodetector before subtracting it from the current of the primary detector. This would require that each photodetector have its own bias voltage.

Note that, with independent biasing, the channel mismatch problem could be actively mitigated using d.c. feedback to match the d.c. current from the reference detector with that of the primary detector. In fact, this approach eliminates the need for duplicating the primary-channel optics for the reference channel altogether; a smaller photodetector with an inexpensive filter and no optical antenna would suffice.

3.6 SUMMARY AND CONCLUSIONS

This chapter has looked at various aspects of the receiver design problem. We presented a detailed noise and bandwidth analysis of a transimpedance preamplifier using a current-feedback pair. We discussed general design strategies and presented specific design procedures for two typical design scenarios. The results indicate that, for a bandwidth of 100 MHz and detector areas near $1\,\mathrm{cm}^2$, the preamplifier can be designed to incur a small penalty (2 dB or less) with respect to the shot-noise limit, using transistors with high transconductance and high f_T. Larger detector areas can be accommodated by using higher-performance transistors. From a systems standpoint, therefore, the effects of electrical noise in the preamplifier can be lumped into a small penalty of at most a few dB and then forgotten.

We have considered FETs only, primarily because of their superior noise performance in fiber-optic receivers [41][42]. When compared to an FET, a bipolar transistor offers a higher transconductance at the expense of higher noise (arising from base current shot noise) and higher capacitance [44]. In our application, however, the background shot noise dominates other noises, and the detector capacitance dominates other input capacitances, and so future work should examine the suitability of bipolar transistors for a wireless receiver.

The design philosophy presented here has concentrated on achieving acceptable performance, with expense a secondary concern. In a commercial application, however, expense will be of primary concern. Future work on this problem should address this fact, and explore receiver designs that do not rely on expensive, discrete transistors. To this end, a combination of the design alternatives discussed in the last section (zero-feedback compensation, photodetector/preamplifier arrays, differential detection, and APDs) may be beneficial.

CHAPTER 4

MULTIPATH DISPERSION

In this chapter we examine the severity of multipath dispersion in typical indoor environments and assess its effect on system performance. We present a recursive method for calculating the impulse response of a room with Lambertian reflectors [51][52]. The method, which accounts for multiple reflections of any order, enables accurate analysis of the effects of multipath dispersion on high-speed indoor optical communication. We present a simple algorithm for computer implementation of the technique. We present computer simulation results for both line-of-sight and diffuse transmitter configurations. In both cases we find that reflections of multiple order are a significant source of intersymbol interference. We also report experimental measurements of optical multipath, which help verify the accuracy of our simulations.

4.1 INTRODUCTION

A non-directed wireless optical communication system can be categorized as either line-of-sight (LOS) or diffuse. In chapter 2 we presented a link analysis for a LOS system, which is designed under the assumption that the LOS path between transmitter and receiver is unobstructed. We define a diffuse system as one which does not rely upon the LOS path, but instead relies on reflections from a large diffusive reflector such as the ceiling. In both cases, the optical signal in transit from transmitter to receiver undergoes temporal dispersion due to reflections from walls and other reflectors; the intersymbol interference (ISI) that results is a primary impediment to communication at high speeds.

For fixed transmitter and receiver locations, multipath dispersion is completely characterized by an impulse response $h(t)$, defined such that the intensity of the received optical signal is the convolution of $h(t)$ with the intensity of the transmitted optical signal. Mobile transmitters, receivers, and reflectors will result in a time-varying channel, but we will ignore this effect because the channel will vary slowly relative to the bit rate for most indoor applications. In this chapter we present a method for calculating the impulse response of a room with an arbitrarily placed transmitter and receiver. Once calculated, the impulse response can be used to analyze or simulate the effects of multipath dispersion on the performance of indoor optical communications systems.

Other researchers have modeled indoor reflections of infrared with the purpose of determining the distribution of power throughout a room. Gfeller and Bapst present such an analysis in [11] that accounts for single reflections only; Hash *et al.* extended their procedure to include double reflections as well [53]. The simulations in these works were meant for a link budget analysis; thus, only the total power reaching the receiver was estimated. In other words, they were concerned only with the time integral of $h(t)$.[1] Since power budgets typically have built-in safety margins, the accuracy provided by considering only first- and second-order reflections was sufficient. Hortensius extended the

1. For this intensity-in, intensity-out channel, the zero-frequency value of its frequency response, $H(0) = \int_{-\infty}^{\infty} h(t)dt$, is the fraction of power emitted from a continuous-wave transmitter that is detected by the receiver.

Gfeller and Bapst model to calculate an impulse response, accounting for single reflections only [54].

In contrast to prior work, the method described here can compute the impulse response accounting for any number of reflections. This allows accurate power distribution analysis, and, perhaps more importantly, accurate impulse-response analysis. The latter is necessary because signal energy undergoing two or more reflections, although having a reduced amplitude, arrives at the receiver much later than first-order reflections. This temporal spread is critical in high-speed applications, in which case higher-order reflections cannot be ignored.

In the next section we define the models upon which our procedure is based. In section 4.3 we describe our recursive algorithm and present a computer implementation. In section 4.4 we present simulation results and compare them to experimentally measured results. Finally, to illustrate the impact of the multipath dispersion on system design, we examine the multipath-induced power penalty in section 4.5.

4.2 MODELS

In this discussion we limit consideration to empty rectangular rooms, although our techniques can be extended to other rooms in a straightforward manner. We next define the models for the source, reflectors, and receiver.

4.2.1 Source and Receiver Models

A wide-beam optical source can be represented by a position vector $\mathbf{r}_S$, a unit-length orientation vector $\hat{\mathbf{n}}_S$, a power P_S, and a radiation intensity pattern $R(\phi,\theta)$, defined as the optical power per unit solid angle emitted from the source at position (ϕ,θ) with respect to $\hat{\mathbf{n}}_S$. Following Gfeller [11], we model a source using a generalized Lambertian radiation pattern having uniaxial symmetry (independent of θ):

$$R(\phi) = \frac{n+1}{2\pi} P_S \cos^n(\phi) \qquad \text{for } \phi \in [-\pi/2, \pi/2]. \tag{4-1}$$

Here, n is the *mode number* of the radiation lobe, which specifies the directionality of the source. This is illustrated in Fig. 4-1, where sources with higher directionality are seen to have larger mode numbers. The coefficient $(n+1)/2\pi$ ensures that integrating $R(\phi)$ over the surface of a hemisphere results in the source power P_S. A mode of $n = 1$ corresponds to a traditional Lambertian source.

To simplify notation, a point source $\mathcal{S}$ that emits a unit impulse of optical intensity at time zero will be denoted by an ordered three-tuple:

$$\mathcal{S} = \{\mathbf{r}_S, \hat{\mathbf{n}}_S, n\} \tag{4-2}$$

where $\mathbf{r}_S$ is its position, $\hat{\mathbf{n}}_S$ is its orientation, and n is its mode number. Linearity allows us to consider only unit-impulse sources and scale the results for other sources.

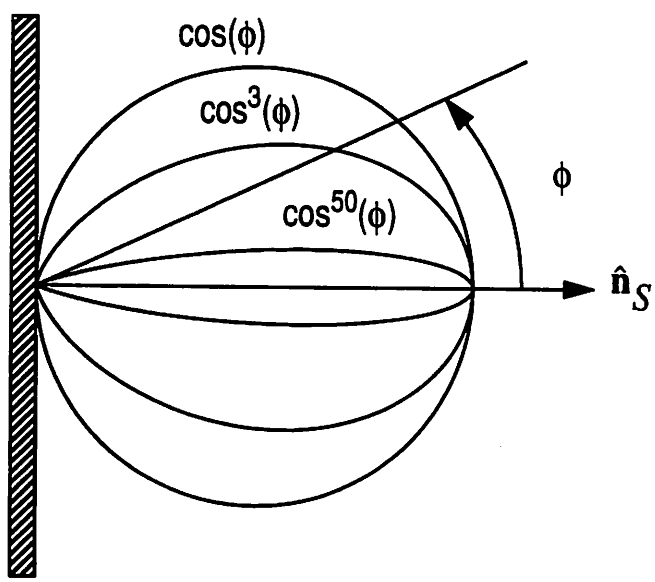


Fig. 4-1. Normalized shape of the generalized Lambertian radiation pattern.

Similarly, a receiving element $\mathcal{R}$ with position $\mathbf{r}_R$, orientation $\hat{\mathbf{n}}_R$, area A_R, and field of view FOV will be denoted by an ordered four-tuple:

$$\mathcal{R} = \{\mathbf{r}_R, \hat{\mathbf{n}}_R, A_R, FOV\}. \tag{4-3}$$

The scalar angle FOV is defined such that a receiver only detects light whose angle of incidence (with respect to the detector normal $\hat{\mathbf{n}}_R$) is less than FOV. A limited field of view may be an inadvertent effect of detector packaging, or it may be used intentionally to reduce unwanted reflections or noise.

4.2.2 Reflector Model

Although true reflections contain both specular and diffusive components [55], we make the simplifying assumption that all reflectors are purely diffusive, ideal Lambertian. Experimental measurements have shown that many typical materials such as plaster walls, acoustic-tiled walls, carpets, and unvarnished wood are well-approximated as Lambertian reflectors [11][53][56].

The radiation intensity pattern $R(\phi)$ emitted by a differential element of an ideal diffuse reflector is independent of the angle of the incident light. This fact is key to our results because it allows us to decompose a reflection into two sequential steps: to model the reflection from a differential reflecting element with area dA and reflectivity ρ, first consider the element as a receiver with area dA, and calculate the power dP it receives. Second, model the differential reflector as a source with total power $P = \rho dP$ and an ideal Lambertian radiation intensity pattern, as given by (4-1) with $n = 1$.

4.2.3 Line-of-Sight Impulse Response

Consider a source $\mathcal{S}$ and receiver $\mathcal{R}$, as specified by (4-2) and (4-3), in an environment with no reflectors; see Fig. 4-2. If the distance R between transmitter and receiver is large relative to the detector size, so that $R^2 \gg A_R$, then the received irradiance is approximately constant over the surface of the detector, and furthermore, all of the signal energy will arrive at the receiver at approximately the same time. Thus, using the models described above, the impulse response for this simple system is approximately a scaled and delayed Dirac delta function:

$$h^{(0)}(t;\, \mathcal{S},\mathcal{R}) \approx \frac{n+1}{2\pi}\cos^{n}(\phi)\mathrm{d}\Omega\ \mathrm{rect}(\theta/\mathrm{FOV})\ \delta(t - R/c), \tag{4-4}$$

where $d\Omega$ is the solid angle subtended by the receiver's differential area (assuming $A_R \ll R^2$):

$$\mathrm{d}\Omega \approx \cos(\theta)A_R/R^2\,, \tag{4-5}$$

R is the distance between the source and receiver:

$$R = \| \mathbf{r}_S - \mathbf{r}_R \| \tag{4-6}$$

θ is the angle between $\hat{\mathbf{n}}_R$ and $(\mathbf{r}_S - \mathbf{r}_R)$:

$$\cos(\theta) = \hat{\mathbf{n}}_R \cdot (\mathbf{r}_S - \mathbf{r}_R)/R, \tag{4-7}$$

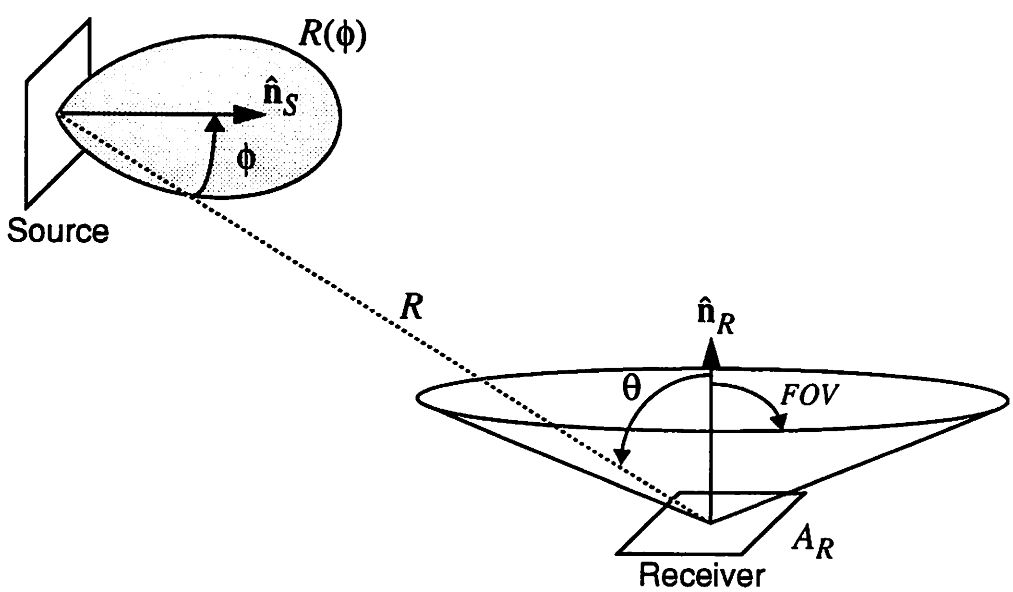


Fig. 4-2. Geometry of source and detector, without reflectors.

ϕ is the angle between $\hat{\mathbf{n}}_S$ and $(\mathbf{r}_R - \mathbf{r}_S)$:

$$\cos(\phi) = \hat{\mathbf{n}}_S \cdot (\mathbf{r}_R - \mathbf{r}_S)/R, \tag{4-8}$$

the rectangular function is defined by:

$$\mathrm{rect}(x) = \begin{cases} 1 \text{ for } |x| \le 1 \\ 0 \text{ for } |x| > 1 \end{cases}, \tag{4-9}$$

and c is the speed of light. The approximation of (4-4) approaches equality as the ratio A_R/R^2 approaches zero.

4.3 MULTIPLE-BOUNCE IMPULSE RESPONSE

We now describe our algorithm for calculating a multiple-bounce impulse response, after which we discuss a computer implementation.

4.3.1 Algorithm

Given a particular source $\mathcal{S}$ and receiver $\mathcal{R}$ in a room with reflectors, light from the source can reach the receiver after any number of reflections. Therefore, the impulse response can be written as an infinite sum:

$$h(t;\mathcal{S},\mathcal{R}) = \sum_{k=0}^{\infty} h^{(k)}(t;\mathcal{S},\mathcal{R}), \tag{4-10}$$

where $h^{(k)}(t)$ is the response of the light undergoing *exactly* k reflections. The line-of-sight response $h^{(0)}(t)$ is given by (4-4), while higher-order terms ($k > 0$) can be calculated recursively:

$$h^{(k)}(t;\mathcal{S},\mathcal{R}) = \int_S h^{(0)}(t;\mathcal{S},\{\mathbf{r}, \hat{\mathbf{n}}, \pi/2, dA\}) \otimes h^{(k-1)}(t;\{\mathbf{r}, \hat{\mathbf{n}}, 1\}, \mathcal{R}) \tag{4-11}$$

where the symbol $\otimes$ denotes convolution. More explicitly, substituting (4-4) and performing the convolution results in:

$$h^{(k)}(t;S,\mathcal{R}) = \frac{n+1}{2\pi}\int_S \frac{\rho_{\mathbf{r}}\cos^n(\phi)\cos(\theta)}{R^2}\mathrm{rect}(2\theta/\pi)h^{(k-1)}(t-R/c;\{\mathbf{r},\hat{\mathbf{n}},1\},\mathcal{R})\,dA\,. \quad (4\text{-}12)$$

The integrations in (4-11) and (4-12) are performed with respect to the position $\mathbf{r}$ on the surface S of all reflectors. Here, $\hat{\mathbf{n}}$ is the normal to the reflector surface S at position $\mathbf{r}$, dA is the differential area of the reflector surface at position $\mathbf{r}$, $\rho_{\mathbf{r}}$ is the reflectivity at position $\mathbf{r}$, $R = \| \mathbf{r} - \mathbf{r}_S \|$, $\cos(\phi) = \hat{\mathbf{n}}_S \cdot (\mathbf{r} - \mathbf{r}_S)/R$, and $\cos(\theta) = \hat{\mathbf{n}} \cdot (\mathbf{r}_S - \mathbf{r})/R$.

Equation (4-11) is the main theoretical result of this chapter. Intuitively, it says that the k-bounce impulse response from a single point-source $\mathcal{S}$ can be found by first finding the distribution and timing of the power from $\mathcal{S}$ onto the reflecting walls; then, using the walls as a distributed light source, computing the $(k-1)$-bounce impulse response.

4.3.2 Implementation

The integral in (4-11) can be calculated numerically by subdividing the reflecting surfaces into numerous small reflecting elements, each with area ΔA. Thus, $h^{(k)}(t)$ can be approximated by:

$$h^{(k)}(t;\mathcal{S},\mathcal{R}) \approx \sum_{i=1}^{N} h^{(0)}(t;\mathcal{S},E_i) \otimes h^{(k-1)}(t;\mathcal{E}_i,\mathcal{R})$$

$$= \frac{n+1}{2\pi}\sum_{i=1}^{N} \frac{\rho_i\cos^n(\phi)\cos(\theta)}{R^2}\mathrm{rect}(2\theta/\pi)h^{(k-1)}(t-R/c;\{\mathbf{r},\hat{\mathbf{n}},1\},\mathcal{R})\Delta A \quad (4\text{-}13)$$

where $\mathcal{E}_i$ signifies the i-th element and N is the total number of elements. Note how $\mathcal{E}_i$ plays the role of both an elemental receiver and an elemental source. This spatial discretization will cause temporal discretization as well, turning the normally piecewise-continuous function of time $h^{(k)}(t)$ into a finite sum of scaled delta functions; temporal smoothing can be achieved by subdividing time into bins of width Δt and summing the total power received in each bin.[1] The resulting histogram closely approximates the actual $h^{(k)}(t)$, achieving equality as ΔA and Δt approach zero.

Direct implementation of (4-13) is not efficient for reflection orders k greater than one, because identical computations would then be performed multiple times. To see this, consider a room with its reflectors subdivided into a total of N elements. Then, in calculating (4-13), a total of N^k *elementary* computations are performed, where one elementary computation is defined as the calculation of differential power and delay from a point source to a differential receiver, as in (4-4). Thus, an elementary computation consists of the collection of multiplications and vector dot-products described in (4-5) through (4-9). However, there are only $(N+1)^2$ *unique* elementary computations, corresponding to the line-of-sight impulse response from any element (including source) to any element (including receiver). Therefore, a more efficient approach would be to construct two look-up tables, each consisting of $(N+1)^2$ entries. The first, call it `dP(i,j)`, should contain the differential power between element i and element j; the second, call it `tau(i,j)`, should contain the delay between element i and element j. With these two tables, a numerical procedure for calculating the k-bounce impulse response is easily implemented, as illustrated by the following pseudocode:

```
function h(t;i,j,k)
begin
    if (k=0)
       return  dP(i,j) × delta(t - tau(i,j))
    else
       return  sum from e = 1 to N
               rho(e) × dP(i,e) × h(t - tau(i,e);e,j,k-1)
end
```

Here, `h(t;i,j,k)` is a function that returns the k-bounce impulse response $h^{(k)}(t)$ with element i as the source and element j as the receiver. The reflectivity ρ_e of the e-th reflecting element is given by `rho(e)`. This algorithm is applicable to rooms of arbitrary shape, although in the next section we present results for empty rectangular rooms only.

---

1. Empirical evidence suggests that a good choice for the bin width is $\Delta t = \sqrt{\Delta A} / c$, which is roughly the time it takes for light to travel between neighboring elements.

The table-lookup approach requires roughly $8N^2$ bytes of memory (assuming a floating-point precision of 4 bytes per entry). Thus, a modest number of reflectors can lead to prohibitive memory requirements. For example, a 10 m $\times$ 10 m $\times$ 5 m room with a spatial resolution of $\Delta A = 100\ \text{cm}^2$ has $N = 4 \times 10^4$ elements and thus requires 12.8 Gbytes of storage. To meet a more realistic storage limit of 32 Mbytes, the number of elements must satisfy $N < 2 \times 10^3$, or equivalently $\Delta A > 45\ \text{cm} \times 45\ \text{cm}$.

When the number of elements exceeds the limit imposed by memory restrictions, the direct approach must be used. The same algorithm outlined in pseudocode above is applicable to the direct approach, except that each occurrence of `dP(i,j)` and `tau(i,j)` must be calculated anew.

The time required to compute $h^{(k)}(t)$ is roughly proportional to the number of bottom-level function calls, which from inspection of the above pseudocode is N^k. The run time is thus exponential in k, which severely limits the number of reflections that can be computed in a reasonable amount of time. Using the table-lookup approach in the C programming language on a Sun Sparcstation 2, we derived an empirical run-time estimate of $N^k \times 4\ \mu\text{s}$. Thus, for example, to compute the $k = 3$ bounce impulse response with $N = 2776$ elements (these numbers are extracted from results of the next section; see the last three columns of Table 4-1), the run time is roughly 24 hours, whereas to compute the $k = 4$ bounce impulse response with the same number of elements would require about N days or 7.5 years. Reducing N would shorten the run time at the expense of reduced accuracy. We resist this temptation and present results for reflections up to third-order only.

4.4 RESULTS

In the next two subsections we present impulse responses from both simulation and experimental measurement. These impulse responses $h(t)$ are defined as the received optical intensity when the transmitted optical intensity is a unit-area Dirac delta function. Therefore, the d.c. gain $H(0) = \int_{-\infty}^{\infty} h(t)dt$ is related to the average received power P_r by

$P_r = P_T H(0)$, where P_T is the average transmitted power. We will find it convenient to compare results under the assumption that the transmitted power is 1 W, and so we often associate the integral of $h(t)$ with the average received power due to a 1-W transmitter.

4.4.1 Simulation Results

A computer program was written that implements the algorithm described in the previous section. The user can specify the various parameters listed in Table 4-1. We equate north with $\hat{\mathbf{x}}$ and west with $\hat{\mathbf{y}}$. The elevation angles are defined with respect to the horizontal plane, so that a source pointing straight down has an elevation of $-90°$, and a receiver pointing straight up has an elevation of 90°. The azimuth angle at position **r** is defined as the angle between $\hat{\mathbf{x}}$ and the projection of **r** onto the x-y plane, with a sign defined so that $\hat{\mathbf{y}}$ has an azimuth of 90°. The final set of parameters in the table control the resolution of the simulation. Here, Δt is the bin width of the power histogram that approximates the impulse response, and *bounces* is the number of reflections that are considered. The spatial resolution of the simulation is specified by the number of partitions per dimension; the total number of differential reflecting elements is then given by:

$$N = 2(N_x N_y + N_x N_z + N_y N_z)\ . \tag{4-14}$$

In Fig. 4-3 we show the simulated impulse responses $h^{(k)}(t)$, $k \in \{0,1,2,3\}$, for the configuration A given in the first column of Table 4-1. (We consider the other columns in the next section.) We were able to use higher spatial resolutions for the lower-order reflections, because their run-times are short; the number of partitions per dimension for each bounce are indicated in the last three rows of Table 4-1. The time origin is defined by the arrival of the line-of-sight impulse. Each of the responses is labeled by the total power it would carry if the source emitted 1 W in continuous-wave mode. Thus, $h^{(0)}(t)$ is a delta function, scaled by 1.23×10^{-6}. The numbers in parenthesis specify the percentage of power due to that pulse. The first-order response is seen to have four peaks, corresponding to the four walls of the room; assuming a 1-W source, the total power from once-reflected

TABLE 4-1: Parameters for Simulation and Experiment.

	parameter	Configuration A			Configuration B			Configuration C			Configuration D		
room:	length (x)	5 m			7.5 m			7.5 m			7.5 m		
	width (y)	5 m			5.5 m			5.5 m			5.5 m		
	height (z)	3 m			3.5 m			3.5 m			3.5 m		
	ρ_{NORTH}	0.8			0.30			0.58			0.58		
	ρ_{SOUTH}	0.8			0.56			0.56			0.56		
	ρ_{EAST}	0.8			0.30			0.30			0.30		
	ρ_{WEST}	0.8			0.12			0.12			0.12		
	$\rho_{CEILING}$	0.8			0.69			0.69			0.69		
	ρ_{FLOOR}	0.3			0.09			0.09			0.09		
source:	mode	1			1			1			1		
	x	2.5			2.0			5.0			3.75		
	y	2.5			4.0			1.0			2.75		
	z	3			3.3			3.3			1.0		
	elevation	−90°			−90°			−70°			+90°		
	azimuth	0°			0°			10°			0°		
receiver:	area	1 cm^2			1 cm^2			1 cm^2			1 cm^2		
	FOV	85°			70°			70°			70°		
	x	0.5 m			6.6 m			2.0 m			6.0 m		
	y	1.0 m			2.8 m			4.0 m			0.8 m		
	z	0 m			0.8 m			0.8 m			0.8 m		
	elevation	90°			90°			90°			90°		
	azimuth	0°			0°			0°			0°		
resolution:	Δt	0.2 ns			0.2 ns			0.2 ns			0.2 ns		
	bounces	1	2	3	1	2	3	1	2	3	1	2	3
	N_x	500	100	25	750	150	30	750	150	30	750	150	30
	N_y	500	100	25	550	110	22	550	110	22	550	110	22
	N_z	300	60	15	350	70	14	350	70	14	350	70	14

light is 0.505 μW. Total power is seen to decrease for each of the higher-order impulse responses; however, they tend to add to a significant amount, as shown in the sum impulse response at the bottom of the figure. Furthermore, this energy arrives much later than that from lower-order reflections.

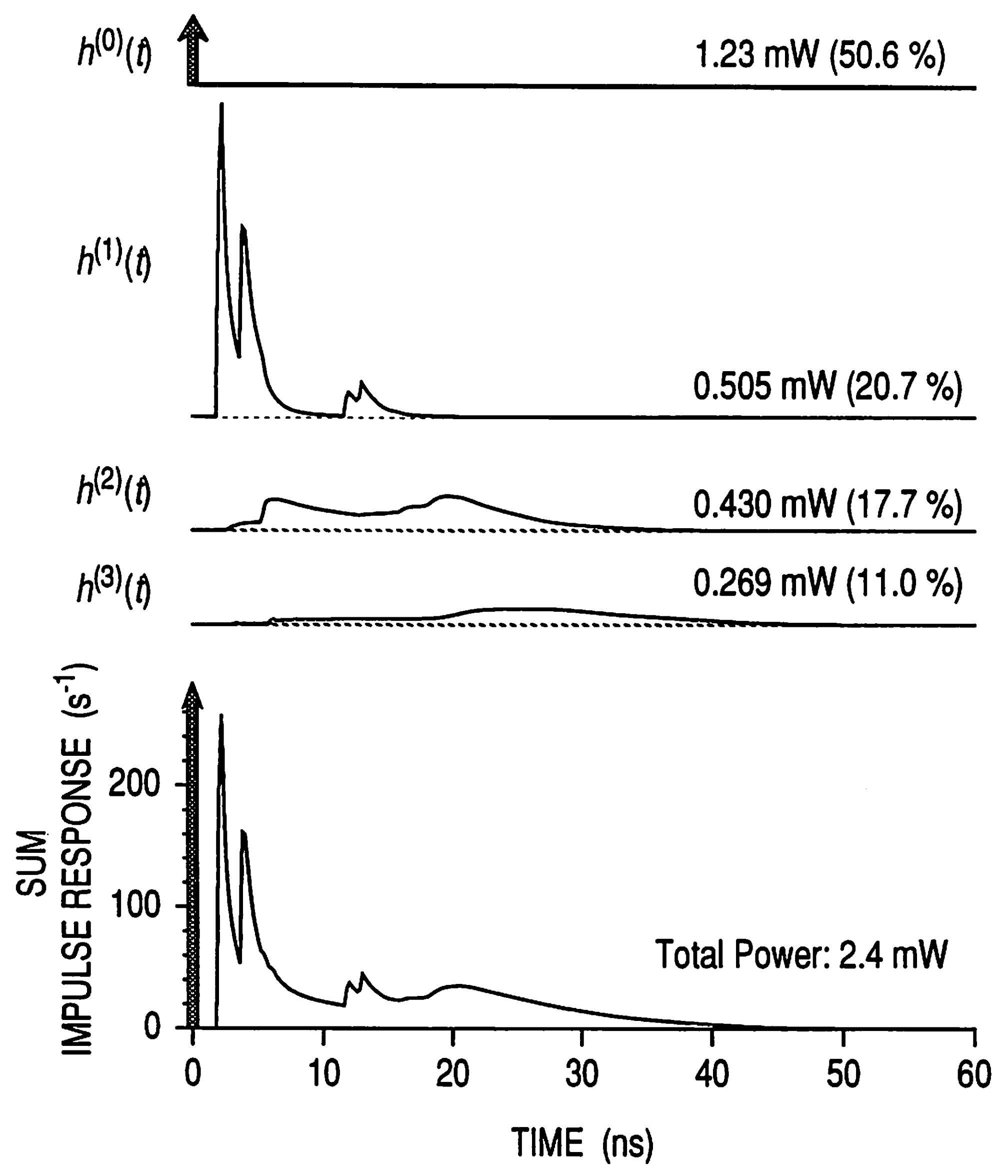


Fig. 4-3. Impulse responses for light undergoing $k \in \{0,1,2,3\}$ reflections and their sum; for configuration A, assuming a source power of 1 W.

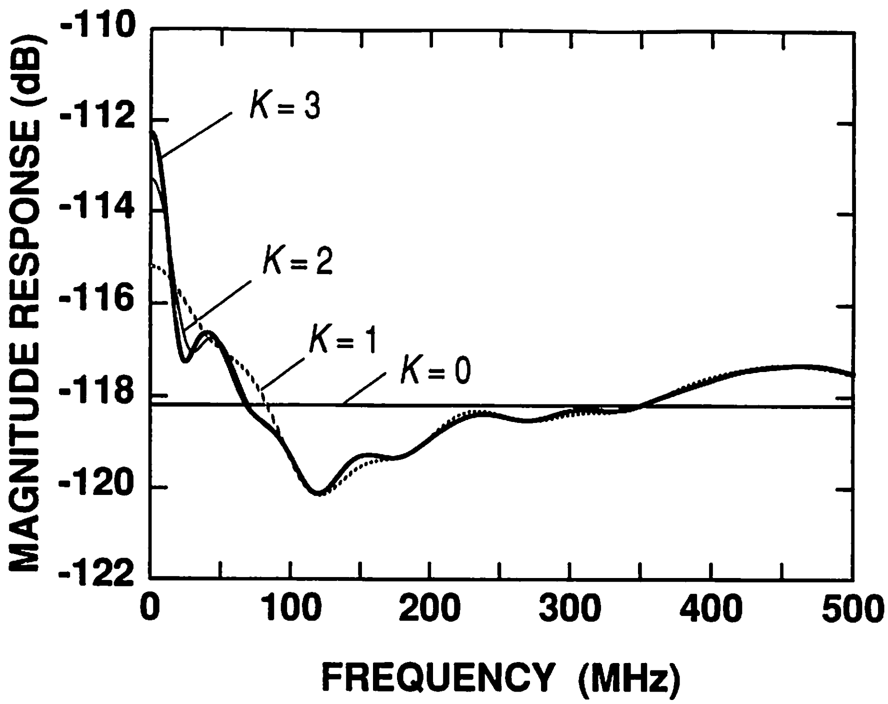
The net result is that higher-order reflections are significant. This may be easier to see in the frequency domain. Using the results of Fig. 4-3, we can estimate the frequency response of the channel:

$$H(\omega) = \int_{\infty}^{\infty} h(t)e^{-j\omega t}\,dt \approx \sum_{n=-\infty}^{\infty} h(n\Delta t)e^{-j\omega n\Delta t}\Delta t = \Delta t H(e^{j\omega\Delta t}) \tag{4-15}$$

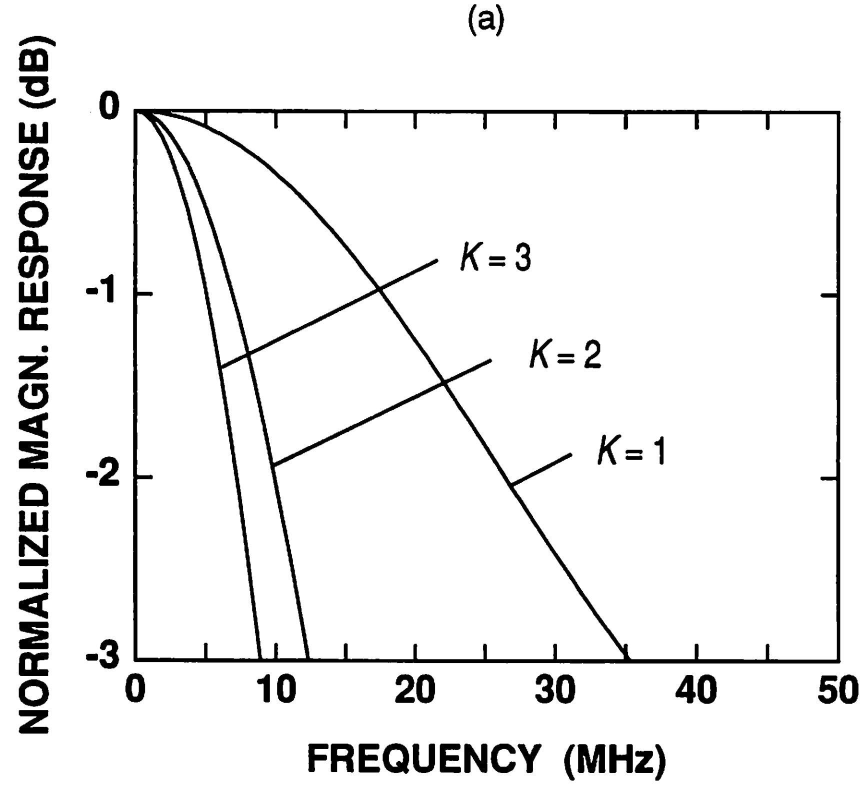
where $H(e^{j\omega\Delta t})$ is the discrete-time fourier transform of the discrete-time signal $h(n\Delta t)$. In Fig. 4-4 we plot $|\Delta t H(e^{j\omega\Delta t})|$ versus ω for impulse responses accounting for up to zero through three reflections. In other words, we approximate $h(t)$ by replacing the upper limit in the summation of (4-10) by $K \in \{0,1,2,3\}$. The curves in Fig. 4-4-a show how the higher-order reflections increase the d.c. component of the frequency response while decreasing its component at other frequencies. The d.c. gain of -112.3 dB $= 20\cdot\log_{10}(2.4 \times 10^{-6})$ for $K = 3$ implies that the receiver detects 2.4 μW for a 1-W continuous-wave transmitter. From Fig. 4-4-a we also see that the higher-order reflections have significant impact only at low frequencies; the high-frequency magnitude response is characterized by the first-order reflection only.

To highlight the effects of higher-order bounces on the -3 dB bandwidth, we normalize the magnitude responses to have unity d.c. gain in Fig. 4-4-b. This figure illustrates the need for considering higher-order reflections. Each additional reflection tends to lengthen the duration of the impulse response, which decreases the bandwidth of the channel. The $K = 1$ channel has a -3 dB bandwidth 30 MHz, while the $K = 3$ channel has a bandwidth of only 9 MHz.

In Fig. 4-5 we show the phase response and group-delay response of the channel. The group delay is defined as $-(\partial/\partial\omega)\angle H(\omega)$, so that a linear-phase channel has constant delay. The line-of-sight impulse response, being a delta function, has zero phase and thus zero delay. As the number of reflections increases, the phase response becomes less linear, and the delay response exhibits more variability.



(a)



(b)

Fig. 4-4. Effects of high-order reflections on frequency response (configuration A).

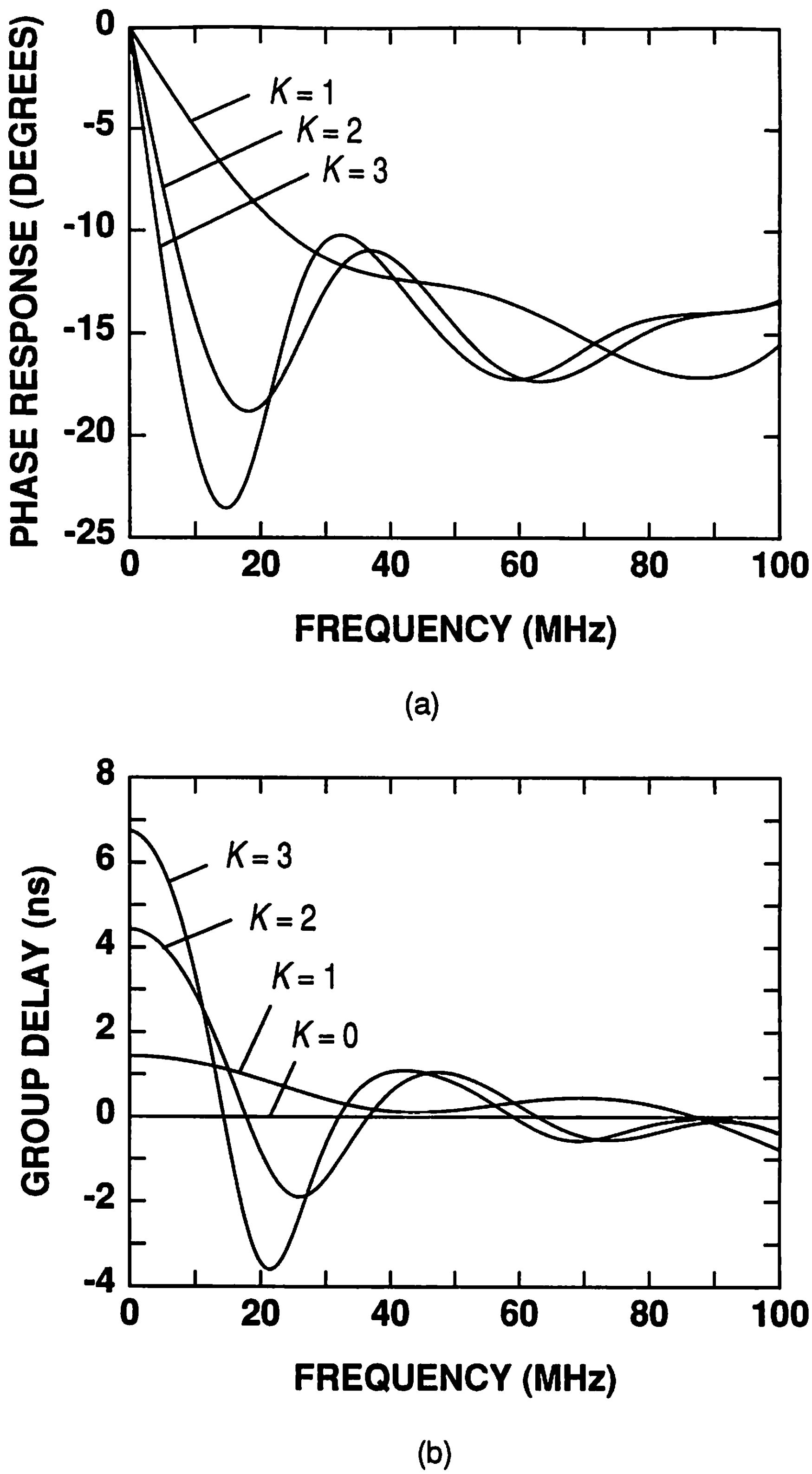


Fig. 4-5. Effects of high-order reflections on (a) phase response and (b) delay response for configuration A.

4.4.2 Experimental Results and Comparison with Simulation

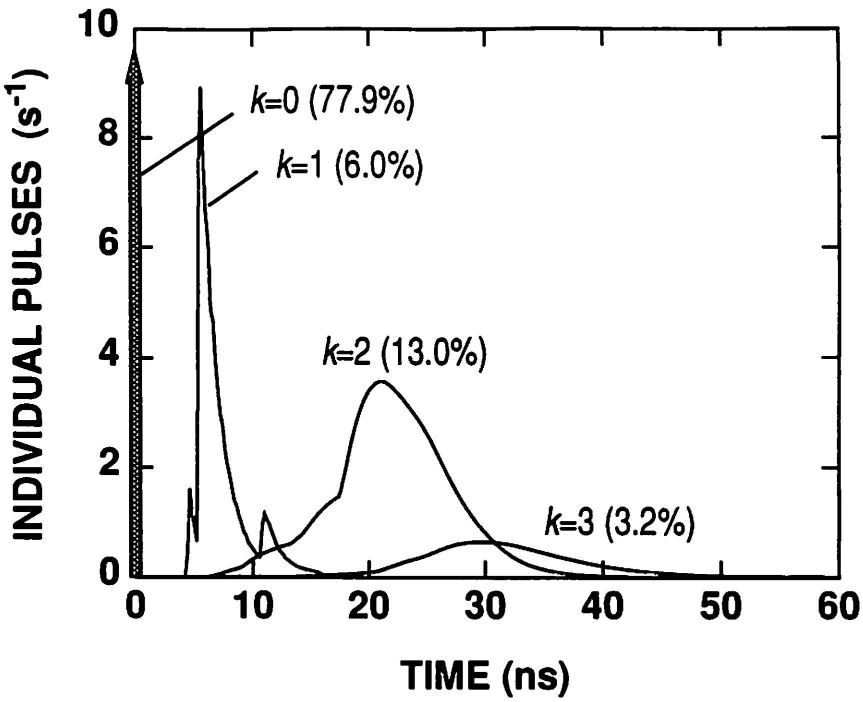
In this section we present simulation results and compare them with experimental results. The experimental data presented in this section were measured by William J. Krause and the author, using an apparatus constructed by the former [56]. The room under measurement is described in the last three columns of Table 4-1. This is an empty conference room with a wide variety of reflecting materials, including textured acoustic-tiled walls (east wall, north wall for configuration B only), rugs (floor, west wall), wood (doors on north and south walls), and painted plaster (remainder of south wall). For configurations C and D we covered the north wall with a white projection screen. The reflectivities of each of these surfaces was measured experimentally as follows: light from a laser was directed towards the material under test at normal incidence, and the reflected power was measured, also at normal incidence. The reflectivity was then chosen so that an ideal Lambertian reflector with identical reflectivity would yield the same reflected power. The results of the reflectivity measurements are shown in the table.

The transmitter in our experimental set-up was a 832-nm laser diode with peak power of 100 mW. The laser was enclosed in a metal box and illuminated a 3 cm $\times$ 10 cm area on a translucent plexiglass window which emitted a broad optical beam with an approximately Lambertian radiation pattern. The receiver front end consisted of a 0.25-cm^2 Si avalanche photodiode and a transimpedance amplifier. We measured the frequency response of the channel using a 300 kHz-3 GHz vector network analyzer. To improve noise immunity we turned off the lights during measurements and used a small IF bandwidth for averaging. To isolate the desired frequency response of the optical channel from that of the measurement system, each measured frequency response was divided by the frequency response of the measurement system. To prevent multipath from corrupting the measurement of the frequency response of the measurement system, a thin 1-m tube was held between the transmitter and receiver during calibration.

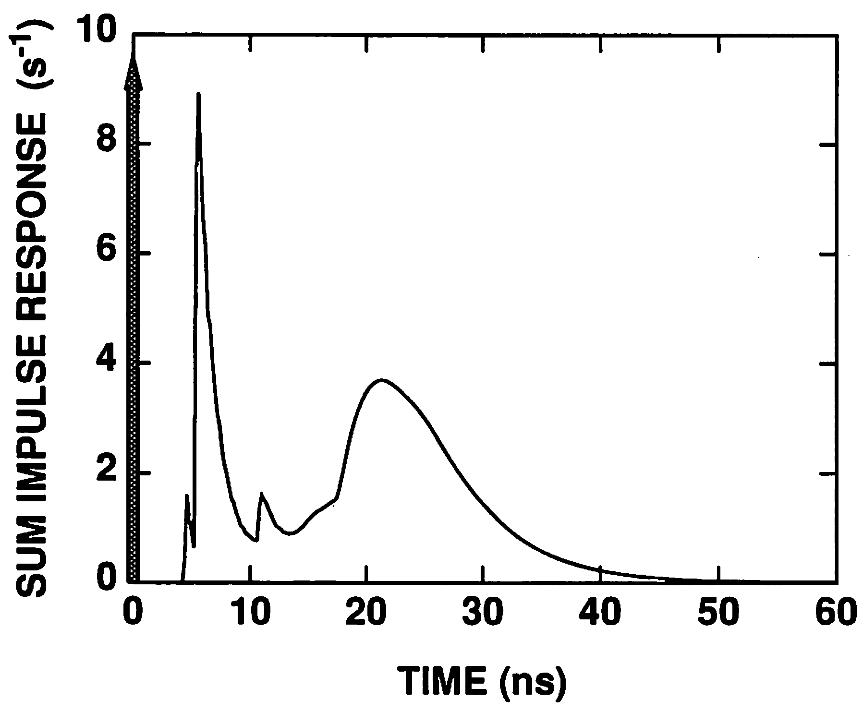
4.4.2.1 Configuration B

Consider the scenario in the second column of Table 4-1, entitled configuration B. Like configuration A, this is also a LOS system, but here the room is somewhat larger and its walls have lower reflectivities. The transmitter is mounted near the ceiling in the south-west corner, pointing straight down, with a mode of $n=1$ (ideal Lambertian). The receiver is at the north end of the room pointing straight up. In Fig. 4-6 we show simulation results using the time and spatial resolution specified in Table 4-1. In Fig. 4-6-a we show separately each of the k-bounce pulses for $k \in \{0,1,2,3\}$. The LOS impulse arrives at time zero and carries 78% percent of the total power, which is 307 nW for a 1-W source. It is interesting to note that the second-order bounce carries about 3.4 dB more power than the first-order bounce. This is likely due to the fact that it takes two bounces for light to reflect from the highly reflective ceiling, which fills the receiver's field of view. The sum of the pulses in Fig. 4-6-a yields the total impulse response of Fig. 4-6-b. The total power of 307 nW is about 9 dB less than the received power for configuration A, the primary reason being that, here, the distance between transmitter and receiver is larger. This illustrates the strict limits on coverage area imposed by path loss.

Experimental measurements were also performed using the configuration B parameters of Table 4-1. In Fig. 4-6-c we illustrate the magnitude response of the experimental channel (solid curve) and of the simulated channel (dashed curve). The experimental result and the simulation result are seen to agree qualitatively; both exhibit high d.c. gain and oscillatory high-frequency components. This structure is in fact common for all multipath channels consisting of a Dirac delta function plus a low-pass impulse response "tail." The experimental and simulation results do not coincide precisely, however. Two reasons stand out as likely candidates. First, we considered only reflections up to third order in our simulations; as illustrated in Fig. 4-4-a, higher-order reflections tend to emphasize the low-frequency components and de-emphasize the "notch" frequencies, which would make the simulation result look more like the experimental one. Second, our simulations were

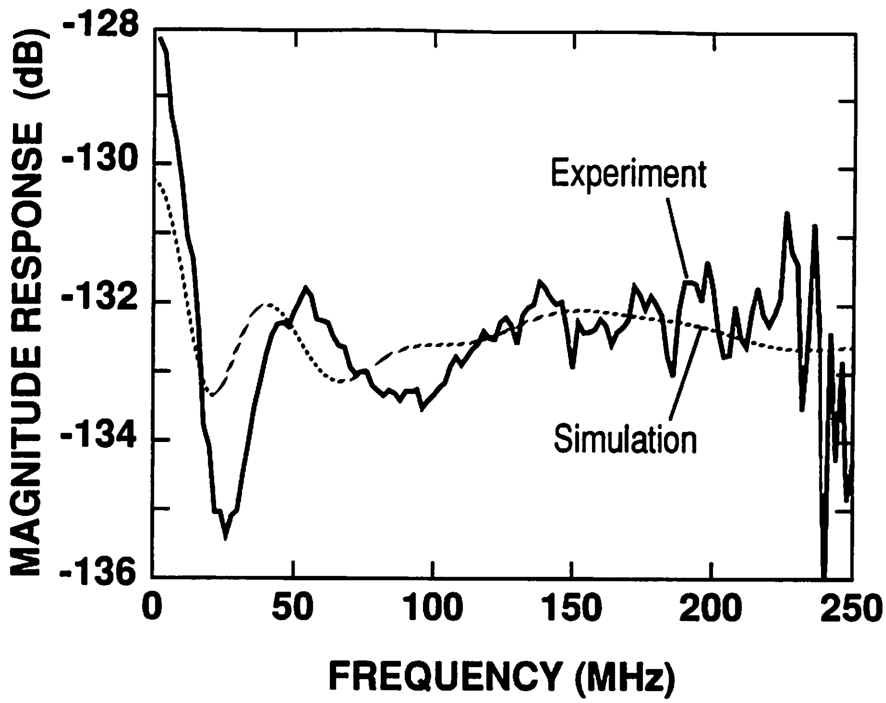


(a)

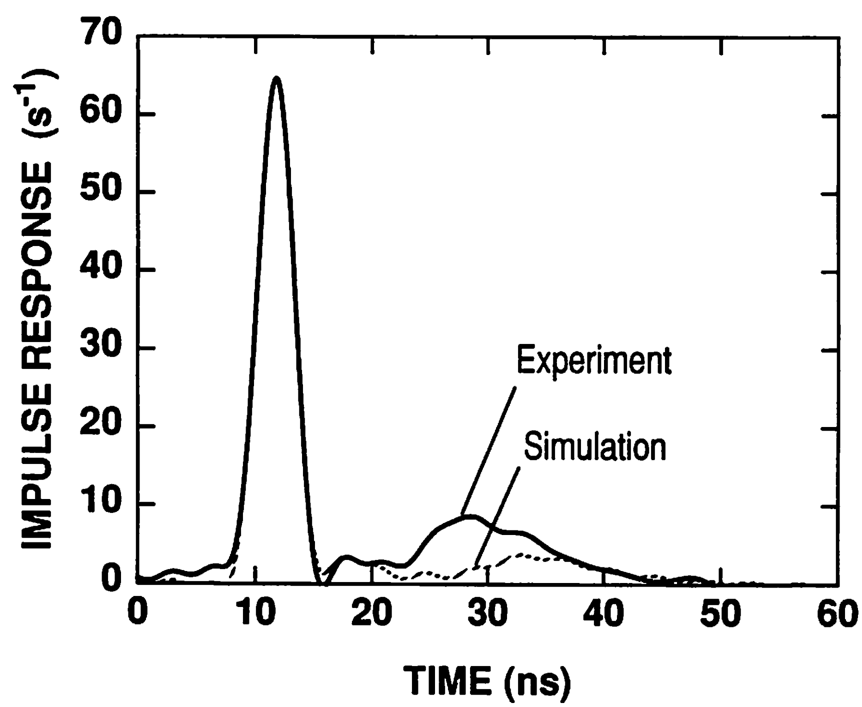


(b)

Fig. 4-6. Configuration B results: (a) separate *k*-bounce pulses; (b) total impulse response. (Continued next page).



(c)



(d)

Fig. 4-6, continued: (c) experimental magnitude response and comparison with simulation; (d) experimental impulse response and comparison with simulation.

based on idealized models of the reflectors, so any discrepancies between simulation and experiment may be attributable to non-ideal or specular reflectors. Other possible reasons for discrepancy include angle-dependent reflections at the air-detector interface and a non-ideal transmitter radiation pattern. The -3 dB bandwidth of the experimental channel is 14 MHz, which is close to the 19.5 MHz predicted by simulation. The bandwidth results for configurations A through D are summarized in Table 4-2.

A time-domain representation of the experimentally measured frequency response can be obtained by performing an inverse Fourier transform. To minimize the effects of the noise, which from Fig. 4-6-a is seen to dominate at high frequencies, we used a 250-MHz Hamming window before performing the inverse transform. The solid curve in Fig. 4-6-d shows the resulting time-domain signal. The dashed curve was obtained from the frequency response of the simulated impulse response in the same manner, i.e, by windowing and inverse transforming. This process removes much of the structure of the original impulse response of Fig. 4-6-b, but facilitates comparison between simulation and experiment. The time-domain signals are seen to agree reasonably well. The full width at half-maximum of the main pulse is about 4 ns, which is the minimum resolution offered by the 250-MHz window. The experimental impulse response exhibits a larger multipath tail than the simulated impulse response.

TABLE 4-2: Summary of Bandwidth and Power Results.

	Config. A (LOS)	Config. B (LOS)	Config. C (LOS)	Config. D (Diffuse)
-3 dB Bandwidth (Simulation)	9 MHz	19.5 MHz	13 MHz	32 MHz
-3 dB Bandwidth (Experiment)	—	14 MHz	12 MHz	34 MHz
Total Received Power (Simulation)	2.4 μW	0.31 μW	0.28 μW	0.69 μW

It is no accident that the experimental curve achieves precisely the same maximum value as the simulation curve in Fig. 4-6-d; in fact, this was how we calibrated the experimental data. The underlying assumption is that our simulation model for LOS path loss matches that of the experiment. This is a reasonable assumption given that the LOS path loss is governed by the geometry of the configuration only, which, unlike non-ideal reflectors, is easily specified.

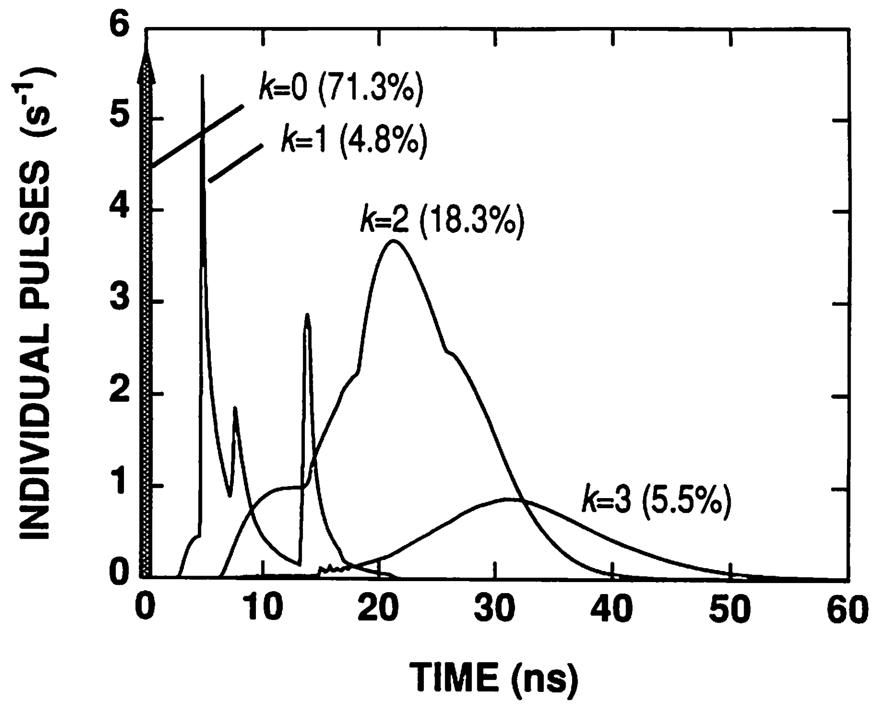
4.4.2.2 Configuration C

Configuration C, like configurations A and B, is a LOS system with the transmitter mounted on the ceiling. Here, however, the acoustic tiles on the north wall was covered by a highly reflective white projection screen, and the transmitter was pointed not straight down but an elevation of -70° and an azimuth of 10°.

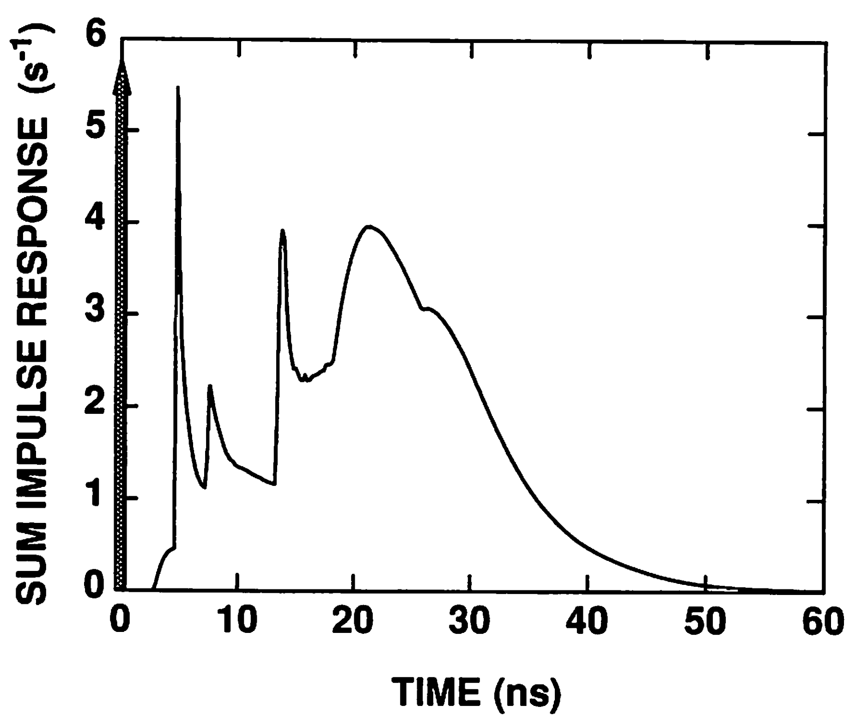
In Fig. 4-7-a we show each of the k-bounce pulses for $k \in \{0,1,2,3\}$ as predicted by simulation. Note again that, as in configuration B, the second-order bounce carries more power than the first-order bounce, this time by about 5.8 dB. The pulses in Fig. 4-7-a combine to yield the total impulse response of Fig. 4-7-b. The total received power for a 1-W source is 283.7 nW, about 0.3 dB less than that for configuration B (see Table 4-2).

The experimentally measured frequency response for configuration C is compared to the simulated response in Fig. 4-7-c. The two curves have the same general shape, but again the simulations seem to underestimate the low-frequency components of the channel, perhaps because only reflections up to third order were considered. The experimental -3 dB bandwidth of 12 MHz is in good agreement with the simulation bandwidth of 13.4 MHz.

The time-domain comparison between simulation and experiment is presented in Fig. 4-7-d, using a procedure identical to that of Fig. 4-6-d. As before, the experimental impulse response exhibits a larger multipath contribution than the simulated impulse response.

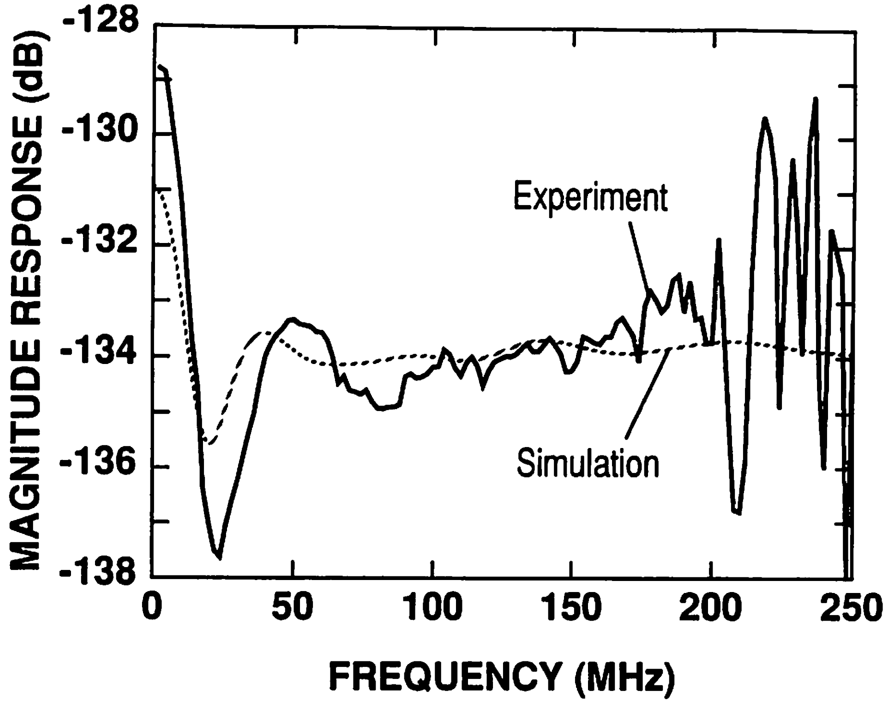


(a)

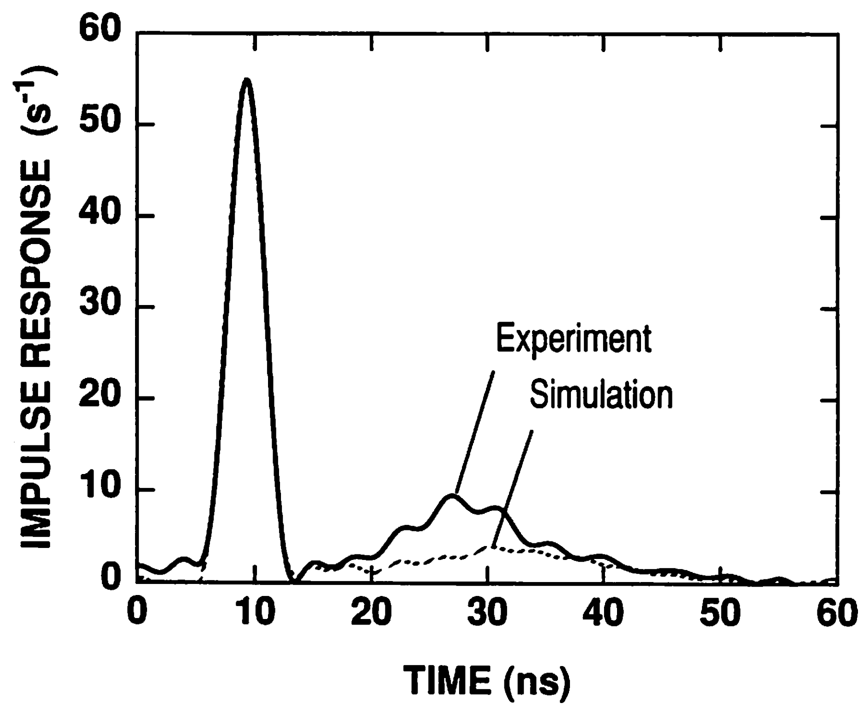


(b)

Fig. 4-7. Configuration C results: (a) separate *k*-bounce pulses; (b) total impulse response. (Continued next page).



(c)



(d)

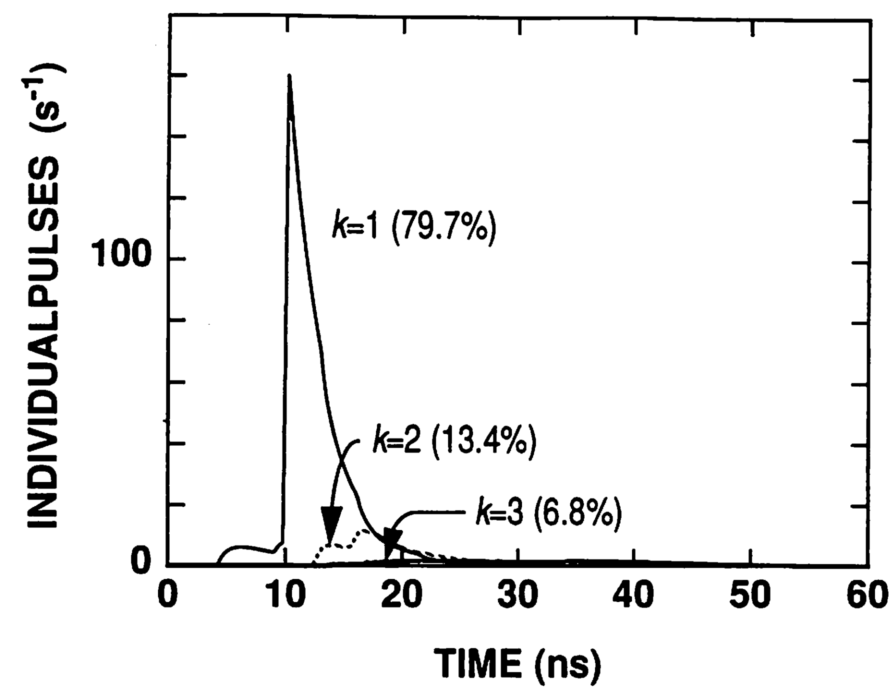
Fig. 4-7, continued: (c) experimental magnitude response and comparison with simulation; (d) experimental impulse response and comparison with simulation.

4.4.2.3 Configuration D

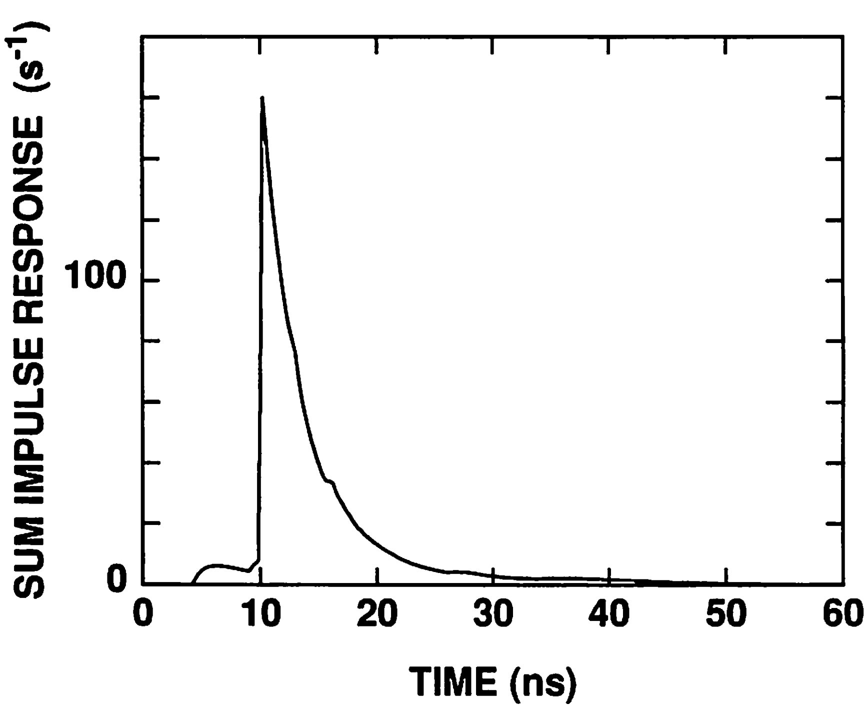
In contrast to the configurations considered so far, configuration D of the fourth column of Table 4-1 represents a diffuse system, with the transmitter in the center of the room near the floor and aimed towards the ceiling; see Fig. 1-1-f. This is similar to the original configuration first proposed by Gfeller over a decade ago [17]. The primary advantage of the diffuse approach is its inherent robustness to shadowing. By illuminating the ceiling with a broad optical beam, the entire ceiling becomes an effective distributed source, making it difficult for an inadvertent obstruction to cast a sharp shadow onto the receiver.

In Fig. 4-8-a we show each of the k-bounce impulse responses, $k \in \{1,2,3\}$, as predicted by simulation for configuration D. There is no LOS contribution, so the $k = 0$ pulse is identically zero. The combined impulse response is shown in Fig. 4-8-b; it corresponds to a total received power of 689.8 nW for a 1-W source. Comparing the received powers for configurations B, C, and D (see Table 4-2), we see that the diffuse system actually provides more power than the LOS systems, despite the lack of a LOS between transmitter and receiver. In fact, this comparison is not completely fair, because the nearly Lambertian transmitter radiation pattern is suboptimal as a transmitter for a LOS system, but it is close to optimal as a transmitter for a diffuse system (see chapter 3). Nevertheless, the relatively high power provided by the diffuse system combined with its robustness to shadowing makes it an attractive candidate for system design.

The experimental and simulation frequency responses for configuration D are compared in Fig. 4-8-c. Because there is no LOS contribution, the LOS-based calibration procedure discussed in the previous section cannot be used directly. Instead, we equated the maximum values of the time-domain signals, which in this case are due to first-order reflections. This has been arbitrarily done for all curves to facilitate comparison of simulation and experiment. This procedure, although justified for the LOS configurations B and

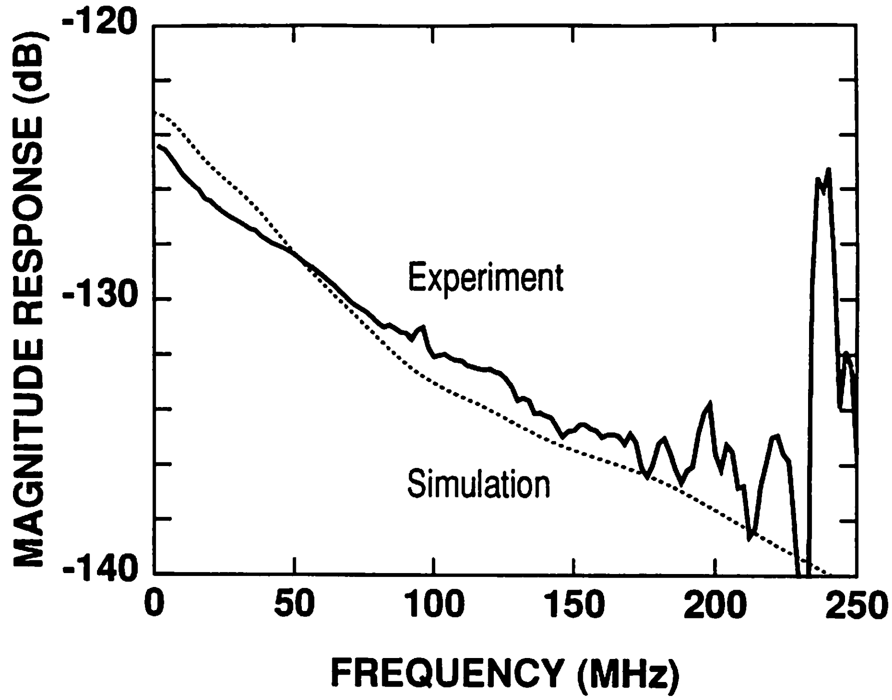


(a)

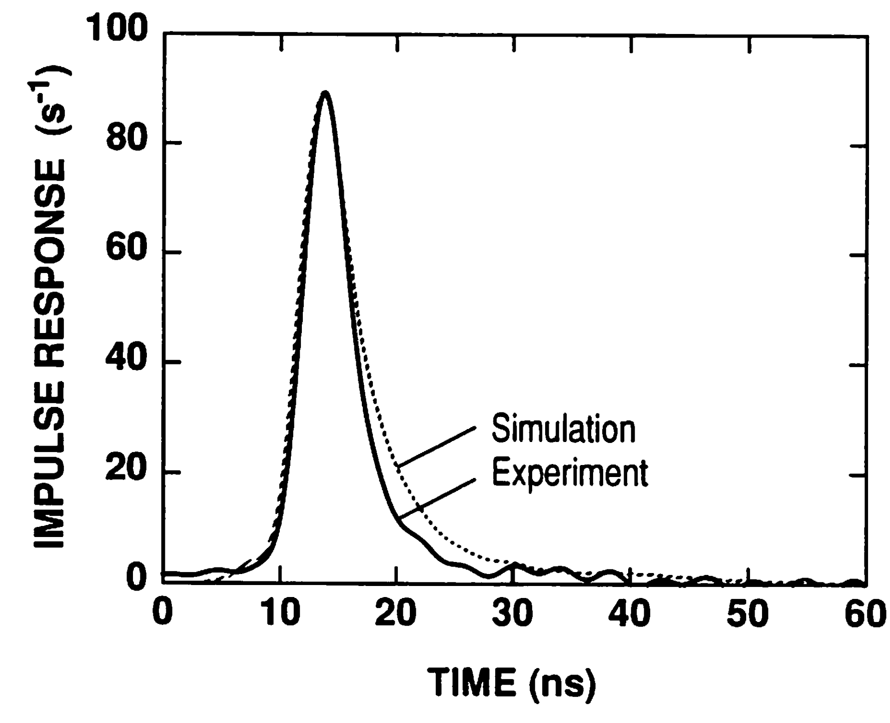


(b)

Fig. 4-8. Configuration D results: (a) separate *k*-bounce pulses; (b) total impulse response. (Continued next page).



(c)



(d)

Fig. 4-8, continued: (c) experimental magnitude response and comparison with simulation; (d) experimental impulse response and comparison with simulation.

C, is less so for configuration D, because it is based on the assumption that the first-order reflector, the ceiling, is ideal Lambertian.

The agreement between simulation and experiment in Fig. 4-8 is good, probably because of the dominant role played by the first-order reflection as shown in Fig. 4-8-a. We found the higher-order reflections play little role in determining the shape of the channel frequency response, so that the simulation curve of Fig. 4-8-c does not change appreciably when only one bounce is considered.

The experimental -3 dB bandwidth of 34 MHz is in good agreement with the simulation bandwidth of 31.8 MHz. As compared in Table 4-2, this bandwidth is over twice the bandwidth of each of the LOS systems. The -3 dB bandwidth is not a fair metric for comparison, however, for the frequency responses of the LOS systems (see Fig. 4-4-a, Fig. 4-6-c, and Fig. 4-7-c) exhibit a narrow peak near d.c. but never roll off significantly even at high frequencies. This characteristic is common to all impulse responses of the form:

$$h(t) = h_0 \delta(t) + p(t), \tag{4-16}$$

with frequency response $H(\omega) = h_0 + P(\omega)$, where $P(\omega)$ is a low-pass response; as frequency approaches infinity, $P(\omega)$ becomes negligible, and the frequency response approaches an asymptote of h_0. The diffuse system, on the other hand, has no LOS Dirac impulse, and so its frequency response rolls off steadily at high frequencies.

The experimental time-domain impulse response is compared to simulation in Fig. 4-8-d, using the same procedure as that for Fig. 4-6-d. There, we see that the pulse is wider than the 4-ns resolution of the 250-MHz window, indicating a broad underlying pulse.

4.5 MULTIPATH-INDUCED POWER PENALTY

To illustrate the adverse effects of multipath dispersion on system performance, consider the model for a baseband on-off keyed (OOK) system shown in Fig. 4-9. The symbols $a_k \in \{0,1\}$ are passed through a transmit filter with impulse response $Ab(t)$ at a bit rate of $1/T$, where A is proportional to the average optical intensity of the transmitter. The output of the transmit filter, which represents the intensity of the transmitted signal, is passed through the multipath channel with impulse response $h(t)$. We assume in this section that $h(t)$ is normalized to have unity area, so that $H(0) = 1$. With this assumption, A becomes the average received optical power. The additive noise $n(t)$ represents the shot noise due to ambient light and is accurately modeled as a Gaussian random process [57]. The received signal plus noise is passed through a receive filter with impulse response $g(t)$, sampled at the bit rate, and quantized to yield the estimate $\hat{a}_k$ of the k-th transmitted bit. We assume that both $b(t)$ and $g(t)$ are normalized to have area T, and that $g(t)$ is a Nyquist pulse [57].

The bit-error rate for this system can be calculated as follows. First, we note that the y_k, the input to the decision device, can be expressed as:

$$y_k = Aa_k \otimes h_k + n_k \tag{4-17}$$

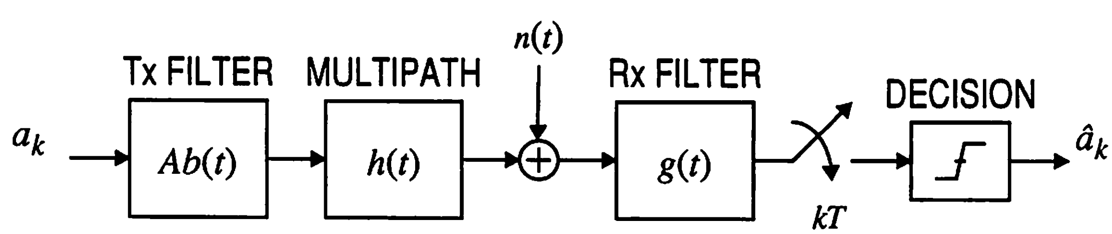


Fig. 4-9. Baseband OOK system.

where the symbol $\otimes$ denotes convolution, h_k is the equivalent discrete-time impulse response of the system, given by:

$$h_k = b(t) \otimes h(t) \otimes g(t)\,|_{t = kT}\,, \tag{4-18}$$

and n_k is given by:

$$n_k = n(t) \otimes g(t)\,|_{t = kT}\,. \tag{4-19}$$

To isolate the power penalty due to intersymbol interference (ISI), we make two assumptions. First, we assume perfect timing recovery; in other words, we assume that the time origin is shifted so as to maximize h_0. Second, we assume an optimal decision threshold. Basic symmetry arguments lead to the conclusion that this optimal threshold is $A/2$.

We can rewrite h_k as $h_k = h_0\delta_k + (1-\delta_k)h_k$, where the second term represents the impulse response "tail," and δ_k is a unit impulse. In general, there will be both precursor and postcursor ISI, so the impulse response tail is not necessarily causal.

For a given sequence of bits $\mathbf{a}_k \equiv (\ldots\, a_{k\text{-}1},\, a_k,\, a_{k+1}\, \ldots)$, the probability that the k-th bit estimate $\hat{a}_k$ is in error is given by:

$$\Pr[\text{error} \mid \mathbf{a}_k] = Q(\rho(1-X_k)) \tag{4-20}$$

where Q is the Gaussian Q-function [57], $\rho \equiv A/(2\sigma)$, σ^2 is the variance of n_k, and X_k represents the ISI:

$$X_k = 2\sum_{i \neq k} a_i h_{k-i}\,. \tag{4-21}$$

The total bit-error rate can be found by averaging over all possible bit sequences:

$$BER = \mathrm{E}[Q(\rho(1-X_k))] \tag{4-22}$$

$$= \frac{1}{2^M}\sum_{\mathbf{a}} Q(\rho(1-X_k)) \tag{4-23}$$

where the expectation is taken over $\mathbf{a}_k$, the elements of which we assume are independent and uniform on $\{0,1\}$. Here, M is the length of the impulse response tail $(1-\delta_k)h_k$, assuming it is finite, and the summation is performed over all $\mathbf{a} \in \{0,1\}^M$.

If there was no multipath dispersion, then X_k would be identically zero and BER would reduce to $Q(\rho)$. The value of ρ required to achieve a desired BER of BER_0 would then be $\rho_0 = Q^{-1}(\mathrm{BER}_0)$. With dispersion, however, a larger value of ρ is required to achieve BER_0. We thus define an optical power penalty as the increased optical signal power required to overcome the multipath ISI and achieve a given BER_0:

$$\text{power penalty} = 10 \cdot \log_{10}\left(\frac{\rho \text{ required for } BER_0}{Q^{-1}(BER_0)} \right) \text{ dB.} \tag{4-24}$$

We emphasize that this is an optical power penalty; the electrical power penalty, in dB, will larger by a factor of two.

In Fig. 4-10 we plot the optical power penalty versus bit rate $1/T$ for each of the four configurations of Table 4-1 using the $K = 3$ simulated impulse responses of Fig. 4-3-b, Fig. 4-6-b, Fig. 4-7-b, and Fig. 4-8-b. We assume a simple OOK system with an integrate-and-dump receiver, so that the transmitter and receiver filters are identical rectangular pulses: $b(t) = g(t) = \mathrm{rect}(2t/T - 1)$, where $\mathrm{rect}(t)$ is defined by (4-9). This receive filter, being matched to the transmit filter, is optimal only when there is no multipath. The results of Fig. 4-10 thus illustrate the performance when multipath is ignored in the system design; equalization can improve performance.

We see that, in all cases, the power penalties are significant for bit rates above 10 Mb/s. Configurations A, B, and C, which are all LOS systems, are seen to be less susceptible to multipath interference. Configuration A has a higher penalty than B or C for two reasons; first, the walls in configuration A are highly reflective, and second, the receiver is on the floor and so a large fraction of the area of the walls are within its field of view. The power penalty for the diffuse system, labeled D in the figure, eventually grows much faster with increasing bit rate than any of the LOS systems. Interestingly, however, at moderate bit rates (below 100 Mb/s) the power penalty's rate of growth is less than the LOS systems. This is due to the relatively low signal energy carried by second- and third-order reflections for configuration D, as compared to the LOS systems.

For bit rates above 100 Mb/s, the power penalties for two of the LOS systems, B and C, are seen to *decrease* as the bit rate increases. This phenomenon is due to the LOS Dirac delta function in the impulse response, and hence does not occur for diffuse systems. We now introduce a second procedure for calculating the optical power penalty based upon a Gaussian approximation, which is useful in explaining the power penalty behavior at high bit rates.

The ISI term X_k as defined in (4-21) is a random variable with mean:

$$\mu = E[2\sum_{i \neq k} a_i h_{k-i}] = \sum_{i \neq 0} h_i = 1 - h_0 \tag{4-25}$$

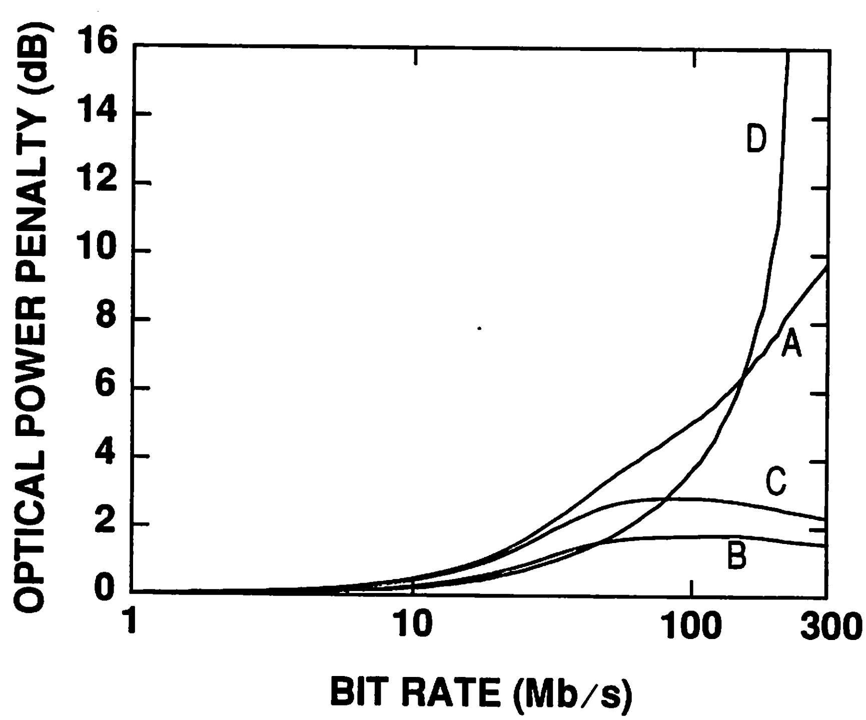


Fig. 4-10. Optical power penalty versus bit rate for configurations A through D, accounting for up to $K = 3$ bounces ($BER_0 = 10^{-6}$).

and variance:

$$E_{tail} = \mathrm{var}[2\sum_{i \neq k} a_i h_{k-i}] = \sum_{i \neq 0} h_i^2 \ . \tag{4-26}$$

Note that E_{tail} is just the energy contained in the impulse response tail. In (4-25) we make use of the assumption that $\sum h_k = 1$.

As the bit rate $1/T$ approaches infinity, the length of the discrete-time impulse response h_k also approaches infinity. Therefore, because $\{a_i\}$ are independent, the central limit theorem tells us that X_k tends towards a Gaussian random variable, with mean μ and variance E_{tail}. With this Gaussian assumption, and rewriting the expectation of (4-22) as an integral with a Gaussian density function, we find that the BER reduces to:

$$BER = Q\left(\frac{\rho h_0}{\sqrt{1+\rho^2 E_{tail}}}\right) . \tag{4-27}$$

Equating the argument of the Q-function here with ρ_0 and solving for ρ/ρ_0, the optical power penalty of (4-24) reduces to the following under the Gaussian assumption:

$$\text{power penalty} = -5 \cdot \log_{10}(h_0^2 - \rho_0^2 E_{tail}) \ \text{ dB}, \tag{4-28}$$

where again $\rho_0 = Q^{-1}(BER_0)$.

In Fig. 4-11 we compare this approximation with the true power penalty for configuration B. The curves are labeled by the maximum number K of reflections considered. The approximate curves, shown with dashed lines, exhibit a more pronounced maximum near 40 Mb/s than do the actual curves. Since configuration B is a LOS system, h_0 will approach asymptotically a nonzero constant as $1/T$ approaches infinity. The energy in the tail, on the other hand, is asymptotically zero when $1/T$ approaches both zero and infinity, and achieves a maximum somewhere in between. The frequency at which the tail energy is maximum is close to the frequency at which the power penalty achieves its maximum.

(The maxima may not coincide exactly when the higher-order reflections arrive soon after the LOS impulse.)

The Gaussian approximation is seen to be inaccurate at moderate frequencies near 40 Mb/s, the reason being that, at this bit rate, the length of the impulse response tail is not sufficient for application of the central limit theorem. As the bit rate increases above 100 Mb/s, however, the Gaussian approximation converges to the actual power penalty. From Fig. 4-6-a we see that $h_0 = 0.779$ for the $K = 3$ impulse response, and so from (4-28) with $E_{tail} = 0$ we calculate the high-bit-rate asymptote of the $K = 3$ curve in Fig. 4-11 to be 1.1 dB.

The three cases $K \in \{1,2,3\}$ are shown in Fig. 4-11 to illustrate the importance of the higher-order reflections on system performance. The curve labeled $K = 1$ accounts for only first-order reflections and is seen to grossly underestimate the true power penalty. The

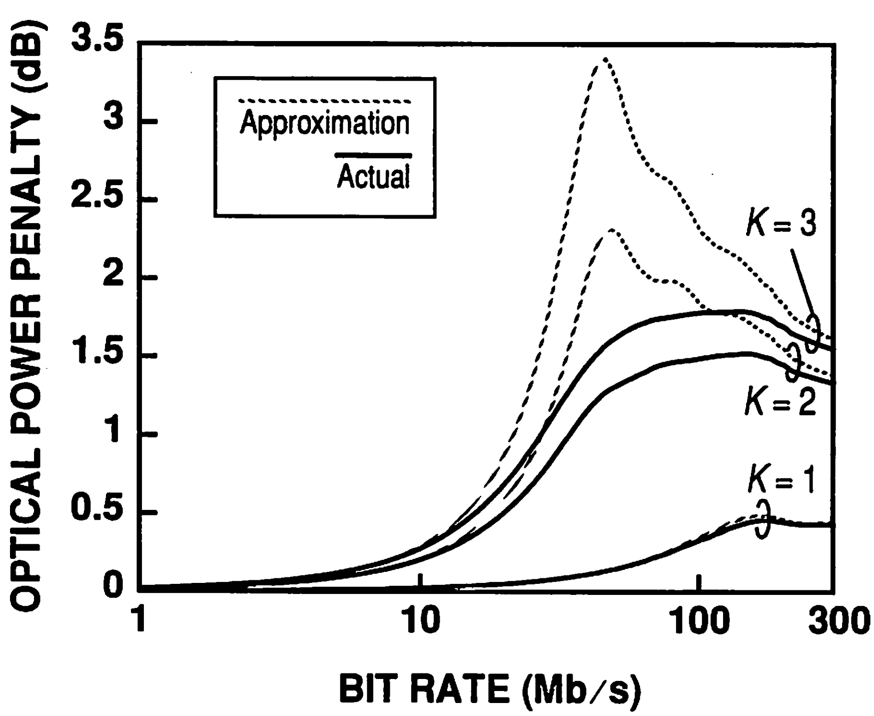


Fig. 4-11. Comparison between Gaussian approximation and actual power penalty for configuration B ($BER_0 = 10^{-6}$).

$K = 2$ curve is more accurate, but still underestimates the power penalty by as much as 0.4 dB. The curve labeled $K = 3$ is identical to the curved labeled B in Fig. 4-10. Reflections of order greater than 3 will further increase the power penalty, although to a lesser extent. We thus conclude that, since most of the power penalty is due to reflections of order greater than one, the high-order reflections are the dominant source of intersymbol interference for configuration B.

4.6 SUMMARY

We have presented a method for evaluating the impulse response of an arbitrary room with Lambertian reflectors. This method can account for any number of reflected paths. A simple algorithm suitable for computer implementation has been presented. The results of computer simulations indicate that reflections of multiple order are a significant source of intersymbol interference for an indoor optical communication system. Our simulations were verified by experimental measurements. The design of a high-speed indoor communication link using infrared requires careful attention to the multipath response described in this chapter.

Our experimental results are applicable only to the particular room configurations specified in Table 4-1, so we cannot make general statements about all room configurations. Future work in channel characterization should concentrate on filling this gap, in particular by examining the effects of irregularly shaped rooms, furniture, non-Lambertian and specular reflectors, and shadowing. From a systems-design standpoint, a statistical characterization of the channel under various conditions would be useful.

CHAPTER 5

MODULATION

In this chapter we define an intensity-modulation channel, which differs from a conventional linear Gaussian-noise channel in two ways: the input cannot be negative, and the input is *amplitude* limited, not power limited. We then compare the power efficiency and bandwidth efficiency of a number of modulation schemes for the intensity-modulation channel. We find that the intensity-modulation channel favors baseband modulation schemes over subcarrier and multiple-subcarrier modulation schemes, particularly those baseband modulation schemes with low duty cycles like pulse-position modulation.

5.1 INTENSITY MODULATION AND DIRECT DETECTION

The simplest way to convey information on an optical carrier is through *intensity modulation*, whereby the intensity of the transmitted optical lightwave is modulated according to a modulating electrical signal $X(t)$, as shown in Fig. 5-1-a. Intensity modulation is easily obtained by varying the bias current of a laser diode or LED. A *direct detection* receiver uses a photodetector, which produces an electrical output $Y(t)$ proportional to the intensity of the detected lightwave. The composite channel from modulating input $X(t)$ to photodetector output $Y(t)$ can be viewed as an equivalent baseband channel, as shown in Fig. 5-1-b. This channel can be characterized by a multipath impulse response $h(t)$ and a noise $n(t)$, so that:

$$Y(t) = X(t) \otimes h(t) + n(t) , \tag{5-1}$$

where the symbol $\otimes$ denotes convolution. The multipath impulse response was discussed at length in chapter 4. As discussed in chapter 2 and chapter 3, the noise $n(t)$ consists of three components: a white Gaussian component from the shot noise due to d.c. background light, an f^2 component due to channel noise in the front-end FET, and a low-fre-

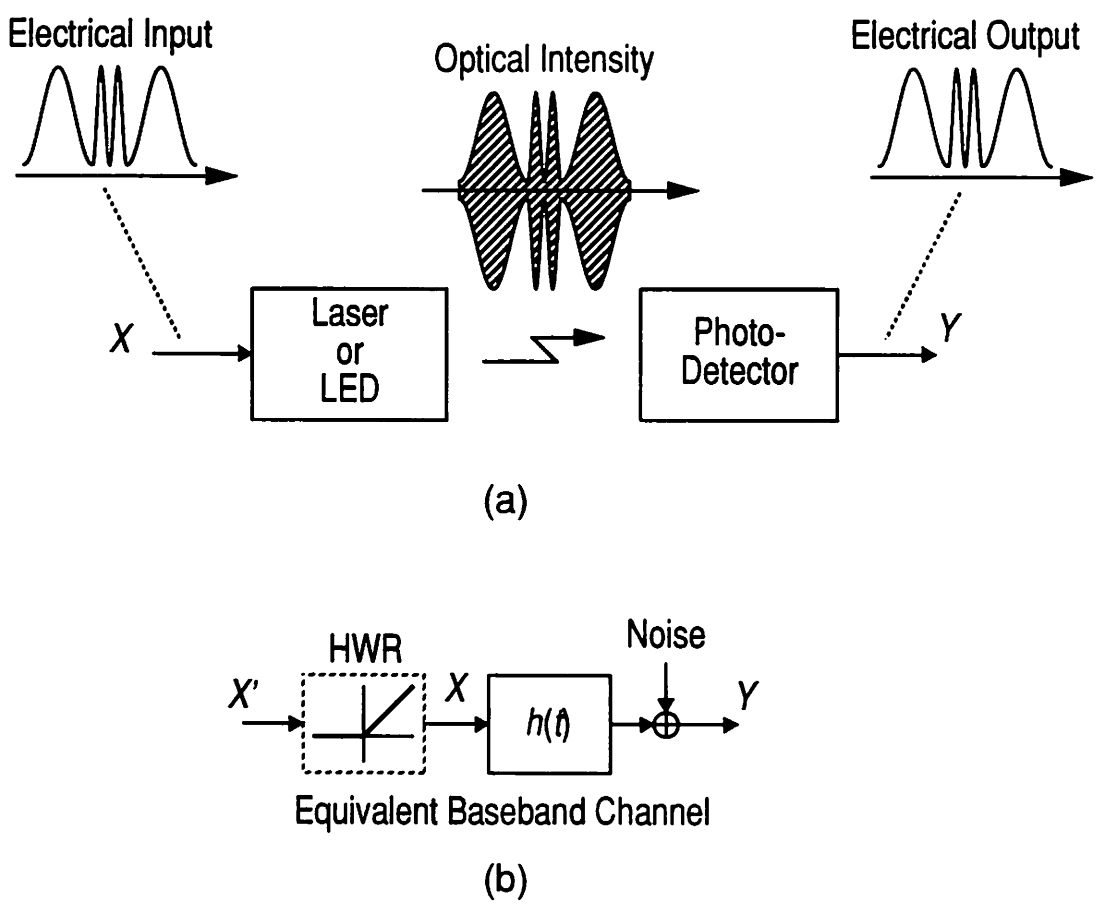


Fig. 5-1. Intensity modulation and its baseband model. The half-wave rectifier (HWR) constrains $X(t)$ to be positive.

quency cyclostationary component due to fluorescent light switching. In this chapter we will ignore the fluorescent-light interference; it does not affect passband systems, and it can be mitigated for baseband systems by combining a line code with a high-pass filter or by using the balanced differential receiver front end proposed in section 3.5.4. We also ignore f^2 noise in this chapter, because it can be made negligible by proper choice of the transconductance of the preamplifier FET (see section 3.3).

The equivalent channel of Fig. 5-1-b is referred to as *baseband* because it is centered at d.c. and not at the carrier frequency of the optical signal (about 3×10^{14} Hz). Nevertheless, the bandwidth of this channel, which is limited primarily by the bandwidth of the receiver electronics, can be quite large (in the hundreds of MHz).

Without loss of generality, this chapter assumes that the proportionality constant between the modulating electrical signal $X(t)$ and the optical intensity of the transmitted lightwave is unity, as is the proportionality constant between the intensity of the received lightwave and the photodetector output $Y(t)$. Under this assumption, $X(t)$ and $Y(t)$ become the transmitted and received optical intensities, respectively.

Although the equivalent baseband channel of Fig. 5-1-b is similar to a conventional linear Gaussian-noise channel, it differs in two important respects:

- The input $X(t)$ must be positive everywhere.
- The average amplitude of the input $X(t)$ is limited. (5-2)

These differences prevent us from blindly applying modulation analysis for the traditional channel, for which there is a wealth of literature, to our application. Instead, we must re-evaluate each candidate modulation scheme anew.

The first constraint in (5-2) follows from the fact that the intensity of the transmitted optical signal cannot be negative. This constraint is represented in the baseband model of Fig. 5-1-b by viewing $X(t)$ as the output of a half-wave rectifier. The second constraint deserves further clarification. The input signal in a traditional communication problem is generally constrained to obey an average power limitation P_X, so that:

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X^2(t)\, dt \;\leq P_X\,. \tag{5-3}$$

In our application, however, the average intensity of the transmitted lightwave is constrained (primarily by eye safety considerations[1]) to a value which we denote P_{avg}, so that:

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X(t)\, dt \;\leq P_{avg}\,. \tag{5-4}$$

As we will see, this difference leads to results that are unique to the intensity-modulation channel. For example, unlike the conventional channel, the intensity-modulation channel can have two candidate transmit pulse shapes with the same energy but different performance.

5.2 BINARY MODULATION

As a precursor to consideration of multilevel signaling, we consider three binary modulation schemes: on-off keying (OOK), two-pulse-position modulation (2-PPM), and subcarrier binary phase-shift keying (BPSK). We examine first their power efficiency, measured by the average optical power required to achieve a given bit-error rate (BER) at a given bit rate, and then their bandwidth efficiency, measured by the electrical bandwidth required to achieve a given bit rate. All three schemes can be analyzed using Fig. 5-2,

---

1. Average intensity, not peak intensity, is the limiting factor for eye hazards when practical signals with a significant duty cycle are used. There are extreme counter-examples, of course, such as a burst of an infinite number of photons during an infinitesimal interval, but such signals are not practical for other reasons. When in doubt, the safeness of a candidate modulation scheme can be ascertained by following the procedures outlined in [58].

which shows three alternative pulse shapes at the transmitter and three alternative matched filters at the receiver.

5.2.1 Power Efficiency

Consider first OOK. Simply stated, $X(t)$ should be "on" during a one bit and "off" during a zero bit. Assuming that ones and zeros are equally likely, an average power limitation of P_{avg} requires that $X(t)$ be $2P_{avg}$ during a one bit.

An equivalent interpretation of OOK is that $X(t)$ consists of a binary pulse-amplitude modulation (2-PAM) signal, taking values of $+P_{avg}$ and $-P_{avg}$ during one and zero bits, respectively, plus a d.c. bias of P_{avg}. This trick will reappear repeatedly in this chapter, where we will find it convenient to decompose an intensity-modulated signal into two components: a d.c. component plus a time-varying component that is symmetric about zero. Thus, the top transmit filter and the top matched filter in are used in Fig. 5-2, and the

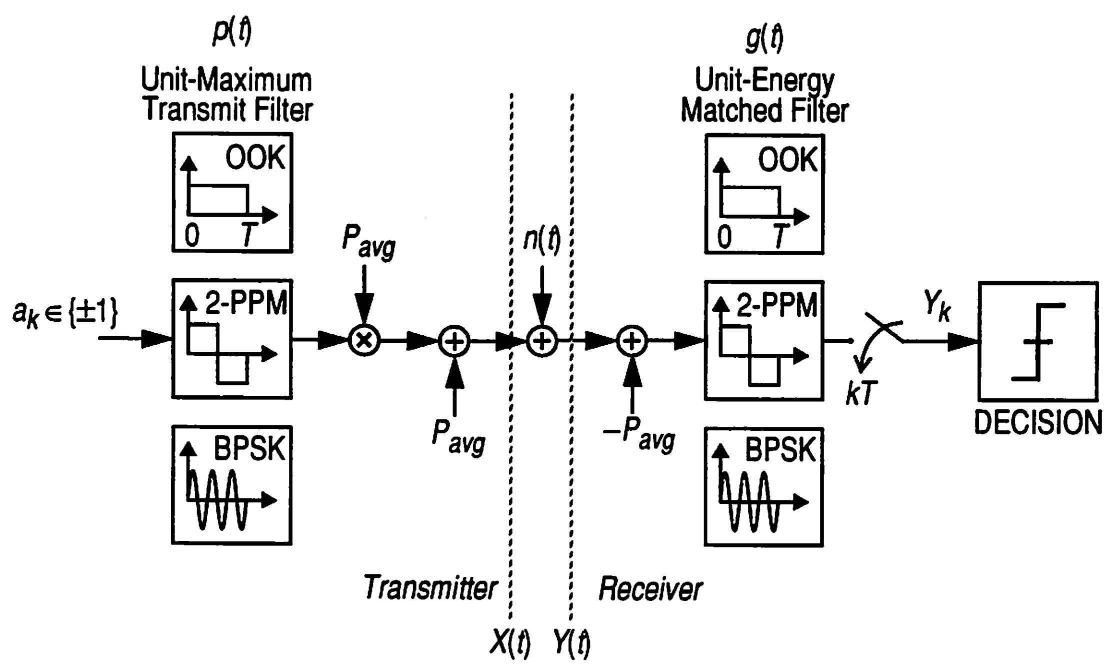


Fig. 5-2. Binary intensity modulation: OOK, 2-PPM, and BPSK.

white discrete-time sequence of symbols $\{a_k\}$ that drives the transmit filter in Fig. 5-2 takes on values from $\{-1, +1\}$.

In this chapter we consistently assume that the transmit pulse shape $p(t)$ is normalized to have a maximum value of unity. The transmit pulse shape for OOK is thus given by:

$$\text{OOK:} \quad p(t) = \begin{cases} 1 & \text{for } t \in [0, T) \\ 0 & \text{elsewhere} \end{cases} . \tag{5-5}$$

In two-pulse-position modulation (2-PPM), a one bit is signified when the optical signal is "on" during the first half of the symbol interval, and a zero bit is signified when the optical signal is "on" during the second half of the symbol interval. As before, we will view 2-PPM as a binary antipodal signal plus a d.c. component. The transmit pulse shape for 2-PPM is given by:

$$\text{2-PPM:} \quad p(t) = \begin{cases} 1 & \text{for } t \in [0, T/2) \\ -1 & \text{for } t \in [T/2, T) \\ 0 & \text{elsewhere} \end{cases} . \tag{5-6}$$

Note that, without the d.c. component, 2-PPM is identical to *biphase* or *Manchester* signaling [57][59].

Similarly, the transmit pulse shape for binary phase-shift keying (BPSK) is given by:

$$\text{BPSK:} \quad p(t) = \begin{cases} \cos(\omega_0 t) & \text{for } t \in [0, T) \\ 0 & \text{elsewhere} \end{cases} , \tag{5-7}$$

where $1/T$ is the symbol rate and ω_0 is the subcarrier frequency.

As shown in Fig. 5-2, the output of the transmitter filter is scaled by a factor P_{avg}, so that the resulting signal takes values in the range $[-P_{avg}, P_{avg}]$. A d.c. bias of P_{avg} is then added to ensure that the result is everywhere positive. The result is $X(t)$, the intensity of the transmitter output:

$$X(t) = P_{avg} + P_{avg} \sum_k a_k p(t - kT) . \tag{5-8}$$

A practical OOK or 2-PPM transmitter need not be implemented in this way, i.e., by adding a d.c. bias to a signal that is symmetric about zero, although a BPSK system must. The BPSK signal as described is modulated fully, so that the intensity achieves a maximum of $2P_{avg}$ and a minimum of zero. To provide a margin against accidental clipping, a practical system may reduce the a.c. amplitude in effect by replacing the second occurrence of P_{avg} in (5-8) with $(P_{avg} - \varepsilon)$. This will reduce the power efficiency. The d.c. offset remains unchanged.

In this section we will ignore path loss and multipath dispersion, so that the optical channel adds noise but otherwise does not distort the intensity of the transmitted signal:

$$Y(t) = X(t) + n(t)\,, \tag{5-9}$$

where $n(t)$ is white Gaussian noise with two-sided power spectral density N_0. Under this assumption, $X(t)$ becomes the intensity of the received, not transmitted, optical signal.

The maximum likelihood receiver, which minimizes the probability of error when ones and zeros are equally likely, is shown in Fig. 5-2 [57]. The receiver first subtracts off the d.c. bias term P_{avg}, which carries no information. The resulting signal is then passed through a receiver filter $g(t)$ that is matched to the transmit filter and has unit energy:

$$g(t) = \frac{1}{\sqrt{E_p}}p(-t) \tag{5-10}$$

where E_p is the energy in the transmit pulse:

$$E_p = \int_{-\infty}^{\infty} p^2(t)\, dt\,. \tag{5-11}$$

The output of the matched filter is sampled yielding a discrete-time sequence $\{Y_k\}$, where:

$$Y_k = a_k P_{avg}\sqrt{E_p} + n_k\,, \tag{5-12}$$

where $\{n_k\}$ is a discrete-time white Gaussian noise with power spectral density (PSD) N_0. The resulting bit-error rate (BER), or probability of a bit error, is given by [57]:

$$BER = Q(P_{avg}\sqrt{E_p/N_0})\,. \tag{5-13}$$

For any modulation scheme, define a power requirement P_{req} as the average optical power required by an ideal system to achieve a given bit rate and BER. Consider first OOK. The energy in the OOK transmit pulse of (5-5) is $E_p = T = 1/R_b$, where R_b is the bit rate, so that BER reduces to:

$$BER = Q\left(\frac{P_{avg}}{\sqrt{N_0 R_b}}\right). \tag{5-14}$$

Therefore, the power requirement for OOK is:

$$P_{req} = \sqrt{N_0 R_b}\,Q^{-1}(BER) \equiv P_{OOK}. \tag{5-15}$$

We will use this parameter as a benchmark to compare the performance of other modulation schemes. In other words, for each new modulation scheme we will ask: For a given noise power N_0 and bit rate R_b, how does P_{req} compare with P_{OOK}?

Note that, when shot noise due to background illumination is the dominant source of noise in the receiver, which is often the case (see chapter 3), the noise power N_0 will be given by:

$$N_0 = \frac{N_{0,\,shot}}{r^2} = \frac{qP_{bg}}{r}\,, \tag{5-16}$$

where P_{bg} is the total detected background power, as given by (3-4), q is the charge of an electron (C), and r is the photodetector responsivity (A/W). For example, from (5-15) with $P_{bg} = 1$ mW, $r = 0.53$ A/W, and $R_b = 100$ Mb/s we find that the average received optical power required by an ideal OOK system to achieve a BER of 10^{-9} is $P_{OOK} = 1.0$ μW, which corresponds to about 35,000 photons/bit.

The power efficiencies of 2-PPM and BPSK follow directly from (5-13). From (5-6) and (5-7), the energy in the 2-PPM transmit pulse is $E_p = T$, while the energy in the BPSK transmit pulse is $E_p = T/2$, so that:

$$P_{req} = \begin{cases} P_{OOK} & \text{for 2-PPM} \\ \sqrt{2}\, P_{OOK} & \text{for BPSK} \end{cases} . \tag{5-17}$$

Thus, 2-PPM has the same sensitivity as OOK, whereas BPSK suffers a 1.5 dB optical power penalty. These results are summarized in Table 5-1, as are the results of later sections.

5.2.2 Bandwidth Efficiency

Of course, power efficiency is not the only measure of performance. As discussed in chapter 3, the high-capacitance of large-area photodiodes makes it difficult to obtain wide

TABLE 5-1: Comparison of Intensity Modulation Techniques.

Modulation Scheme	Average Optical Power Requirement	Optical Power Penalty (Relative to OOK)	Bandwidth Requirement
OOK	P_{OOK} (see (5-15))	0 dB	R_b
2-PPM	P_{OOK}	0 dB	$2R_b$
BPSK	$\sqrt{2}\, P_{OOK}$	1.5 dB	$2R_b$
L-PAM	$\frac{L-1}{\sqrt{\log_2 L}} P_{OOK}$	$5 \cdot \log_{10} \frac{(L-1)^2}{\log_2 L}$ dB	$\frac{R_b}{\log_2 L}$
N-BPSK	$\sqrt{2N} P_{OOK}$	$1.5 + 5 \cdot \log_{10} N$ dB	$2R_b$
N-QPSK	$\sqrt{2N} P_{OOK}$	$1.5 + 5 \cdot \log_{10} N$ dB	R_b
N-L-QAM	$\sqrt{4N} \frac{\sqrt{L}-1}{\sqrt{\log_2 L}} P_{OOK}$	$1.5 + 5 \cdot \log_{10} N + 5 \cdot \log_{10} \frac{(\sqrt{L}-1)^2}{\log_2 \sqrt{L}}$ dB	$\frac{2R_b}{\log_2 L}$
L-PPM	$\frac{1}{\sqrt{\frac{1}{2} L \cdot \log_2 L}} P_{OOK}$	$-5 \cdot \log_{10}(\frac{1}{2} L \cdot \log_2 L)$ dB	$\frac{LR_b}{\log_2 L}$

receiver bandwidths, and so we also view bandwidth efficiency as an important metric. Fig. 5-3 illustrates the power spectra of the transmitted signals for the three binary schemes considered so far. Roughly speaking, the required bandwidth is $B = R_b$ for OOK and $B = 2R_b$ for both 2-PPM and BPSK, where R_b is the bit rate. These values are noted in the last column of Table 5-1. These bandwidths were estimated by the first null in the spectra of Fig. 5-3. They can be reduced by pulse shaping; for example, an ideal zero-excess-bandwidth pulse shape for OOK would require a bandwidth of only $B = R_b/2$. Unless otherwise noted, this chapter assume 100%-excess-bandwidth pulses.

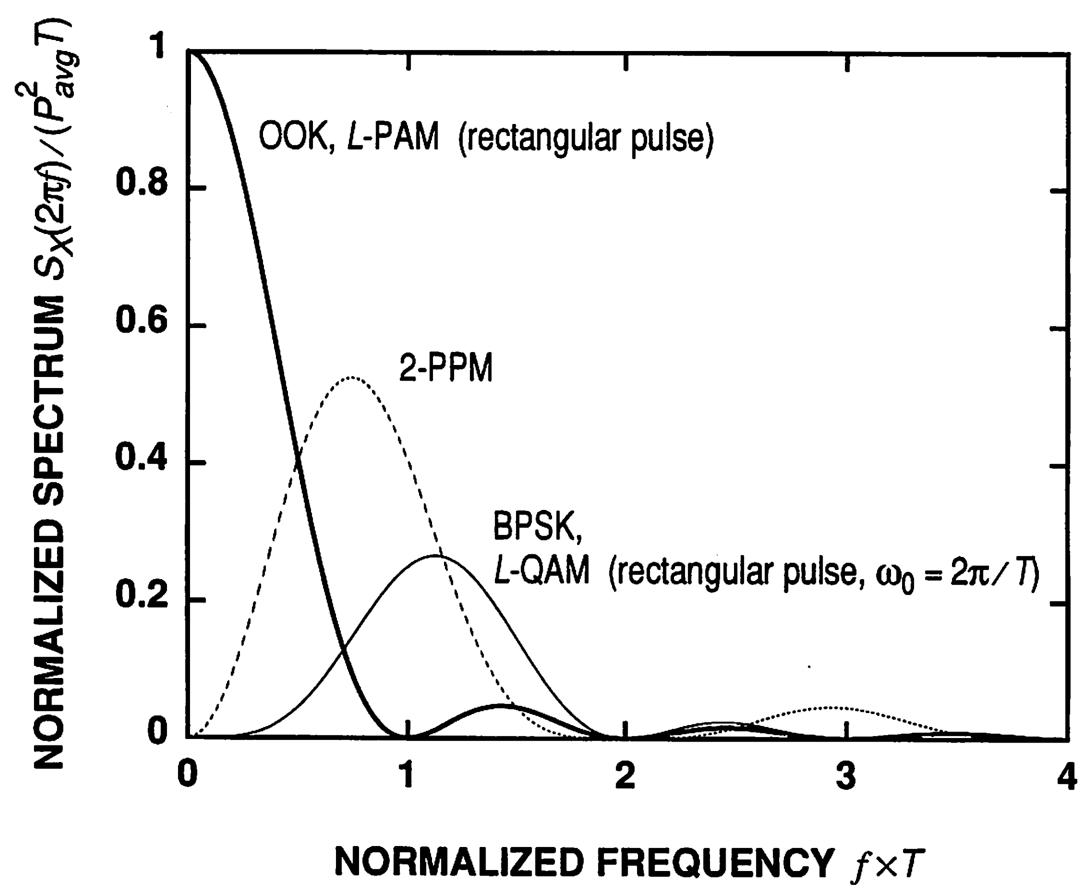


Fig. 5-3. Comparison of power spectra for various modulation schemes; $1/T$ is the symbol rate. (The Dirac impulse at d.c. is not shown.)

We can compare the ability of these three binary modulation schemes to achieve both bandwidth efficiency and power efficiency by plotting power requirement versus bandwidth requirement, as shown in Fig. 5-4. The OOK data point is marked with a circle, the 2-PPM data point is marked with a square, and the BPSK data point is marked with the symbol ×. Other things being equal, the best modulation scheme is the one closest to the lower-left-hand corner. Of the three binary schemes considered so far, OOK is best.

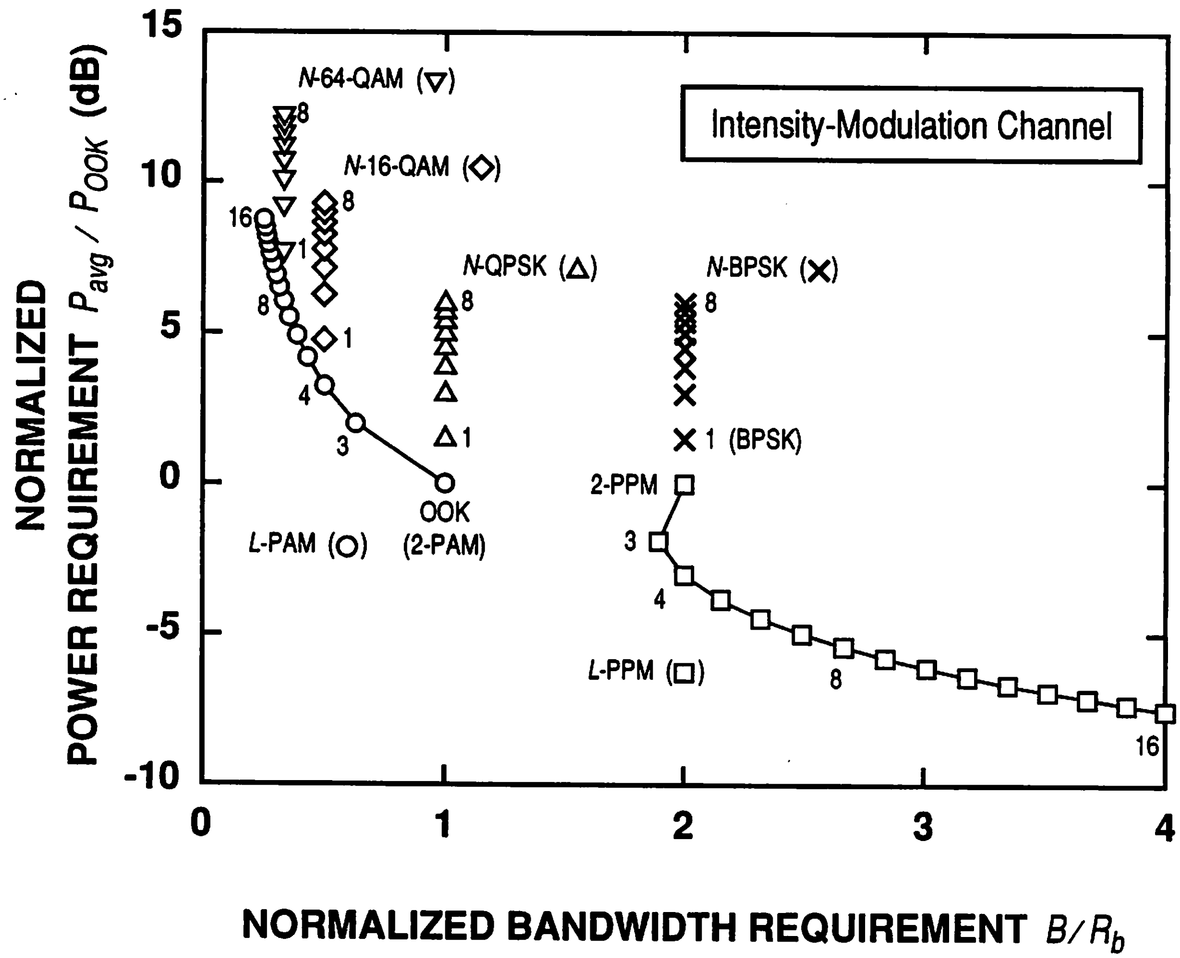


Fig. 5-4. Combined bandwidth efficiency and power efficiency for various modulation schemes on the intensity-modulation channel.

5.2.3 Comparison to the Conventional Channel

Recall the input constraints (5-2) of the intensity-modulation channel: $X(t)$ must be positive, and the average amplitude of $X(t)$ must not exceed a maximum value of P_{avg}. As we examine the performance of various modulation schemes for the intensity-modulation channel, it will be useful to make comparisons with the "conventional" channel, for which the first constraint in (5-2) is omitted, so that negative values of $X(t)$ are allowed, and the second constraint in (5-2) is changed to a power constraint, so that the average *power* of $X(t)$ must not exceed a maximum of P_X. In other words, the conventional channel is an additive-white Gaussian-noise (AWGN) channel with an input power constraint of P_X.

The power efficiencies for a number of modulation schemes on the conventional channel are presented in Appendix A. Just as P_{OOK} is the benchmark for power efficiency on the intensity-modulation channel, $P_{X,2\text{-}PAM}$ is the benchmark for power efficiency on the conventional channel, where $P_{X,2\text{-}PAM}$ is the power (per (5-3)) required by an ideal 2-PAM system to achieve a given BER at a given bit rate on the conventional channel (see (A-9) in Appendix A). The results are presented in Fig. 5-5 by plotting the power requirement versus the bandwidth requirement, making Fig. 5-5 the conventional-channel counterpart to Fig. 5-4. (Note that bandwidth efficiency on the conventional channel is identical to bandwidth efficiency on the intensity-modulation channel.)

Of the three binary modulation schemes considered so far, we see that the 1.5-dB penalty for BPSK relative to OOK (2-PAM) and 2-PPM is unique to the intensity-modulation channel; there is no such penalty for the conventional channel.

Also included in Fig. 5-5 is a curve indicating the Shannon limit of power efficiency and bandwidth efficiency. The capacity of a conventional strictly band-limited channel with bandwidth B is given by the Hartley-Shannon formula [57][60]:

$$C = B \cdot \log_2\left(1 + \frac{P_X}{2N_0 B}\right). \tag{5-18}$$

Solving this equation for P_X yields the following:

$$P_X = \frac{2B/C}{Q^{-1}(BER)^2}(2^{C/B} - 1)\ P_{X,2\text{-}PAM}\,, \tag{5-19}$$

where $P_{X,2\text{-}PAM} = N_0 R_b Q^{-1}(BER)^2$ as defined in Appendix A. Thus, for a given bandwidth B, this expression described how much signal power is needed by an ideal capacity-achieving system to achieve a bit rate of $R_b = C$. The Shannon limit curve was obtained by plotting the ratio $P_X/P_{X,2\text{-}PAM}$ from (5-19) versus B/C, assuming $BER = 10^{-4}$. (Note that, although the ratio $P_X/P_{X,2\text{-}PAM}$ depends on BER, P_X does not.) A similar curve for

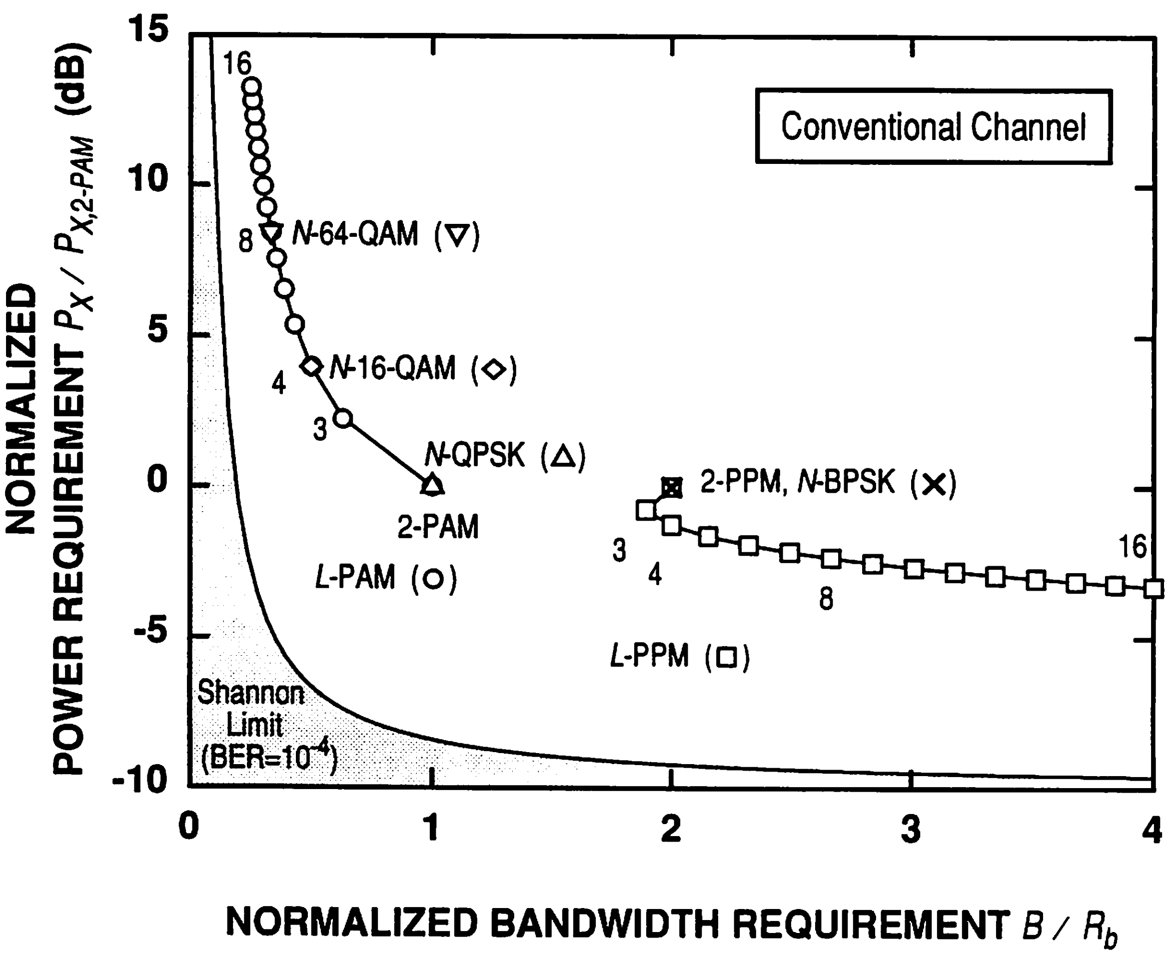


Fig. 5-5. Conventional analog of Fig. 5-4: combined bandwidth and power efficiency for various modulation schemes on the conventional AWGN channel.

the intensity-modulation channel would be useful, but unfortunately the capacity of the intensity-modulation channel is, to the author's knowledge, unknown and thus an open problem.

5.3 MULTI-LEVEL MODULATION

In this section we evaluate the bandwidth efficiency and power efficiency for the following multilevel modulation schemes on the intensity-modulation channel: L-level baseband pulse-amplitude modulation, subcarrier L-level quadrature-amplitude modulation, multiple-subcarrier modulation, and L-level pulse-position modulation.

5.3.1 Baseband Pulse-Amplitude Modulation

We consider first L-level baseband pulse-amplitude modulation (L-PAM), in which one of L possible levels is transmitted each symbol interval. An L-PAM system can be analyzed using the same block diagram as for the binary systems (see Fig. 5-2), but with a symbol alphabet of size L. To ensure that the output of the transmit filter has a maximum value of unity, we normalize the symbol alphabet to have a maximum value of unity:

$$a_k \in \left\{ \frac{\pm 1}{L-1}, \frac{\pm 3}{L-1}, \ldots, \pm 1 \right\} \tag{5-20}$$

As before, the transmit filter $p(t)$ is normalized to have a maximum value of unity. The analysis is identical to that for OOK, with the transmitted signal given by (5-8), and the sampled output of the matched filter given by (5-12). The distance between neighboring received symbols is $d = P_{avg}\sqrt{E_p}\frac{2}{L-1}$, so that the probability of a symbol error is given by [57][60]:

$$\text{Prob[symbol error]} = 2(1-1/L)Q\left(\frac{d/2}{\sqrt{N_0}}\right) = 2(1-1/L)Q\left(\frac{P_{avg}}{L-1}\sqrt{\frac{E_p}{N_0}}\right). \tag{5-21}$$

With Gray coding, each symbol error causes about one bit error on average, so that the BER can be approximated by:

$$BER \approx \frac{2(1-1/L)}{\log_2 L} Q\left(\frac{P_{avg}}{L-1}\sqrt{\frac{E_p}{N_0}}\right) \approx Q\left(\frac{P_{avg}}{L-1}\sqrt{\frac{E_p}{N_0}}\right). \tag{5-22}$$

The last approximation facilitates comparison with OOK and other modulation schemes; it is valid for moderate values of L at high SNR, in which case the error due to this assumption, measured in dB, is small [57].

An advantage of L-PAM is a reduced symbol rate and thus a reduced bandwidth. To achieve a bit rate of R_b, with each symbol carrying $\log_2 L$ bits of information, the symbol rate for L-PAM is:

$$1/T = \frac{R_b}{\log_2 L}. \tag{5-23}$$

The bandwidth requirement is therefore $B = R_b/\log_2 L$, assuming 100% excess bandwidth. This result is included in Table 5-1.

To determine the power efficiency of L-PAM, assume the rectangular pulse shape of (5-5), which has energy $E_p = T$. With (5-23), the BER for L-PAM from (5-22) reduces to:

$$BER \approx Q\left(\frac{P_{avg}}{L-1}\sqrt{\frac{\log_2 L}{N_0 R_b}}\right). \tag{5-24}$$

The power efficiency of L-PAM is therefore:

$$P_{req} = \frac{L-1}{\sqrt{\log_2 L}} \cdot \sqrt{N_0 R_b} Q^{-1}(BER) = \frac{L-1}{\sqrt{\log_2 L}} P_{OOK}. \tag{5-25}$$

Thus, L-PAM suffers a power penalty of $10 \cdot \log_{10}\left(\frac{L-1}{\sqrt{\log_2 L}}\right)$ dB relative to OOK.

The combined bandwidth and power efficiency of L-PAM is illustrated in Fig. 5-4, where the L-PAM points are marked with circles for $L \in \{2, 3, 4 \ldots, 16\}$. Note that 2-PAM is identical to OOK. Higher-level PAM results in a reduced bandwidth requirement at the expense of an increased power penalty: as L increases from 2 to 16, the bandwidth requirement decreases from $B = R_b$ to $B = R_b/4$, while the power penalty relative to OOK increases from 0 dB to 8.8 dB. 4-PAM is an attractive candidate, with a bandwidth requirement of $B = R_b/2$ and an optical power penalty of 3.3 dB.

Comparing Fig. 5-4 with Fig. 5-5, we see that multilevel baseband PAM is better-suited to the intensity-modulation channel than to the conventional channel; as L increases from 2 to 16, the power penalty increases from 0 dB to 13.3 dB on the conventional channel, as compared to a range of 0 dB to 8.8 dB on the intensity-modulation channel.

The PAM signal as constructed above consists of a d.c. bias plus a time-varying term that is symmetric about zero; see (5-8). A practical system need not be implemented in this way, but rather the optical intensity may be modulated directly from a minimum value of zero intensity to a maximum value of $2P_{avg}$, for a total of L levels, with no d.c. offset. In fact, this is no different from the implementation of (5-8); it is an equivalent way of viewing the same PAM signal, and so it has identical performance.

5.3.2 Subcarrier and Multiple-Subcarrier Modulation

In Fig. 5-6 we show a multiple-subcarrier intensity-modulation communication system. By letting $N = 1$, the figure is also applicable to a single-subcarrier system. Each of the N subcarriers is modulated using L-level quadrature-amplitude modulation (L-QAM). We assume that $\sqrt{L}$ is an integer. For example, $L = 4$ corresponds to 4-QAM, which is also known as quadriphase-shift keying (QPSK). We will denote the entire multi-subcarrier modulation scheme as N-L-QAM.

As shown in the figure, the transmitter output $X(t)$ in N-L-QAM is formed by adding together N different L-QAM signals with subcarrier frequencies $\{\omega_1, \omega_2, \ldots, \omega_N\}$, plus a d.c. bias signal to ensure that $X(t)$ is everywhere positive.

The n-th L-QAM signal u_n is formed by adding two $\sqrt{L}$-PAM signals in quadrature. Specifically, let $\{a_{I,n}[k]\}$ and $\{a_{Q,n}[k]\}$ represent the in-phase and quadrature PAM symbol sequences for the n-th subcarrier, respectively. We make the important assumption that, regardless of L, the complex constellation $\{a_I + ja_Q\}$ is inscribed in the unit circle. In other words, a_I and a_Q must take on values from the set:

$$a_I, a_Q \in \left\{ \frac{\pm 1/\sqrt{2}}{\sqrt{L}-1}, \frac{\pm 3/\sqrt{2}}{\sqrt{L}-1}, \ \ldots, \pm\frac{1}{\sqrt{2}} \right\}. \tag{5-26}$$

These symbol sequences drive a transmit pulse shape $p(t)$, which is identical for both in-phase and quadrature channels and for each subcarrier. As before, we assume that $p(t)$ is normalized so that its maximum value is unity. The rectangular pulse shape of (5-5) is a

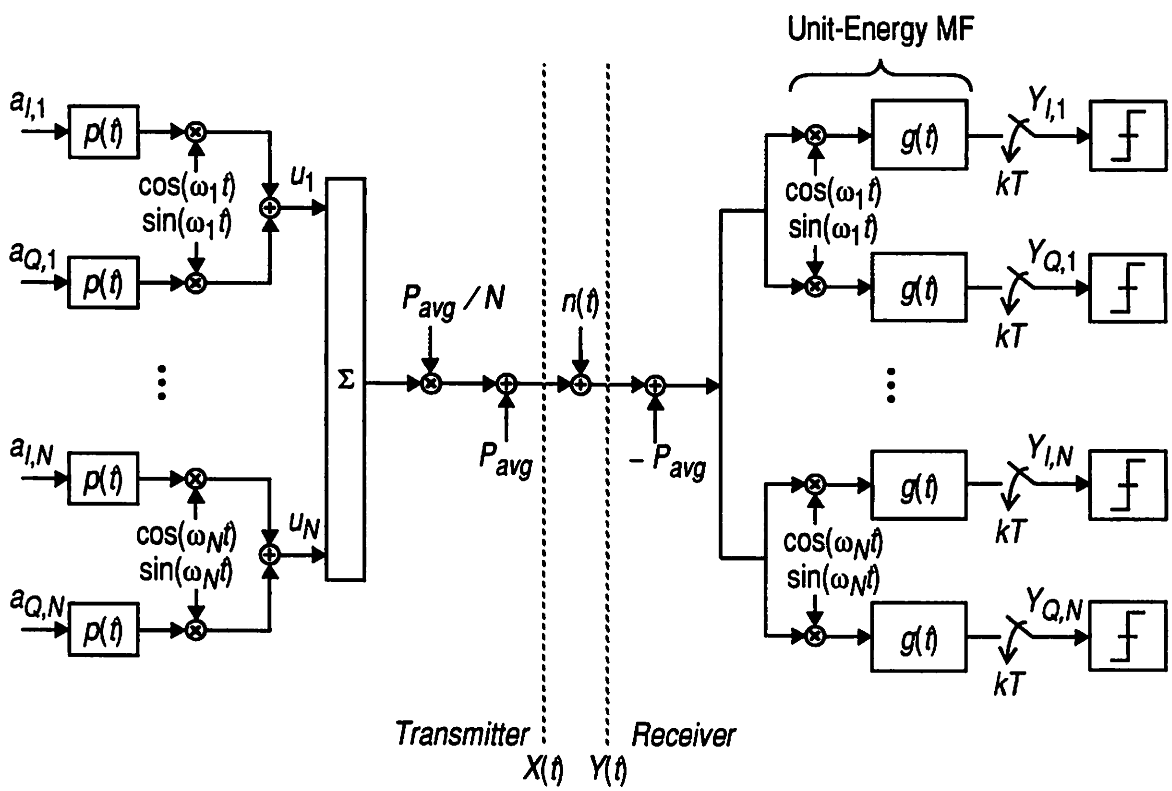


Fig. 5-6. Multiple subcarrier modulation: N-L-QAM.

common choice. The in-phase and quadrature PAM signals are multiplied by $\cos(\omega_n t)$ and $\sin(\omega_n t)$, respectively, and then added to form the n-th L-QAM signal u_n. Because we have constrained the constellation to the unit circle, and because $p(t)$ has a maximum value of unity, the n-th L-QAM signal is symmetric about zero and has a maximum value of unity. Therefore, the output of the summing bar in Fig. 5-6 will also be symmetric about zero, but with a maximum amplitude of N. As shown in the figure, this signal is scaled by a factor P_{avg}/N so that the result takes values from the range $[-P_{avg}, P_{avg}]$. A d.c. bias of P_{avg} is then added to ensure that $X(t)$ is positive everywhere:

$$X(t) = P_{avg} + \frac{P_{avg}}{N} \sum_k p(t - kT) \sum_{n=1}^{N} \left\{ a_{I,n}[k]\cos(\omega_n t) + a_{Q,n}[k]\sin(\omega_n t) \right\}. \tag{5-27}$$

As before, the received signal $Y(t)$ is simply $X(t) + n(t)$, where $n(t)$ is white Gaussian noise. The receiver shown in Fig. 5-6 is the maximum likelihood receiver. It first subtracts off the d.c. bias, which carries no information, and then demodulates each of the in-phase and quadrature PAM signals using a matched filter. The n-th demodulator multiples the received signal by $\cos(\omega_n t)$ and $\sin(\omega_n t)$ and passes the results through identical filters with impulse response $g(t)$. Here, $g(t)$ is matched to the transmit pulse shape $p(t)$ and normalized so that the combination of demodulator and baseband filter is equivalent to a unit-energy passband matched filter:

$$g(t) = \sqrt{\frac{2}{E_p}}\, p(-t). \tag{5-28}$$

(This normalization is natural because it causes the discrete-time noise sequence at the output of the sampler to have the same PSD as the continuous-time noise source of the channel, namely N_0.) The outputs of the two receiver filters in the n-th demodulator are then sampled at the baud rate yielding two discrete-time sequences $\{Y_{I,n}[k]\}$ and $\{Y_{Q,n}[k]\}$, where:

$$Y_{I,n}[k] = a_{I,n}[k]\frac{P_{avg}}{N}\sqrt{\frac{E_p}{2}} + n_{I,n}[k]\,, \tag{5-29}$$

$$Y_{Q,n}[k] = a_{Q,n}[k]\frac{P_{avg}}{N}\sqrt{\frac{E_p}{2}} + n_{Q,n}[k]\,, \tag{5-30}$$

where $\{n_{I,n}[k]\}$ and $\{n_{Q,n}[k]\}$ are independent, identically distributed white Gaussian noise sequences with PSD N_0. The probability of error for the n-th L-QAM signal can be approximated by [57]:

$$BER \approx Q\left(\frac{d}{2\sqrt{N_0}}\right) = Q\left(\frac{1/2}{\sqrt{L}-1}\cdot\frac{P_{avg}}{N}\sqrt{\frac{E_p}{N_0}}\right) \tag{5-31}$$

where d is the distance between neighboring received symbols in the complex plane:

$$d = \frac{\sqrt{E_p}}{\sqrt{L}-1}\cdot\frac{P_{avg}}{N}\,. \tag{5-32}$$

The BER of the composite N-L-QAM signal will be larger than (5-31) by a factor N, but we will ignore this difference to facilitate comparison with other modulation techniques. The error due to this assumption, measured in dB, is small for moderate values of N [57].

We assume that the bit rate is fixed at some value R_b, but that the number of subcarriers N and the number of QAM levels L can take on any value. A given N and L will therefore fix the symbol rate $1/T$. Using N subcarriers, each carrying $\log_2 L$ bits per symbol, an aggregate bit rate of R_b requires a symbol rate of:

$$1/T = \frac{R_b}{N\cdot\log_2 L}\,. \tag{5-33}$$

Thus, as N and L are varied, the symbol rate changes, which in turn changes E_p, which in turn changes the BER per (5-31). We will now narrow the scope to transmit pulse shapes $p(t)$ with energy $E_p = T$, which includes the rectangular pulse shape of (5-5) as well as the zero-excess-bandwidth pulse $p(t) = \sin(\pi t/T)/(\pi t/T)$. Substituting $E_p = T$ from (5-33) into (5-31) yields a BER of:

$$BER \approx Q\left(\frac{1}{\sqrt{4N}} \cdot \frac{\sqrt{\log_2 L}}{\sqrt{L}-1} \cdot \frac{P_{avg}}{\sqrt{N_0 R_b}}\right) \tag{5-34}$$

Solving for P_{avg} yields the following average optical power requirement for N-L-QAM multi-subcarrier modulation to achieve a given BER:

$$P_{req} = \sqrt{4N}\frac{\sqrt{L}-1}{\sqrt{\log_2 L}}P_{OOK}, \tag{5-35}$$

where $P_{OOK} = \sqrt{N_0 R_b}Q^{-1}(BER)$ (see (5-15)). This result is included in Table 5-1. The optical power penalty for N-L-QAM multiple-subcarrier modulation (relative to OOK) can thus be decomposed into three components:

$$\text{power penalty} = 10\cdot\log_{10}(P_{req}/P_{OOK})$$

$$= \underbrace{1.5}_{\substack{\text{offset}\\\text{penalty}}} + \underbrace{5\cdot\log_{10}N}_{\substack{\text{multi-subcarrier}\\\text{penalty}}} + \underbrace{5\cdot\log_{10}\left(\frac{(\sqrt{L}-1)^2}{\log_2\sqrt{L}}\right)}_{\substack{\text{multi-level}\\\text{penalty}}} \text{ dB}. \tag{5-36}$$

The offset penalty is due to power wasted in the d.c. offset required by all subcarrier techniques. The multi-subcarrier penalty, which is zero for $N = 1$, represents the penalty when more than one subcarrier is used. The multi-level penalty, which is zero for 4-QAM (QPSK), represents the penalty due to larger symbol alphabets.

The bandwidth required by N-L-QAM multi-subcarrier modulation to achieve a bit rate of R_b, assuming rectangular pulses or 100% excess bandwidth pulses, is roughly the number of carriers times twice the symbol rate: $B \approx 2N/T = 2R_b/\log_2 L$. The required bandwidth is thus independent of the number of subcarriers, depending only on the alphabet size.

The bandwidth and power efficiency results for N-L-QAM are illustrated in Fig. 5-4 for $N \in \{1, 2, \ldots, 8\}$ and $L \in \{4, 16, 64\}$. Consider first the case of $L = 4$. Since 4-QAM is equivalent to QPSK, the N-4-QAM points in the figure, which are marked with a triangle, are labeled N-QPSK. The figure shows that a single-subcarrier QPSK system (1-QPSK) requires the same bandwidth as OOK but 1.5 dB more optical power. As the number of QPSK subcarriers increases from one to eight, the bandwidth requirement remains unchanged, but the power penalty increases from 1.5 dB to 6 dB according to (5-36). Therefore, from the standpoint of bandwidth and power efficiency, there is no advantage to using more than one QPSK subcarrier, but rather there is a power penalty.

Consider next the 16-QAM data points, which are marked with a diamond in Fig. 5-4. A single subcarrier 16-QAM system requires half the bandwidth required by OOK to achieve a given bit rate, but requires 4.8 dB more power—a significant penalty. As the number of subcarriers increases from one to eight, the bandwidth requirement is unchanged and the power penalty increases from 4.8 dB to 9.3 dB.

The 64-QAM data points are marked with a downward-pointing triangle. The bandwidth required by all N-64-QAM signals is one third of the bandwidth required by an OOK signal, with power penalties ranging from 7.6 dB to 12.1 dB as the number of subcarriers N is increased from one to eight.

We now compare the N-L-QAM power penalty, which is given by (5-36), with the L-PAM penalty, which is given by (5-25):

$$\text{power penalty} = 5 \cdot \log_{10}\left(\frac{(L-1)^2}{\log_2 L}\right) \text{ dB}. \tag{5-37}$$

We see that the "multilevel" component of the penalty in (5-36) is equal to the entire penalty for $\sqrt{L}$-PAM. Therefore, for all L, there is a 1.5-dB offset penalty for single-subcarrier L-QAM relative to $\sqrt{L}$-PAM. This is evident in Fig. 5-4, where each of the 1-L-QAM data points is 1.5 dB above the locus of L-PAM data points.

It is instructive to note that there is no such penalty on the conventional channel, as shown in Fig. 5-5: the performance of L-QAM is identical to the performance of $\sqrt{L}$-PAM.

Furthermore, unlike the intensity-modulation channel, for which the penalty due to multiple subcarriers grows as $5 \cdot \log_{10} N$, the conventional-channel results hold for all values of N; the number of subcarriers is irrelevant.

The above N-L-QAM analysis assumed that the transmit pulse shape had energy $E_p = T$. This is true for the rectangular pulse of (5-5) and the ideal sinc pulse $p(t) = \sin(\pi t/T)/(\pi t/T)$, but not true in general. From (5-31) it is clear that it is desirable for E_p to be as large as possible, given the constraint that the maximum value of $p(t)$ must be unity, and that the pulse satisfy the Nyquist criterion for no intersymbol interference. Consider the raised cosine pulse with 100% excess bandwidth with Fourier transform [57]:

$$P(\omega) = \begin{cases} \frac{T}{2}(1 + \cos(\omega T/2)) & \text{for } |\omega| \leq 2\pi/T \\ 0 & \text{for } |\omega| > 2\pi/T \end{cases} . \tag{5-38}$$

From (5-11) we find that, in this case, $E_p = 3T/4$, and so systems using this pulse shape will suffer a penalty of $5 \cdot \log_{10}(3/4) = 1.2$ dB relative to pulses satisfying $E_p = T$ (such as the rectangular pulse or the zero-excess-bandwidth pulse). Whether there exists a pulse shape satisfying the Nyquist criterion and having a maximum value of unity that has energy $E_p > T$ is an open question and warrants further investigation.

Instead of using L-QAM modulation on each subcarrier, one could also use BPSK. The resulting modulation scheme, which we denote N-BPSK, where N is the number of subcarriers, can be analyzed in a manner similar to the N-L-QAM analysis. It is easy to show that N-BPSK has the same power efficiency as N-4-QAM (N-QPSK), but requires twice the bandwidth. The N-BPSK data points in Fig. 5-4 are marked with the symbol $\times$. Because N-BPSK offers no advantages over BPSK, other than an insignificant decrease in complexity, an N-QPSK system will always be preferable to an N-BPSK system.

5.3.3 Pulse-Position Modulation

The multilevel modulation schemes considered in previous sections were able to achieve higher bandwidth efficiency at the expense of decreased power efficiency. In this section we consider L-level pulse-position modulation (L-PPM), a fundamentally different modulation scheme that achieves high power efficiency at the expense of reduced bandwidth efficiency.

Consider Fig. 5-7, which shows a block diagram for an ideal L-PPM system. The input bits with bit rate R_b are grouped in blocks of length $\log_2 L$ at a symbol rate of $1/T$, and from each block one of L possible signals is chosen to transmit. Simply stated, the symbol interval of duration T is divided into L sub-intervals, and to signify the l-th symbol, the optical intensity is "on" during the l-th sub-interval and "off" everywhere else. In other words, the output of the "encoder" block of Fig. 5-7 during the k-th baud interval $t \in [kT, (k+1)T)$ is given by $p_{l[k]}(t - kT)$, where $l[k] \in \{1, 2, \ldots, L\}$ denotes the position of the "on" sub-interval during the k-th baud interval, and where $\{p_l(t)\}$ is a family of pulse shapes given by:

$$p_l(t) = \begin{cases} 1 & \text{for } t \in [(l-1)T/L,\, lT/L) \\ 0 & \text{elsewhere} \end{cases}, \quad \text{for } l \in \{1, 2, \ldots, L\}\,. \tag{5-39}$$

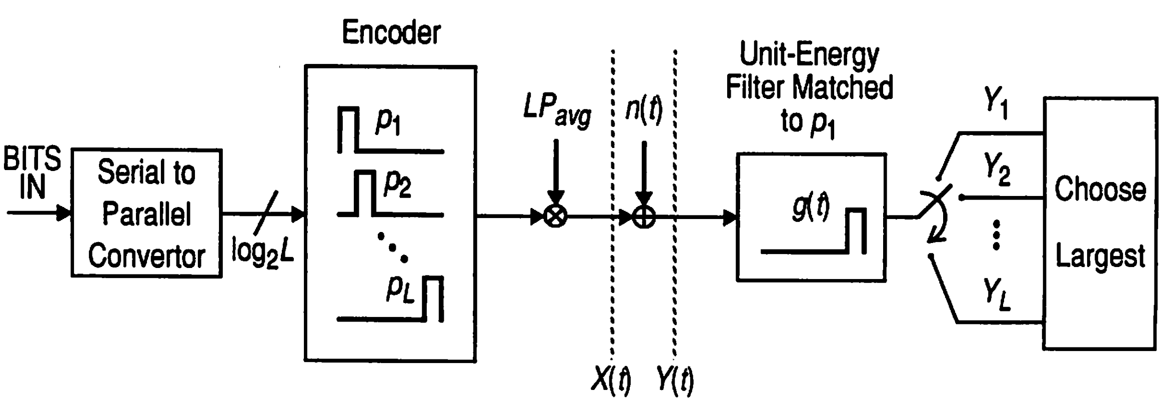


Fig. 5-7. Block diagram for a pulse-position modulation system.

These pulse shapes are normalized to have a maximum value of unity. The output of the encoder is scaled by a factor LP_{avg} to ensure that the resulting intensity $X(t)$ has an average value of P_{avg}:

$$X(t) = LP_{avg} \sum_k p_{l[k]}(t - kT) . \tag{5-40}$$

As before, we ignore path loss and multipath dispersion, so that the received intensity $Y(t)$ is simply $X(t)$ plus a white Gaussian noise with PSD N_0. The receiver shown in the figure is the correlation receiver, again a maximum likelihood receiver [57]. It consists of a filter $g(t)$ that has unit energy and is matched to the first-position pulse shape $p_1(t)$:

$$g(t) = \sqrt{\frac{L}{T}} p_1(-t) . \tag{5-41}$$

The L different receiver branches are able to share this single filter by sampling its output at a rate of L/T, yielding the same L sufficient statistics Y_1 through Y_L that would result from a receiver employing a bank of L matched filters. The receiver compares Y_1 through Y_L and decides on the l-th symbol when Y_l is the largest.

We will now derive the bit-error rate for this system. Because we neglect multipath dispersion and timing-error effects, we need consider only a single symbol transmission rather than an infinite sequence of symbols. Furthermore, the problem is symmetric, so we can assume that the first symbol $p_1(t)$ was sent without loss of generality. In this case, the L sufficient statistics are given by:

$$\begin{aligned} &Y_1 = S + n_1 \\ &Y_k = n_k \qquad \text{for } k \in \{2, 3, \ldots, L\} , \end{aligned} \tag{5-42}$$

where $S = P_{avg}\sqrt{LT}$ and where $\{n_k\}$ for $k \in \{1, 2, \ldots, L\}$ are independent, identically distributed zero-mean Gaussian random variables with variance N_0. The probability of a symbol error can be calculated as follows:

$$\text{Prob[symbol error]} = 1 - \text{Prob[correct decision]} \tag{5-43}$$

$$= 1 - \text{Prob}[(n_2 < S{+}n_1)\ \&\ (n_3 < S{+}n_1)\ \&\ \ldots\ \&\ (n_L < S{+}n_1)] \tag{5-44}$$

$$= 1 - \text{E}\{\text{Prob}[(n_2 < S{+}n_1)\ \&\ (n_3 < S{+}n_1)\ \&\ \ldots\ \&\ (n_L < S{+}n_1) \mid n_1]\} \tag{5-45}$$

where the expectation E{ · } is taken over the random variable n_1 [61]. With n_1 a known quantity, the $L-1$ events in (5-45) are independent, so that the probability of a symbol error reduces to:

Prob[symbol error] =

$$1 - \text{E}\{\text{Prob}[(n_2 < S{+}n_1)|n_1]\cdot\text{Prob}[(n_3 < S{+}n_1)|n_1]\ldots\text{Prob}[\,(n_L < S{+}n_1)|n_1]\} \tag{5-46}$$

$$= 1 - \text{E}\{\text{Prob}[(n_2 < S{+}n_1) \mid n_1]^{\,L-1}\} \tag{5-47}$$

$$= 1 - \text{E}\{\left[1 - Q\left(\frac{S + n_1}{\sqrt{N_0}}\right)\right]^{L-1}\}\,. \tag{5-48}$$

This expression is identical to the symbol-error rate expression for L-ary frequency-shift keying (L-FSK) given in [57][60], as expected, because both L-FSK and L-PPM are orthogonal signaling schemes. This expression can be simplified by noting that, at high SNR, the $Q(\cdot)$ term will be very small, in which case $(1 - Q(\cdot))^{L-1} \approx 1 - (L-1)Q(\cdot)$, so that:

$$\text{Prob[symbol error]} \approx 1 - \text{E}\{1 - (L-1)Q\left(\frac{S + n_1}{\sqrt{N_0}}\right)\} \tag{5-49}$$

$$= (L-1)\text{E}\{Q\left(\frac{S + n_1}{\sqrt{N_0}}\right)\} = (L-1)Q\left(\frac{S}{\sqrt{2N_0}}\right), \tag{5-50}$$

where the last equality follows from the identity $\text{E}[Q(x)] = Q\,(\mu/\sqrt{\sigma^2 + 1})$ for a Gaussian random variable x with mean μ and variance σ^2.

The last expression in (5-50) can also be arrived at by a signal space argument. The L possible signals in L-PPM are orthogonal and thus form an orthogonal basis for the signal subspace spanned by the transmitted signals. An ML receiver projects the received signal onto this L-dimensional subspace and decides on the signal vector closest (in a Euclidean sense) to the result. With proper normalization, the $L-1$ incorrect signal vectors are all a distance $d = S\sqrt{2}$ from the correct signal vector, and the L components of the noise vector are independent Gaussian random variables with variance N_0. Application of the union bound thus leads to (5-50) [57].

Because the L possible signals are equally likely and orthogonal, the BER is related to the probability of a symbol error by [60]:

$$BER = \frac{L/2}{L-1}\text{Prob[symbol error]} \tag{5-51}$$

$$= \frac{L}{2}Q\left(\frac{S}{\sqrt{2N_0}}\right) \approx Q\left(\frac{S}{\sqrt{2N_0}}\right), \tag{5-52}$$

the last approximation being valid at high SNR for moderate L.

The symbol rate required by L-PPM to achieve a bit rate of R_b is simply:

$$1/T = \frac{R_b}{\log_2 L}\ . \tag{5-53}$$

The bandwidth required by L-PPM can be approximated by the bandwidth of a single chip pulse: $B \approx L/T = LR_b/\log_2 L$. This result is noted in Table 5-1.

Substituting (5-53) into $S = P_{avg}\sqrt{LT}$ and substituting the result into (5-52) yields the following BER expression for L-PPM:

$$BER \approx Q\left(\sqrt{\frac{1}{2}L \cdot \log_2 L}\frac{P_{avg}}{\sqrt{N_0 R_b}}\right). \tag{5-54}$$

The average optical signal power required to achieve a given BER for an L-PPM system can be found by solving this expression for P_{avg}, yielding:

$$P_{req} = \frac{\sqrt{N_0 R_b}Q^{-1}(BER)}{\sqrt{\frac{1}{2}L \cdot \log_2 L}} = \frac{P_{OOK}}{\sqrt{\frac{1}{2}L \cdot \log_2 L}}, \tag{5-55}$$

where P_{OOK} was first defined in (5-15). This result is also noted in Table 5-1. Substituting $L = 2$ yields a sensitivity for 2-PPM that is identical to OOK, a fact that we had derived in section 5.2.1. From (5-55) we see that, for any L greater than two, the optical power required by L-PPM is smaller than that required by OOK. In principle, the power requirement can be made arbitrarily small by making L suitably large.

In Fig. 5-4 we illustrate the combined bandwidth and power efficiency of L-PPM by plotting power penalty versus required bandwidth for $L \in \{2,3,\ldots,16\}$ using square symbols. Note that 2-PPM has the same power efficiency as OOK but requires twice the bandwidth. 3-PPM is slightly better than 2-PPM both in terms of bandwidth and power efficiency, but its implementation is complicated by the fact that $\log_2 3$ is not an integer. It is apparent that 4-PPM is particularly attractive because it has the same bandwidth requirement as 2-PPM but requires 3.8 dB less optical power. Therefore, from the viewpoint of bandwidth and power efficiency, 4-PPM is always preferable over 2-PPM. As L increases from 4 to 16, the bandwidth requirement increases from $2R_b$ to $4R_b$, while the sensitivity increases from 3 dB better than OOK to 7.5 dB better than OOK. Evidently, the high-sensitivity of high-level PPM motivated Photonics Corporation to use 16-PPM in their 1-Mb/s infrared transceiver [20].

It is difficult to predict whether bandwidth efficiency or power efficiency will be the more important goal in an indoor wireless link. One could argue that power efficiency is more important, because the channel is fundamentally power limited; the background noise is unavoidable, and safety considerations preclude the possibility of making the transmitter power arbitrarily large. For a given transmitter power, background illumination power, and bit rate, it is desirable to maximize the allowable distance between transmitter and receiver, which is equivalent to maximizing the power efficiency. The primary impediment to achieving high receiver bandwidths, on the other hand, is the high capacitance of large-area photodetectors. Unlike the background noise, this impediment is not fundamental; high bandwidths can be achieved at the expense of increased electronic complexity and power consumption. For this reason, low-to-moderate-speed wireless links will benefit from modulation schemes like PPM. High-speed systems having chip rates higher than about 10 MHz, however, must contend with intersymbol interference due to multipath dispersion as discussed in chapter 4. The 1-Mb/s 16-PPM Photonics transceiver, for example, is not affected appreciably by multipath dispersion, because it uses a chip rate of only 4 MHz. The performance of PPM in the face of multipath dispersion is unknown and must be investigated carefully before adopting PPM in high-speed systems.

The observations of this section were based almost exclusively on bandwidth efficiency and power efficiency. There are other important criteria by which to assess modulation performance, however, such as resistance to multipath dispersion effects and amenability to multiple access. These issues are discussed further in section 5.4.

5.3.4 Alternative Modulation Schemes

The following modulation schemes were not included in the above discussion because their performance was found to be inferior. Their bandwidth and power efficiencies are included here for completeness, although to prevent clutter they are not included in Table 5-1 and Fig. 5-4.

5.3.4.1 *L*-FSK

The transmitted intensity for L-level frequency-shift keying (L-FSK) is given by:

$$X(t) = P_{avg} + P_{avg}\sum_k p_{l[k]}(t - kT) \tag{5-56}$$

where $\{p_l(t)\}$ is a family of pulse shapes given by:

$$p_l(t) = \begin{cases} \cos(\omega_l t) & \text{for } t \in [0, T) \\ 0 & \text{elsewhere} \end{cases}, \quad \text{for } l \in \{1, 2, \ldots, L\}, \tag{5-57}$$

and $\omega_l = 2\pi(l - 0.5/T)$, assuming a modulation index of unity [60]. The BER for L-FSK can be derived in a manner similar to that for L-PPM, with the result given by (5-52) with

$S = P_{avg}\sqrt{T/2}$. The symbol rate required to achieve a given bit rate is $1/T = R_b/\log_2 L$, so that the bandwidth requirement for L-FSK is [60]:

$$B = \frac{L}{T} = \frac{LR_b}{\log_2 L}. \tag{5-58}$$

The power requirement for L-FSK is:

$$P_{req} = \frac{2}{\sqrt{\log_2 L}} P_{OOK}. \tag{5-59}$$

Compared with L-PPM, L-FSK requires the same bandwidth but requires more optical power by a factor $\sqrt{2L}$. Thus, as L increases from 2 to 16, the L-FSK data points in Fig. 5-4 would lie directly above the corresponding L-PPM data points, with a power penalty (relative to OOK) decreasing from 3 dB to 0 dB.

5.3.4.2 RZ-OOK

A variation of OOK proposed by Kotzin [25] for the intensity-modulation channel is return-to-zero on-off keying with duty cycle δ (δ-RZ-OOK). This is similar to the NRZ OOK scheme considered earlier, except that the pulse shape is high for only a fraction $\delta \in (0, 1]$ of the baud interval. The transmitted intensity for δ-RZ-OOK is given by:

$$X(t) = \frac{2}{\delta} P_{avg} \sum_k a_k p(t - kT) \tag{5-60}$$

where $a_k \in \{0, 1\}$ are the data symbols and the pulse shape is:

$$p(t) = \begin{cases} 1 & \text{for } t \in [0, \delta T) \\ 0 & \text{elsewhere} \end{cases}. \tag{5-61}$$

The bandwidth and power requirements for δ-RZ-OOK are:

$$B = \frac{R_b}{\delta} \tag{5-62}$$

$$P_{req} = \sqrt{\delta} P_{OOK}. \tag{5-63}$$

When $\delta = 1$, δ-RZ-OOK reduces to the NRZ OOK scheme examined earlier, so that the bandwidth and power requirements are identical to those for OOK. As δ decreases, the bandwidth requirement grows faster than the power requirement decreases. Thus, δ-RZ-OOK is inferior to L-PPM, because it requires more bandwidth to achieve a given sensitivity improvement. As δ decreases from 1 to 0.25, the trajectory of the δ-RZ-OOK data points in Fig. 5-4 would be almost linear, starting at the OOK point and intersecting with the right-hand boundary of the plot with a power penalty of – 3 dB.

5.3.4.3 Spread Spectrum

The immense bandwidth requirement of direct-sequence and frequency-hopped spread spectrum modulation schemes makes them inappropriate for high-speed non-directed infrared links. For example, a modest spreading factor of ten would require an electrical bandwidth of 1 GHz for a 100 Mb/s system. For this reason, spread spectrum modulation is not considered here.

5.4 OTHER CRITERIA

If power efficiency and bandwidth efficiency were our only concern, then we could reduce the field of candidate modulation schemes with the aid of Fig. 5-4, by eliminating those data points for which another data point has both higher power efficiency and higher bandwidth efficiency. This would eliminate all of the subcarrier modulation schemes, leaving only L-PPM with $L \in \{3, 4, \dots\}$, and L-PAM with $L \in \{2, 3, \dots\}$. Before we rule any modulation schemes out, however, we must consider other criteria besides power efficiency and bandwidth efficiency, such as resistance to multipath and amenability to multiple access.

Ignoring for the moment the reduced power and bandwidth efficiency of subcarrier schemes, they do offer some advantages over baseband schemes. For example, by assigning different subcarrier frequencies to different users, asynchronous multiple access can be achieved. Also, a multi-subcarrier system will be resistant to multipath-induced ISI when the symbol rate of each sub-band is smaller than about 10 Mbaud for typical indoor applications. This follows because, as shown in chapter 4, the multipath impulse response is about 50-ns long for typical office environments. Thus, data transmitted every 100 ns, or at a rate of 10 Mbaud, will not experience appreciable ISI. This is clearly illustrated in Fig. 4-10, where power penalty is plotted versus baud rate for a number of different transmitter configurations, and the power penalty increases significantly as the baud rate increases above about 10 Mbaud.

Alternatively, the mitigation of ISI through baud rate reduction can be explained by a frequency domain argument: there is little variation in the frequency response of the multipath channel over a bandwidth of about 10 MHz, as shown in Fig. 4-6 through Fig. 4-8, and so a 10-Mbaud sub-carrier (or baseband) signal will see a nearly flat frequency response.

Unfortunately, as shown in the last section, the power penalty for all multiple-subcarrier systems is $5 \cdot \log_{10} N$ dB larger than for a single-subcarrier system, so the resistance to multipath provided by multiple subcarriers is paid for by an increased power penalty. In fact, the net effect may be an increased power penalty, depending on the severity of the ISI in the first place. The following numerical example illustrates this point.

Consider the plot of ISI penalty versus baud rate in Fig. 4-10; although it assumes OOK modulation and an unequalized integrate-and-dump receiver, it can be used to approximate the ISI penalty for other modulation schemes as well. At 100 Mb/s, the unequalized OOK system suffers a penalty of at most about 6 dB for the four transmitter configurations considered. By switching from OOK to a multi-subcarrier technique, say five 10-Mbaud QPSK subcarriers, the ISI penalty is reduced to a fraction of a dB per channel, for a total of about 0.5 dB. This sensitivity gain is obliterated almost completely by the d.c. offset and multi-subcarrier penalty of $1.5 + 5 \cdot \log_{10} 5 = 5$ dB. Thus, in going from a baseband OOK system without an equalizer to a five-subcarrier QPSK system, the net gain is about 0.5 dB.

Observations like these lead to the conclusion that multipath immunity alone is not sufficient justification for the increased complexity of a multi-subcarrier system. Particularly in light of the fact that an adaptive equalizer such as a decision-feedback equalizer (DFE) can reduce the 6-dB ISI penalty of an OOK system considerably [62]. In other words, while the subcarrier system in the above example is only 0.5 dB better than the unequalized OOK system, adding an equalizer to an OOK system can reduce the ISI penalty by 4 dB or more [62][63].

Another disadvantage of the subcarrier techniques relative to baseband modulation is the increased attenuation of the multipath frequency response at high frequencies. From Fig. 4-6-c, Fig. 4-7-c, and Fig. 4-8-c, we see that in each case the frequency response is maximum at d.c. and decreases steadily as the frequency increases to about 30 MHz. Thus, each sub-carrier in a multi-subcarrier system will experience a "flat-fading" loss relative to a baseband system. This penalty is not reflected in the analysis of the previous section, and thus represents a further reduction in power efficiency for multi-subcarrier systems.

Despite the poor power efficiency of multi-subcarrier modulation, it may still find use in certain applications. For example, in some applications the transmitted signal consists of a number of data streams multiplexed together, and a receiver is interested in detecting only one of the data sub-streams. This is the case, for example, in fiber-optic video distribution systems [64]. Multi-subcarrier modulation allows each receiver to detect a single subcarrier only, obviating the need for each receiver to have the high-speed digital circuitry necessary to detect the composite signal and perform the demultiplexing operation.

For applications in which a single receiver may require high data rates near 100 Mb/s, it seems that multi-subcarrier modulation is not the best candidate. Rather, one of the baseband modulation schemes, L-PAM or L-PPM, with OOK being a special case of L-PAM, should be adopted. Without detailed specifications of the application, such as the expected room size, multipath response, and multiple-access protocol, it is not possible at this point to isolate a single modulation scheme as the definitive choice; all we can do is point out the relative merits of the various modulation schemes in terms of bandwidth efficiency, power efficiency, and multipath immunity, and leave it to the system designer to choose the scheme that best fits their needs. Nevertheless, we can say that the following modulation schemes are all good candidates: OOK, 2-PPM, 4-PPM, 8-PPM, 4-PAM, and 8-PAM. OOK is an obvious choice because of its simplicity. 2-PPM is only slightly more complicated to implement, but 4-PPM will usually be preferable over 2-PPM because of its higher sensitivity. Further research is needed to examine the effects of multipath on the higher-level PPM signaling schemes and to assess the prospects of adaptive equalization. The other multilevel schemes trade of bandwidth efficiency and power efficiency, and will be appropriate for channels that are either highly bandwidth limited or highly power limited.

5.5 COHERENT OPTICAL COMMUNICATION

Coherent optical transmitters modulate the phase or frequency of the roughly 3×10^{14} Hz optical carrier directly, and coherent (heterodyne) optical detectors add light from a local laser to the received lightwave as part of the detection process. The lasers in a coherent system must be single mode and have narrow spectral linewidths. Because the background illumination has a broad spectrum, its effect on coherent detection is almost negligible [65], allowing a coherent wireless system to achieve acceptable error performance with as few as ten received photons per bit [66].

Unfortunately, application of coherent detection to non-directed systems is hampered by the requirement that the polarization and amplitude of the local optical signal be matched with the received optical field over the entire photodetector surface. In a non-directed LOS application, the received lightwave will have a single spatial mode, but both its direction and its polarization will be unknown, so that these attributes must be tracked actively by the receiver and matched by the local lightwave — an expensive proposition. For light undergoing diffuse reflections, coherent detection is even more difficult because the received lightwave will have multiple spatial modes [67].

Consider a received lightwave with power P_S, electric field $\mathbf{E}_S$, and frequency ω_S and a local lightwave with power P_{LO}, electric field $\mathbf{E}_{LO}$, and frequency ω_{LO}. A coherent detector couples the two lightwaves together, typically with a beam splitter, and directs the result towards a photodetector, which produces an a.c. current given by:

$$i_{IF}(t) = 2r\sqrt{\eta_{mix}P_S P_{LO}}\cos(\omega_{IF}t) + \text{shot noise} \tag{5-64}$$

where r is the detector responsivity, $\omega_{IF} = |\omega_S - \omega_{LO}|$ is the intermediate frequency and η_{mix} is the *mixing efficiency* [65][68][69][70]:

$$\eta_{mix} = \frac{\left|\int (\mathbf{E}_S \cdot \mathbf{E}_{LO}^{*})\, dA\right|^2}{\int |\mathbf{E}_S|^2 dA \int |\mathbf{E}_{LO}|^2 dA}\ . \tag{5-65}$$

The integrals are performed over the surface of the photodetector. In (5-64) we see that the signal power is multiplied by η_{mix}, implying an inverse relationship between power requirement and mixing efficiency; as $\eta_{mix} \rightarrow 0$, $P_{req} \rightarrow \infty$.

A heterodyne BPSK system, for example, requires 18 photons/bit when the mixing efficiency is unity [66]. At 100 Mb/s and 810 nm, this corresponds to detected optical power of 0.5 nW. From (5-15) and (5-16), on the other hand, an ideal direct-detection receiver requires:

$$rP_{OOK}T/q = \sqrt{\frac{rP_{bg}T}{q}}Q^{-1}(BER) \approx 35{,}000 \text{ photons/bit}, \tag{5-66}$$

assuming a responsivity of 0.53 A/W, a detected background power of P_{bg} = 1 mW, a bit rate of 100 Mb/s, and an error rate of $BER = 10^{-9}$. The difference in sensitivity is about 33 dB.

The above numerical example shows that a mixing efficiency of only 0.005% is needed to make the coherent system more sensitive than the direct-detection system. Further research is needed to determine whether such mixing efficiencies can be achieved in practical non-directed applications. In any event, the high complexity of a coherent receiver makes intensity modulation and direct detection the better choice for a low-cost system.

5.6 CONCLUSIONS

We defined an intensity modulation channel, for which the input is constrained to be positive, and for which the average amplitude of the input cannot exceed a given maximum value. We compared the power efficiency and bandwidth efficiency for a number of modulation schemes on this channel, and showed how the results differed from those of the conventional AWGN channel. Our results are applicable to all systems using intensity modulation, including fiber-optic systems. We showed that a single subcarrier system suffers a 1.5-dB penalty due to the required d.c. offset, and that multi-subcarrier systems suffer an additional loss of $5 \cdot \log_{10} N$ dB, where N is the number of subcarriers. These penalties make baseband schemes better suited for the intensity modulation channel. For example, L-PPM was shown to be the most sensitive of the modulation schemes considered, while OOK and L-PAM are less sensitive but are more bandwidth efficient. Considering bandwidth efficiency, power efficiency, receiver complexity, and multipath immunity, OOK with decision-feedback equalization seems to be a good choice.

CHAPTER 6

SYSTEM ISSUES

Chapters 2 through 5 have concentrated on the physical-layer problem of point-to-point communication using non-directed infrared radiation. Although the results presented there stand alone, the purpose of this chapter is to address some of the issues that arise when one tries to use non-directed infrared links as building blocks for a wireless LAN. Because a complete design of a wireless LAN is beyond the scope of this thesis, the discussion presented here is not an in-depth one; nevertheless, it should convince the reader that there are no fatal obstacles to an infrared LAN lurking in the higher network layers.

6.1 INTRODUCTION

From the perspective of higher network layers, an infrared-based network is similar in most respects to a micro-cellular (or nano-cellular) radio-based network, and so many of the principles of radio networks can be applied to an infrared network. For example, when choosing a modulation scheme, its suitability to the multiple access and multiplexing protocols should be considered. There are two notable differences between infrared networks and radio networks, however. First, infrared has virtually unlimited bandwidth, and second, neighboring infrared links are less likely to interfere due to the opaqueness of most objects at infrared wavelengths. In principle, these differences should make an infrared network easier to design than a micro-cellular radio network.

Refer to Fig. 1-2, which shows the layout of a typical wireless LAN environment. The network can be decomposed into two levels in a tree-like fashion, with the wired network of base stations forming the backbone network, and the wireless links at each base station forming the link level. Depending on the application, the backbone network may have distributed control or it may have a central controller. When communication between different portables (peer-to-peer communication) accounts for the majority of the traffic, the distributed approach is more natural. It is unlikely, however, that peer-to-peer traffic alone would justify a high-speed network. Rather, high communication speeds would be needed only when accessing high-speed data from a information source that is wired to the backbone network. In this case, a central controller, coordinating the transmissions of each of the base stations, may be beneficial. In any event, the backbone network can use any conventional protocol; in their discussion of protocols for wireless infrared networks, Lessard and Gerla recommend a token bus protocol [71]. A conventional protocol may not be the best choice for a low-cost system, however; because the cell radius in an infrared LAN will be only a few meters, large geographical areas will require numerous base stations to provide ubiquitous coverage. It is important, therefore, to choose the protocol so that the interface between the backbone network and wireless links is as transparent as possible, alleviating the burden placed on base stations to translate data and convert protocols.

The fundamental non-physical-layer problems faced in the design of a wireless infrared network, like any wired or wireless network, are two-fold: first, how to multiplex data for multiple portables onto one or more down-link signals, and second, how to provide access to the uplink for multiple portables (multiple access). The solutions depend on the exact network architecture, which in turn depends on the application and the environment.

6.2 SINGLE-CELL ARCHITECTURES

We consider first the network topology for the specific case of an isolated cell. The results are applicable to multiple cells as well, provided that neighboring cells do not overlap. Topologies for the case of overlapping cells will be examined in section 6.3.

6.2.1 Unconstrained

Without a priori constraints on the wavelengths and subcarrier frequencies of the up link and down link, communication between a single base station and the portable units within its cell can be viewed as taking place on a shared bidirectional bus, as illustrated in Fig. 6-1-a. We assume that every portable unit can "see" the base station, and that the base station can "see" every portable unit. It it likely that some portable units will be "hidden" from the remaining portable units, however, in which case the shared bus of Fig. 6-1-a may be broken. The possibility of hidden nodes makes it inconvenient, but not impossible, to use protocols based on a carrier sense mechanism; a carrier-sense protocol can still be used if the base station rebroadcasts all up-link signals onto the down-link — an effective but inefficient procedure.

With the exception of the hidden node possibility, the unconstrained architecture for a single cell is identical to cable-based networks using bus topologies. In principle, therefore, it should be possible to adapt any bus protocol to the wireless case. Of course, there are numerous effects not reflected in the simplified picture of Fig. 6-1-a, such as multipath dispersion and path loss, but these effects do not change the underlying network architecture, and so the adoption of conventional protocols is still a valid approach. It may also an academic approach, however, because a practical system may isolate the up-link transmissions from the down-link transmissions to prevent interference. This greatly simplifies the

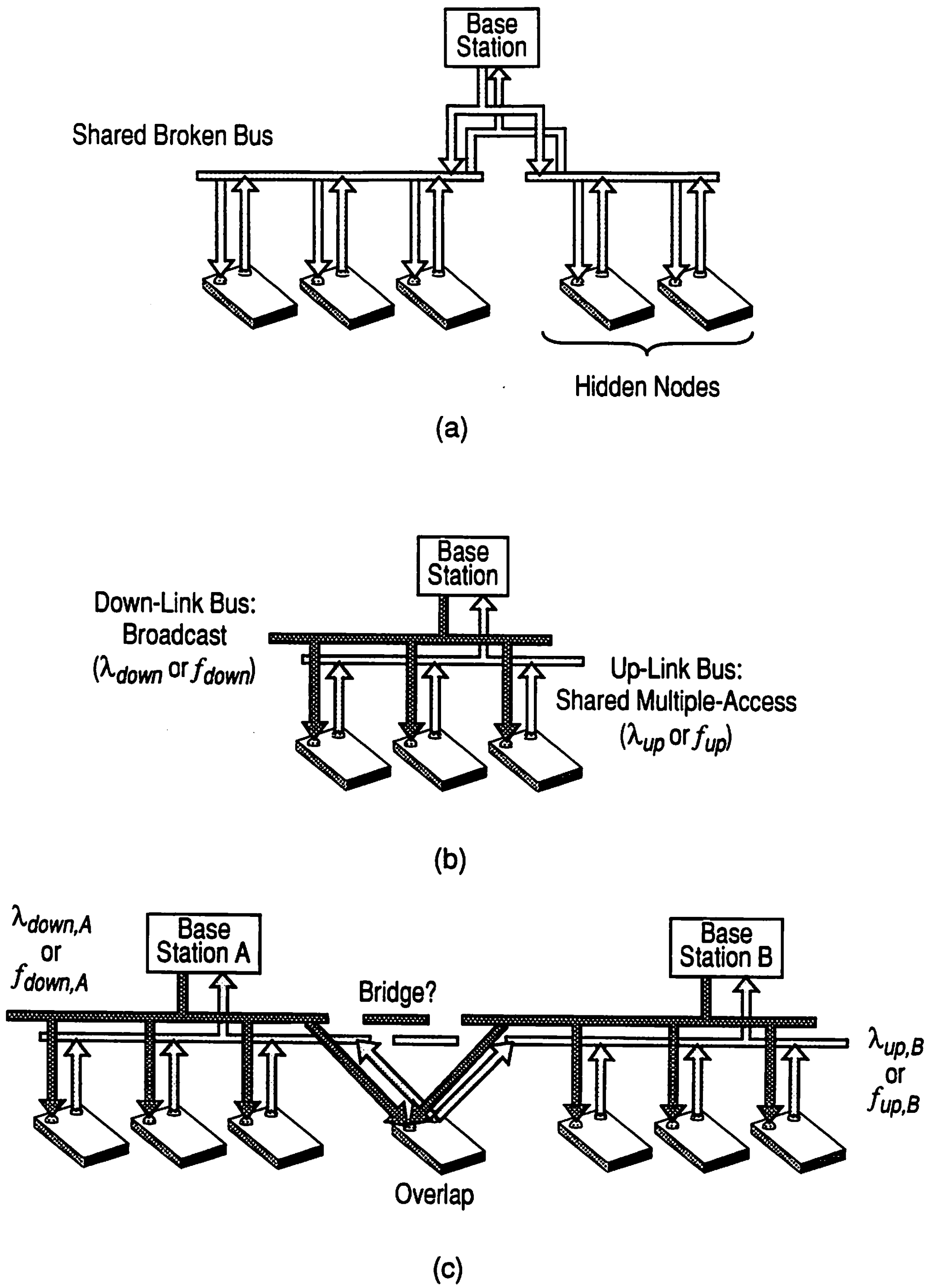


Fig. 6-1. Single-cell and multiple cell equivalent network architectures: (a) unconstrained with single cell; (b) wavelength duplex or subcarrier duplex with single cell; (c) wavelength duplex or subcarrier duplex with multiple cells.

network architecture, bypassing the problems of echo cancellation and near-end cross-talk. Furthermore, given the unlimited infrared bandwidth available, this isolation can be achieved without sacrificing capacity. (A similar strategy would be beneficial in radio- and wire-based networks if not for the significant capacity penalty incurred due to the limited amount of available bandwidth.)

6.2.2 Wavelength Duplex

The up-link transmissions can be isolated from the down-link transmissions by using different wavelengths for the up and down links, a technique referred to as *wavelength duplex*. In this configuration, the base station transmits at wavelength λ_{down}, and the portable units transmit at wavelength λ_{up}. Narrowband optical filters prevent interference, so that the equivalent architecture of a singe-cell network is the dual-unidirectional bus of Fig. 6-1-b. As before, carrier-sense protocols are still possible, provided that the base station retransmits all up-link signals on the down link.

It is desirable for two portable units to communicate without an intervening base station, a concept referred to as ad hoc networking. Unfortunately, ad hoc networking is quite costly when wavelength duplex is used, because each portable unit would need the capability to transmit and detect at two distinct wavelengths. An alternative to wavelength duplex that is better-suited to ad hoc networking is to use different subcarrier frequencies for the up and down links, a technique referred to as *subcarrier duplex*.

6.2.3 Subcarrier Duplex

In a subcarrier duplex system, the base stations and the portable units use the same optical wavelength, but up-link and down-link transmissions use different subcarrier frequency bands. For example, the up and down links could transmit at subcarrier frequencies f_{up} and f_{down}, respectively. For multiple-subcarrier systems, the up and down links could transmit at sets of subcarrier frequencies $\{f_{up,1}, \ldots, f_{up,N}\}$ and $\{f_{down,1}, \ldots, f_{down,N}\}$, respectively, but we use the f_{up} and f_{down} notation for simplicity. A baseband channel corresponds to a zero subcarrier frequency. Electrical filtering provides isolation between the up and down links. The network architecture is thus identical to that for wavelength duplex, as shown in Fig. 6-1-b; a dual unidirectional bus with the base station acting as a central controller.

To accommodate ad hoc networking, the receivers on the portable terminals must have the capability to tune to both f_{down} and f_{up}. Thus, in adopting the subcarrier duplex scheme over the unconstrained scheme of Fig. 6-1-a, we have added complexity to the physical layer for the benefit of simplifying the higher layers of the network. It should be noted, however, that this added receiver complexity is much less than that required to accommodate ad hoc networking on a wavelength duplex system.

In a subcarrier duplex system, the electrical bandwidth is divided into two sub-bands, with one sub-band allocated to the up link and the other to the down link. A drawback of subcarrier duplex is the reduced amount of bandwidth available to each link. Recall that, as mentioned in chapter 1, the bit-rate requirements for the up and down links will likely be highly asymmetric, since the down link carries executable files, graphics, and video images, whereas the up link carries only key strokes, pen strokes, and voice. In this case, only a small fraction of the electrical bandwidth should be allocated to the up link, leaving the majority for use by the down link.

There are two ways that carrier-sense protocols can be used with subcarrier duplex. The first, as in the wavelength duplex case, is to have the base station retransmit all uplink signals on the down link. The second is to incorporate specialized carrier-sense circuitry independent of the data bus: besides listening to subcarrier frequency f_{down} for data, the portable terminals could also listen to subcarrier frequency f_{up} to sense transmissions from other portable units. The carrier-sense circuitry need be nothing more than an energy detector, and thus will not increase appreciably the overall receiver complexity. Unfortunately, this latter approach is not always reliable, because some terminals may be hidden from others. A reliable carrier sense is not required, however, because the protocol accounts for collisions anyway.

6.3 OVERLAPPING CELLS

The discussion so far has considered a single cell only. Extensions to large areas requiring multiple base stations can be handled in a number of different ways. The simplest scenario is a building consisting of small rooms only, such as single- and double-occupant offices. In this case, a single base station is sufficient to cover each office, and interference between adjacent cells is prevented by the office walls. But what happens as a user moves from one office to another? There must either be a "dead zone" separating the two cells, presumably an undesirable condition, or the two cells must overlap. More generally, for large conference rooms, cafeterias, factory floors, and open office environments, there will not be opaque boundaries isolating neighboring cells from each other. If dead zones are not an acceptable solution, then the problems due to overlapping cells must be addressed.

In the most general cell-overlap scenario, j portable terminals can see m different base stations, and k portable terminals can be seen by n different base stations. In Fig. 6-1-c we illustrate cell overlap for the specific case of two base stations A and B, with one portable unit in the overlap region, and this portable unit is seen by both base stations and can see both base stations. We assume subcarrier duplex in each of the two cells, so that cell A up-link and down-link transmissions use subcarrier frequency $f_{up,A}$ and $f_{down,A}$, respectively, while cell B up-link and down-link transmissions use subcarrier frequency $f_{up,B}$ and $f_{down,B}$, respectively.

6.3.1 Unison Broadcast

A straightforward solution to the cell-overlap problem is to effectively expand the cell size to fill the entire room or building. This procedure, which we refer to as *unison broad-*

cast, can be accomplished be having all base stations transmit identical signals. All base stations transmit and receive at the same wavelength and subcarrier frequency, so that the multiple-cell topology collapses to one large effective cell. This unison approach was adopted by Gfeller in the design of the Infranet network [19]. In Fig. 6-1-c, the unison approach effectively bridges the gaps between neighboring up-link and down-link busses, forming two large up-link and down-link busses that permeate the entire network.

There are two drawbacks to the unison approach: first, because it doesn't exploit the network-layer/link-layer hierarchy that is inherent to a cellular network, it can be very inefficient for networks containing a large number of portable units. For example, a unison broadcast system broadcasts a signal intended for a single portable unit in every cell, whereas a more intelligent system would transmit that signal only in the cell containing the destination unit. The second drawback is the intersymbol interference due to the multipath propagation that results when a portable unit receives multiple delayed versions of the same signal from several different base stations, each with a different propagation delay.

This latter effect is not appreciable for small rooms or for low-speed links, and thus poses no problem to the Infranet network [19]. To illustrate its effect on a high-speed network, we simulated the effective impulse response from two base stations transmitting in unison to a single portable unit. We chose a room of moderate size that would likely require only two base stations, as shown in the inset of Fig. 6-2-a. The room is 24 m long, 12 m wide, with a 4 m ceiling. The two base stations are placed on the ceiling at the centers of the two cells formed by an imaginary wall that divides the room in half. Equating one of the floor corners with the origin, transmitter A is located at position (6 m, 6 m, 4 m) and transmitter B is located at position (18 m, 6 m, 4 m), and the receiver is located at position (0.1 m, 0.1 m, 1 m). Using the notation of chapter 4, the radiation pattern of both transmitters was a Lambertian sphere, the receiver was pointing straight up with a 90° field of view, the walls and ceiling reflectivities were 70%, and the floor reflectivity was 30%.

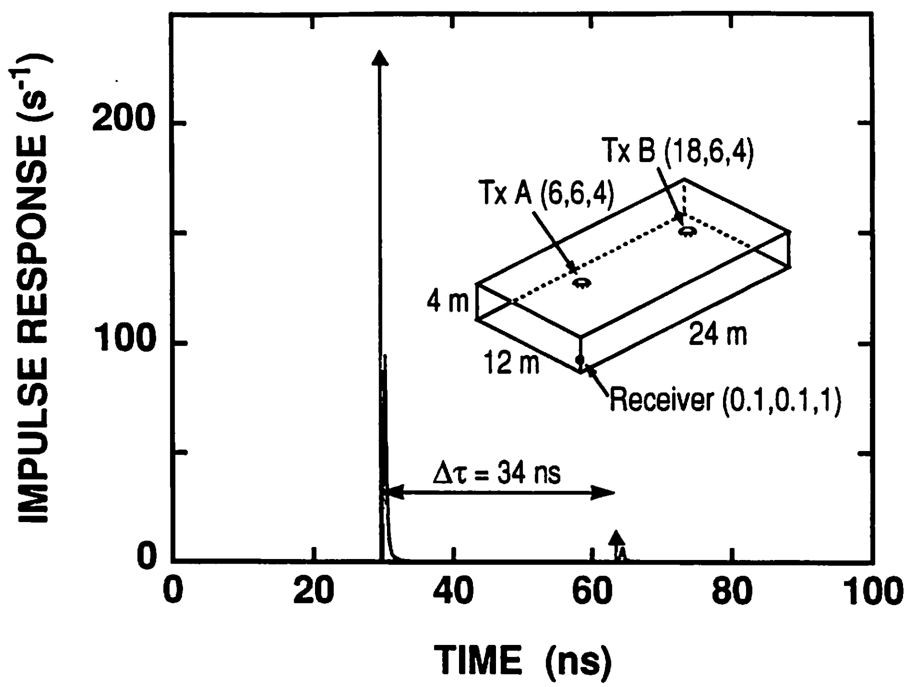
The walls, ceiling, and floor were subdivided into square elements of dimension 10 cm × 10 cm.

The impulse response was simulated using the method described chapter 4, accounting for single reflections only, and the results are shown in Fig. 6-2-a. The composite impulse response $h(t)$ is just the sum of the impulse responses from transmitter A to the receiver and transmitter B to the receiver: $h(t) = h_A(t) + h_B(t)$. The first Dirac impulse shown in the figure is due to the LOS path from transmitter A, and the second is due to the LOS path from transmitter B. The time difference between the two LOS impulses is $\Delta\tau = 34$ ns. This would certainly cause problems for data rates of 100 Mb/s and higher, for which the baud interval is 10 ns. Of course, the second impulse will not be detrimental if it is sufficiently attenuated relative to the first impulse. Define α as the attenuation of $h_B(t)$ relative to $h_A(t)$:

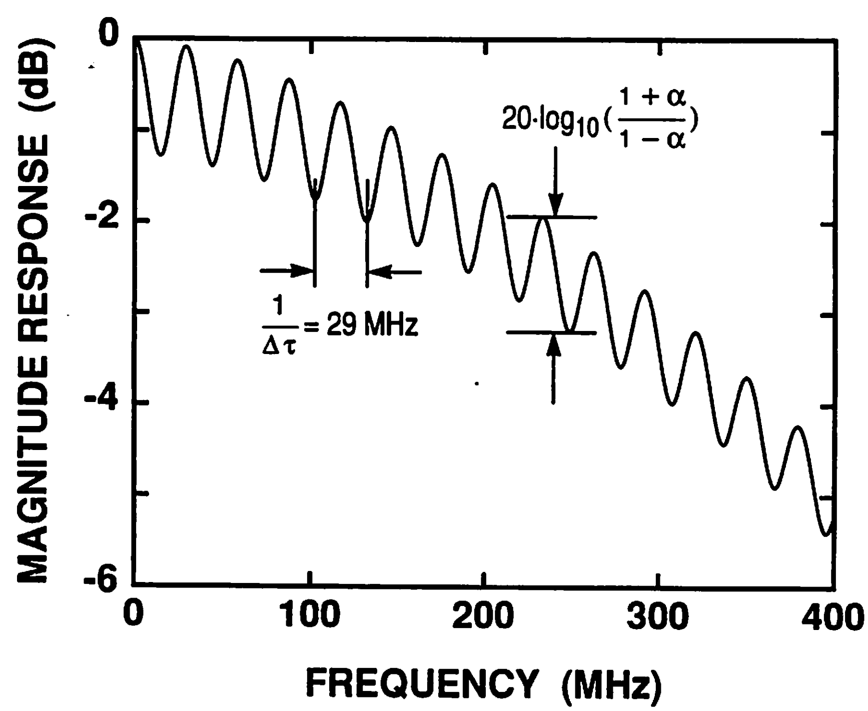
$$\alpha = \frac{\int h_B(t)\,dt}{\int h_A(t)\,dt} . \tag{6-1}$$

The magnitude response of the channel is shown in Fig. 6-2-b for frequencies up to 400 MHz, where we see that there is a severe ripple, or *frequency-selective fading*. The overall shape of the frequency response is governed by the non-LOS tails of $h_A(t)$ and $h_B(t)$, while the ripple is governed by the differential delay $\Delta\tau$ and attenuation α. The distance between peaks is approximately $1/\Delta\tau$ = 29 MHz, while the peak ripple is about $20 \cdot \log_{10}\left(\frac{1+\alpha}{1-\alpha}\right) = 1.3$ dB.

Note that $\Delta\tau$ and α are not independent; larger differential delays imply more path loss and thus smaller values for α. Therefore, a steeper ripple almost surely implies a larger spacing between ripple peaks, and conversely smaller spacing between ripple peaks implies a shallower ripple variation. Unfortunately, there are scenarios in which the undesirable condition of α near unity and $\Delta\tau$ large can occur; for example, when the LOS from the receiver to the nearest base station is obstructed, so that the initial signal is weak and of



(a)



(b)

Fig. 6-2. Impulse response and magnitude response for two base stations transmitting in unison.

comparable magnitude to the delayed signal from a distant base station. Another scenario is when the signals from a number of equidistant base stations combine coherently such that it is of comparable magnitude to the signal from the local base station. In such cases, the multipath frequency response will have a deep fade at a frequency of $1/\Delta\tau$.

The Infranet network, which uses unison broadcast, assumes a maximum differential delay of $\Delta\tau = 2.5$ μs, which corresponds to a frequency-selective fade at 400 kHz [19]. Since Infranet uses subcarrier frequencies up to about 500 kHz (it uses subcarrier duplex with f_{up} = 400 kHz and f_{down} = 200 kHz, with a symbol rate of 100 kbaud), a delay of $\Delta\tau = 2.5$ μs would in fact be catastrophic if α were near unity, although a delay as large as 1 μs would be harmless, regardless of α. For high-speed systems, however, when $1/\Delta\tau$ is not large relative to the electrical signal bandwidth, and when α may be near unity, the unison broadcast approach results in a harsh channel with severe frequency-selective fading. The strategies for fighting this problem are identical to those for fighting the multipath dispersion due to multiple refections as discussed in chapter 4, and include adaptive equalization and multiple-subcarrier modulation.

6.3.2 Independent Cells

Instead of unison broadcast, an alternative solution to the overlapping cell problem is to make the signals in any given cell be orthogonal to the signals in each of its neighboring cells. This can be accomplished in a number of ways, a few of which are outlined below.

- *Wavelength-division multiplexing* — Assign different wavelengths to neighboring cells.
- *Subcarrier-division multiplexing* — Assign different subcarrier frequencies to neighboring cells.
- *Code-division multiplexing* — Assign different pseudorandom code sequences to neighboring cells.

- *Time-division multiplexing* — Coordinate the timing of transmissions in neighboring cells so that simultaneous transmissions in two neighboring cells is impossible.

Wavelength multiplexing is highly impractical, because a portable unit roaming from cell to cell would need the capability to transmit and receive at every possible cell wavelength, complicating immensely the optics of the portable unit. Subcarrier multiplexing is more practical, since it is conceivable to construct a low-cost portable transceiver with the ability to transmit and receiver at N different subcarrier frequencies, where N is the frequency-reuse factor. The drawback of subcarrier-division multiplexing is its bandwidth inefficiency, which is important because high-capacitance photodiodes makes electrical bandwidth a scarce commodity (see chapter 3). Relative to the single-cell case, each cell in a subcarrier-division-multiplexed network with overlapping cells can use only a fraction $(1/N)$ of the total electrical bandwidth. In situations where overlapping cells are common, such as large open offices or factory floors, this bandwidth penalty may be an acceptable price to pay. In situations where overlapping cells are rare, however, such as in a network of single-occupant offices, the bandwidth penalty may not be too high a price to pay.

For the special case when multiple-subcarrier modulation schemes as described in chapter 5 are used, the bandwidth penalty due to subcarrier-division multiplexing described above can be alleviated by using all N subcarriers for non-overlapping cells, and using N/k subcarriers when k cells overlap. With this approach, terminals in non-overlapping cells can access higher data rates than terminals in overlapping cells. Alternatively, for the case when subcarriers are used to multiplex lower-speed signals to multiple users, non-overlapping cells can accommodate more terminals than overlapping cells.

Code-division multiplexing usually involves spread-spectrum modulation, which is not appropriate for high-speed applications because of its excessive electrical-bandwidth requirements. A discussion of code-division multiplexing for a low-speed (64 kb/s) system is given in [72].

In a traditional time-division multiplexing system, time is divided into frames, and each frame is subdivided into N slots, where $N-1$ is the number of nearest neighboring cells. Each cell is then assigned one of the time slots so that neighboring cells do not share the same slot. This approach has two primary disadvantages: first, it requires synchronization between portable units and the base stations, and second, it again results in an effective capacity penalty of $1/N$ relative to the single-cell case. A reservation scheme, which can be viewed as a form of time-division multiplexing, may be better-suited for preventing adjacent-cell interference. One such scheme, proposed by Demers for use in a nano-cellular network under development at Xerox PARC, requires each node to request and receive permission before transmitting any data. The protocol can be designed in such a way that neighboring base stations are never transmitting at the same time [73].

When multiplexing neighboring cells, the equivalent architecture for the two-base-station one-overlap case is shown in Fig. 6-1-c with no bridge between neighboring up and down links. For the specific case of subcarrier multiplexing, we have $f_{A,up} \neq f_{B,up}$ and $f_{A,down} \neq f_{B,down}$. In any case, the portable unit in the overlap region will receive orthogonal signals from two different base stations. Likewise, the same portable unit will transmit the same signal to two different base stations.

In the general case of multiplexed neighboring cells with cell overlap, the portable unit should monitor each of the N orthogonal signal subspaces, determine which base station is transmitting the cleanest signal, and keep that base station notified of its presence. The portable unit should transmit signals in the signal subspace expected by its base station. In this way, as a portable unit roams out of one cell and into another, it stops communicating with the base station of the old cell, and starts communicating with the base station of the new cell, a process referred to as *cell hand-off.* Thus, each portable is heard by only one base station at any given moment.

6.4 SUMMARY

The physical-layer link as designed in chapter 2 results in nearly omnidirectional transmitters and receivers. The relationship between the physical layer and higher layers for an infrared-based wireless network are thus similar to those for a micro-cellular radio-based network. The designer of an infrared network has more degrees of freedom, however, because of the immense bandwidth of infrared, and also because infrared links are less susceptible to adjacent-cell interference.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

The dissertation has addressed the general problem of designing a high-speed wireless link using non-directed infrared radiation. The results suggest that speeds near 100 Mb/s are practical. The only true test of this hypothesis, of course, is to build such a link and verify it experimentally. This dissertation has laid the groundwork for such an experiment, and perhaps has convinced the reader that its outcome will be successful as well.

7.1 CONCLUSIONS

The important conclusion to draw from the link analysis of chapter 2 is that a lot can be gained through careful design of the transmitter and receiver optics. Although from a systems standpoint it is a useful conceptualization to view the baseband system from modulating input at the transmitter to photodetector output at the receiver as a black box, the system designer should not lose sight of the fact that the path loss and noise properties of this box can be controlled through careful design. For example, the hemispherical lens can provide an optical gain of approximately n^2 at all angles of incidence, where n is its refractive index. The hemispherical thin-film filter proposed in chapter 2 was shown to outperform a planar filter by providing both a narrow bandwidth and a wide field-of-view. Reflection losses can be minimized through the use of index matching gels and antireflection coatings. Finally, to make the best use of a given transmitted signal power, it is beneficial to design the optical gain of the transmitter and receiver jointly, not separately.

The f^2 noise power in a wideband preamplifier is roughly proportional to the square of the photodetector surface area; this fact complicates low-noise receiver design for non-directional links, which require large-area detectors. The results of chapter 3 indicate that the preamplifier can be designed so that the shot noise due to background light, a fundamental noise source that will always be present, is dominant. This simplifies other aspects of system design, such as budget analysis and modulation design, because it allows the electrical noise in the preamplifier to be neglected.

Intersymbol interference due to multipath dispersion is a fundamental impediment to high-speed non-directed communication. Chapter 4 presented a Lambertian model for multipath optical propagation accounting for multiple reflections. Simulations based on this model were found to agree with experimental measurements. The high-order bounces were found to carry significant signal energy, affecting both the power budget and the intersymbol interference. The importance of this theory lies not in the capability to simulate multipath dispersion, although this is an important tool in system design. Rather, the agreement between the theory and experiment suggests that the physical processes underlying multipath optical propagation can be well-approximated by the simple Lambertian model. This conceptual tool will likely prove more useful than the simulation tool in the long run.

Chapter 5 defined the intensity-modulation channel as an additive white Gaussian noise channel with two constraints on its input: the input cannot be negative, and the average amplitude of the input cannot exceed a given maximum value. This definition is applicable to all systems using intensity modulation and direct detection, including fiber-optic systems. (It should be noted, however, that a fiber-optic transmitter may have limitations on its peak power as well as its average power.) We evaluated the bandwidth efficiency and power efficiency of a number of modulation schemes on this channel. The results indicate that modulation schemes well-suited for a conventional AWGN channel are not necessarily well-suited for the intensity-modulation channel. In terms of power efficiency and bandwidth efficiency, baseband modulation schemes such as OOK, L-PAM, and L-PPM were found to significantly outperform subcarrier and multiple-subcarrier modulation schemes.

7.2 FUTURE WORK

An obvious conclusion to draw from the results of this dissertation is that there remains much more work to be done. The following sections suggests areas for future research in non-directed wireless communication.

7.2.1 Link Analysis

The link analysis and optics design presented in chapter 2 was based on the non-directed LOS configuration of Fig. 1-1-e. This configuration represents a compromise between the convenience of non-directed links with the reduced-path-loss advantages of the LOS links. Future research should address the practical problem of implementing the optimal radiation pattern derived in section 2.4.4. Whether some combination of discrete sources, lambertian reflectors, mirrors and lenses is sufficient to approximate the optimal pattern is an open question and warrants further investigation.

The major drawback of relying on an unobstructed LOS path, obviously, is the inconvenience posed to the user of the portable receiver. The diffuse (non-directed non-LOS) configuration of Fig. 1-1-f is much preferable from the user's standpoint. Therefore, a complete link analysis and design for the diffuse channel, analogous to that in chapter 2, is needed for the diffuse configuration. While the advantages of the hemispherical lens and hemispherical will undoubtedly remain, the optimal transmitter radiation pattern for a diffuse link will likely by quite different. A complete analysis would take into account the losses due to partial shadowing in addition to path losses, reflection losses, tilt losses, and filter losses.

7.2.2 Receiver Design

The results of chapter 3 indicate that a wide-band low-noise preamplifier can be designed without introducing noise stronger than the shot noise due to background illumination. The implication of this result is that the receiver design can be effectively decoupled from the development of the rest of the system. At the circuit-design level, however, there remain a number of challenges, most of which stem from the requirements of low power consumption and low cost that accompany a portable receiver. For example, while the FET devices assumed in chapter 3 were adequate to provide shot-noise limited operation, it is likely that bipolar transistors would be as well. In fact, when compared to an FET, the higher transconductance offered by a bipolar transistor may outweigh its disadvantages of higher base-current shot noise and higher capacitance [44], leading to a more cost-effective design. Future work in preamplifier design should explore other transimpedance configurations using bipolar as well as FET devices in search for a design that not only meets the noise and bandwidth requirements, but also minimizes some suitably defined cost function that accounts for power consumption and monetary cost.

7.2.3 Multipath Dispersion

The algorithm presented in chapter 4 for simulating the multipath impulse response was simple to describe but not very efficient; it may take several days to simulate one impulse response to a high degree of accuracy. The algorithm is recursive and therefore difficult to parallelize. Future work may look into ways to speed up the simulation process.

Considerable speed up is needed before the simulation tool can compete with the speed of the experimental means for measuring multipath described in [56].

One aspect of multipath that is difficult to deal with experimentally is the effect of non-Lambertian transmitter radiation pattern; an arbitrarily shaped radiation pattern can be easily simulated, but it cannot be easily implemented. The multipath characterization of chapter 4 assumed a Lambertian radiation pattern, while the link analysis of chapter 2 assumed an optimal radiation pattern. Future work should examine the effects of non-Lambertian radiation patterns on multipath dispersion.

Despite the advantages of simulation, many issues can only be resolved through experimentation. For example, the design of an adaptive equalizer must take into account the speed with which the impulse response varies, which is best determined experimentally. The effects of shadowing and tilt, especially in diffuse configurations, are also best determined experimentally. In a related area, better characterization of the background noises due to sunlight, incandescent light, and fluorescent light is needed.

7.2.4 Modulation

Further study in modulation for a wireless optical link should concentrate on examining in detail some of the issues touched on in chapter 5. For example, the effect of multipath interference on the performance of pulse-position modulation schemes should be addressed, along with the prospects for adaptive equalization. Modifications of the L-PPM scheme may be beneficial in this regard. For example, a weight-m L-PPM scheme, in which m chips are on during each baud interval instead of just one, would improve multipath immunity at the expense of an increased bandwidth requirement. An interesting problem in information theory is to calculate the Shannon capacity of the intensity-modulation channel defined in chapter 5. This may lead to useful insight into the signal set design problem. Finally, the viability of coherent detection in non-directed applications is an intriguing question, although perhaps an academic one.

7.2.5 Systems Issues

The scope of this thesis was limited to the physical layer problem of establishing a high-speed point-to-point link using non-directed infrared radiation. Widening the scope to include the design of a wireless LAN, as discussed in chapters 1 and 6, opens up an abundance of new problems too numerous to mention here. A complete design of a wireless LAN based on non-directed infrared links would include the design of multiple-access and cell hand-off protocols. When designing these protocols, the constraints imposed by a battery-powered terminal, such as the desire to power down (and stay powered down) when idle, should be taken into account.

REFERENCES

[1] D. C. Cox, "A Radio System Proposal for Widespread Low-Power Tetherless Communications," *IEEE Transactions on Communications*, vol. 39, No. 2, pp. 324-335, February 1991.

[2] D. J. Goodman, "Evolution of Wireless Information Networks," *European Transactions on Telecommunications and Related Technologies*, Vol. 2, No. 1, pp. 104-113, January-February 1991.

[3] A. Ramirez, "Mapping Out the Wireless-Phone Future," *The New York Times*, p. 1, November 12, 1992.

[4] B. Tuch, "An ISM Band Spread Spectrum Local Area Network: WaveLAN," *Proceedings of IEEE Workshop on Wireless Local Area Networks*, Worcester Polytechnic Institute, MA, pp. 103-111, May 1991.

[5] P. R. Chevillat, H. R. Rudin, "Guest Editorial," *IEEE Network Magazine*, pp. 10, November 1991.

[6] C.-S. Yen and R. D. Crawford, "The Use of Directed Optical Beams in Wireless Computer Communications," *IEEE Globecom '85*, pp. 1181-1184, December 2-5, 1985.

[7] T. S. Chu, M. J. Gans, "High Speed Infrared Local Wireless Communication," *IEEE Communications Magazine*, Vol. 25, No. 8, pp. 4-10, August 1987.

[8] G. Berline and E. Perratore, "Wireless LANs," *PC Magazine*, February 1992.

[9] N. Baran, "Wireless Networking," *BYTE*, pp. 291-294, April 1992.

[10] "Photolink User's Manual," Photonics Corp., San Jose, CA 95131.

[11] F. R. Gfeller and U. H. Bapst, "Wireless In-House Data Communication via Diffuse Infrared Radiation," *Proceedings of the IEEE*, Vol. 67, No. 11, pp. 1474-1486, November 1979.

[12] G. Yun and M. Kavehrad, "Spot Diffusing and Fly-Eye Receivers for Indoor Infrared Wireless Communications," *Proc. IEEE International Conference on Selected Topics in Wireless Communications*, Vancouver, B. C., pp. 262-265, June 1992.

[13] T. Minami, K. Yano, and T. Touge, "Optical Wireless Modem for Office Communication," *National Computer Conference*, pp. 721-728, 1983.

[14] O.Takahashi and T. Touge, "Optical Wireless Network for Office Communication," *JARECT (Japan Electron. Rev. Electron. Comput. Telecomm.)*, Vol. 20, pp. 217-228, 1985/1986.

[15] Y. Nakata, J. Kashio, T. Kojima, and T. Noguchi, "In-House Wireless Communication System Using Infrared Radiation," *Proceedings of the International Conference on Computer Communication*, pp. 333-337, 1984.

[16] T. Fuji and Y. Kikkawa, "Optical Space Transmission Module," *National Technical Report*, Vol. 34, No. 1, February 1988.

[17] F. R. Gfeller, H. R. Müller, and P. Vettiger, "Infrared Communication for In-House Applications," *IEEE COMPCON*, Washington D. C., pp. 132-138, September 5-8, 1978.

[18] I. A. Parkin and J. Zic, "An Application of Infra-red Communications," *Journal of Electrical and Electronics Engineering, Australia - IE Aust. & IREE Aust.*, Vol. 4, No. 4, pp. 331-336, December 1984.

[19] F. R. Gfeller, "Infranet: Infrared Microbroadcasting Network for In-House Data Communication, "*IBM Research Report*, RZ 1068-38619, April 1991.

[20] Photonics Corporation, promotional literature, San Jose, CA 95131.

[21] Spectrix Corp., Evanston, IL 60201.

[22] D. R. Pauluzzi, P. R. McConnell, and R. L. Poulin, "Free-Space Undirected Infrared Voice and Data Communications with a Comparison to RF Systems," *Proc. IEEE International Conference on Selected Topics in Wireless Communications*, Vancouver, B. C., pp. 279-285, June 1992.

[23] R. L. Poulin, D. R. Pauluzzi, and M. R. Walker, "A Multi-Channel Infrared Telephony Demonstration System for Public Access Applications," *Proc. IEEE International Conference on Selected Topics in Wireless Communications*, Vancouver, B. C., pp. 286-291, June 1992.

[24] J. R. Barry, J. M. Kahn, E. A. Lee, and D. G. Messerschmitt, "High-Speed Nondirective Optical Communication for Wireless Networks," *IEEE Network Magazine*, pp. 44-54, November 1991.

[25] M. D. Kotzin, "Short-Range Communications Using Diffusely Scattered Infrared Radiation," Ph.D. Dissertation, Northwestern University, 1981.

[26] M. E. Marhic, M. D. Kotzin, and A. P. van den Heuvel, "Reflectors and Immersion Lenses for Detectors of Diffuse Radiation," *Journal of the Optical Society of America*, Vol. 72, No. 3m pp. 352-355, March 1982.

[27] M. D. Kotzin and A. P. van den Heuvel, "A Duplex Infra-Red Systems for In-Building Communications," *IEEE VTC '86*, pp. 179-185, 1986.

[28] G. Smestad, H. Ries, R. Winston, and E. Yablonovitch, "The thermodynamic limits of light concentrators," *Solar Energy Materials*, Vol. 21, pp. 99-111, 1990.

[29] H. A. Ankermann, "Transmission of Audio Signals by Infrared Light Carrier," *SMPTE Journal*, Vol. 89, pp. 834-837, November 1980.

[30] S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, John Wiley & Sons, New York, 1984.

[31] H. A. Macleod, *Thin-Film Optical Filters*, Adam Jilger Ltd, London, 1969.

[32] J. D. Rancourt, *Optical Thin Films*, Macmillan Publishing, New York, 1987.

[33] A. Yariv, *Optical Electronics*, Holt, Rinehart, and Winston, New York, 1985.

[34] Melles Griot Product Catalog, 1991.

[35] OCLI Product Catalog, 1990.

[36] E. D. Palik, ed., *Handbook of Optical Constants of Solids*, Orlando, Academic Press, 1985.

[37] T. V. Muoi, "Optical Receivers," Chapter 16 in *Optoelectronic Technology and Lightwave Communications Systems*, ed. C. Lin, pp. 441-472, Van Nostrand Reinhold, 1989.

[38] S. D. Personick, "Receiver Design for Digital Fiber Optic Communications Systems, I and II," *Bell System Technical Journal*, Vol. 52, No. 6, pp. 843-886, July-August 1973.

[39] S. D. Personick, "Receiver Design for Optical Fiber Systems," *Proceedings of the IEEE*, Vol. 65, No. 12, pp. 1670-1678, December 1977.

[40] J. E. Goell, "Input Amplifiers for Optical PCM Receivers," *Bell System Technical Journal*, Vol. 53, No. 9, pp. 1771-1793, November 1974.

[41] B. L. Kasper, "Receiver Design," Chapter 18 in *Optical Fiber Telecommunications II*, ed. S. E. Miller and I. P. Kaminow, pp. 689-723, Academic Press, San Diego, 1988.

[42] T. V. Muoi, "Receiver Design for High-Speed Optical- Fiber Systems," *Journal of Lightwave Technology*, Vol. LT-2, No. 3, pp. 243-267, June 1984.

[43] M. Aiki, "Low-Noise Optical Receiver for High-Speed Optical Transmission," *Journal of Lightwave Technology*, Vol. LT-3, No. 6, pp. 1301-1306, December 1985.

[44] M. Brain and T. P. Lee, "Optical Receivers for Lightwave Communication Systems," *Journal of Lightwave Technology*, Vol. LT-3, No. 6, pp. 1281-1300, December 1985.

[45] P. R. Gray, R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, John Wiley & Sons, New York, 1993.

[46] R. G. Smith, C. A. Brackett, H. W. Reinbold, "Optical Detector Package," *Bell System Technical Journal*, Vol. 57, No. 6, pp. 1809-1822, July-August 1978.

[47] A. B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design*, John Wiley & Sons, New York, 1984.

[48] J. M. Kahn, J. R. Barry, M. D. Audeh, E. A. Lee, D. G. Messerschmitt, "Design of High-Speed Wireless Links Using Non-Directional Infrared Radiation," in *Wireless Communications: Future Directions*, Kluwer Academic Publishers, Boston, 1992.

[49] G. Marsh, "Experimental Demonstration of High-Speed Nondirectional Wireless Infrared Communication," Ph.D. Dissertation, U. C. Berkeley, expected 1994.

[50] B. C. Kuo, *Automatic Control Systems*, Prentice Hall, Englewood Cliffs, N. J., 1982.

[51] J. R. Barry, J. M. Kahn, W. J. Krause, E. A. Lee, D. G. Messerschmitt, "Simulation of Multipath Impulse Response for Indoor Wireless Optical Channels," accepted in *IEEE Journal of Selected Areas in Communications*, June 1992.

[52] J. R. Barry, J. M. Kahn, E. A. Lee, D. G. Messerschmitt, "Simulation of Multipath Impulse Response for Indoor Diffuse Optical Channels", *Proc. IEEE Workshop on Wireless Local Area Networks*, Worcester MA, May 9-10, 1991, pp. 81-87.

[53] D. Hash, J. Hillery, and J. White, "IR RoomNet: Model and Measurement," *IBM Communication ITL Conference*, June 1986.

[54] P. Hortensius, "Research and Development Plan of the Infrared Portable Data Link," IBM T. J. Watson Research Center, Yorktown Heights, New York, January 4, 1990. Internal Report.

[55] A. S. Glassner, ed., *An Introduction to Ray Tracing*, Academic Press, San Diego, 1989.

[56] W. J. Krause, "Experimental Characterization of Non-Directed Indoor Infrared Channels," Master's Report, University of California, Berkeley, expected December 1992.

[57] E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer Academic Publishers, Boston, 1988.

[58] American National Standards Institute, *American National Standard for the Safe Use of Lasers*, ANSI Z136.1-1986.

[59] S. Haykin, *Digital Communications*, John Wiley & Sons, New York, 1988.

[60] J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1983.

[61] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1984.

[62] M. D. Audeh, "Performance Evaluation of Baseband OOK for Wireless Indoor Infrared LANs Operating at 100 Mb/s," Master's Report, U. C. Berkeley, December 1992.

[63] J. M. Kahn, J. R. Barry, W. J. Krause, M. D Audeh, J. B. Carruthers, G. W. Marsh, E. A. Lee, and D. G. Messerschmitt, "High-Speed Non-Directional Infrared Communication for Wireless Local-Area Networks," *Proceedings of 26th Asilomar Conference on Signals, Systems and Computers*, October 26-28, 1992, Pacific Grove, CA.

[64] T. E. Darcie, "Subcarrier Multiplexing for Lightwave Networks and Video Distribution Systems," *IEEE Journal of Selected Areas in Communications*, Vol. 8, No. 7, pp. 1240-1248, September 1990.

[65] W. R. Leeb, "Degradation of Signal-to-Noise Ratio in Optical Free-Space Data Links due to Background Illumination," *Applied Optics*, Vol. 28, No. 15, pp. 3443-3449.

[66] J. R. Barry and E. A. Lee, "Performance of Coherent Optical Receivers," *Proceedings of the IEEE*, Vol. 78, No. 8, August 1990, pp. 1369-1394.

[67] G. Gould, S. F. Jacobs, J. T. LaTourrette, M. Newstein, and P. Rabinowitz, "Coherent Detection of Light Scattered from a Diffusely Reflecting Surface," *Applied Optics*, Vol. 3, No. 5, pp. 648-649, May 1964.

[68] R. H. Kingston, *Detection of Optical and Infrared Radiation,* Springer Verlag, Berlin, 1976.

[69] A. E. Siegman, "The Antenna Properties of Optical Heterodyne Receivers," *Proceedings of the IEEE*, Vol. 54, No. 10, October 1966.

[70] J. J. Degnan and B. J. Klein, "Optical Antenna Gain 2: Receiving Antennas," *Applied Optics*, Vol. 13, No. 9, pp. 2397-2401, 1974.

[71] A. Lessard and M. Gerla, "Wireless Communications in the Automated Factory Environment," *IEEE Network Magazine*, Vol. 2, No. 3, pp. 64-69, May 1988.

[72] Y. Yamauchi, M. Sato, and T. Namekawa, "In-House Wireless Optical Digital SSMA," *Electronics and Communications in Japan*, Part 1, Vol. 70, No. 6, pp. 87-101, 1987.

[73] A. Demers, "Protocols for the PARC Nanocellular Network," seminar at U.C. Berkeley, November 1992.

Appendix A: POWER EFFICIENCY ON THE LINEAR GAUSSIAN-NOISE CHANNEL

In chapter 5 we defined an intensity-modulation channel with two constraints on its input $X(t)$: first, $X(t)$ must be positive everywhere, and second, the average amplitude of $X(t)$ must not exceed a value of P_{avg}. We derived the bandwidth efficiency and power efficiency for a number of modulation schemes on the intensity modulation channel. We then defined a "conventional" channel by omitting the first constraint, allowing negative values of $X(t)$, and by changing the second constraint from an amplitude constraint to a power constraint, so that the average *power* of $X(t)$ cannot exceed P_X. With the aid of Fig. 5-4 and Fig. 5-5, the bandwidth and power efficiencies of the intensity-modulation channel were shown to differ both quantitatively and qualitatively from their conventional-channel counterparts. The purpose of this appendix is to present the equations used to arrive at Fig. 5-5, and to clarify the assumptions made in the process.

The bandwidth requirements for the conventional channel are identical to those for the intensity-modulation channel, hence we need consider only power efficiency here. In all cases we assume that the received signal is $Y(t) = X(t) + n(t)$, where $n(t)$ is a white Gaussian noise with PSD N_0, and we assume ideal maximum-likelihood detection.

L-PAM

Assume that the transmitted signal for L-PAM is given by:

$$X(t) = A \sum_{k=-\infty}^{\infty} a_k p(t-kT) \ , \qquad \text{(A-1)}$$

where A is the signal amplitude, $\{a_k\}$ is the white symbol sequence with alphabet defined by (5-20), and $p(t)$ is the rectangular pulse shape given by (5-5). Recall that both the symbols $\{a_k\}$ and the pulse shape $p(t)$ have a maximum value of unity. To make the average power of $X(t)$ equal to P_X, the amplitude must be $A = \sqrt{\frac{L-1}{L+1} 3P_X}$. For a given bit rate, the BER for this system can be found by substituting $P_{avg} = A$ into (5-24), yielding [57][60]:

$$BER \approx Q\left(\sqrt{\frac{3\log_2 L}{L^2-1}}\sqrt{\frac{P_X}{N_0 R_b}}\right). \tag{A-2}$$

The power requirement for L-PAM is therefore:

$$P_{X,req} = \frac{L^2-1}{3\log_2 L} P_{X,2\text{-}PAM} \tag{A-3}$$

where $P_{X,2\text{-}PAM} \equiv N_0 R_b Q^{-1}(BER)^2$ is the power requirement for 2-PAM. This parameter will be the benchmark for the conventional channel, in the same way that P_{OOK} was the benchmark for the intensity modulation channel. Comparing (A-3) with (5-25), we see that, as L increases, the power penalty due to multilevel PAM grows faster for the conventional channel than for the intensity-modulation channel.

N-L-QAM

Assume that the transmitted signal for N-L-QAM is given by:

$$X(t) = A \sum_{k=-\infty}^{\infty} a_k p(t-kT) \sum_{n=1}^{N} \left\{a_{I,n}[k]\cos(\omega_n t) + a_{Q,n}[k]\sin(\omega_n t)\right\}, \tag{A-4}$$

where A is the signal amplitude, N is the number of subcarriers, $\{a_{I,n[k]} + ja_{Q,n[k]}\}$ is the complex-valued white symbol sequence for the n-th L-QAM signal with alphabet defined by (5-26), $p(t)$ is the rectangular pulse shape given by (5-5), and ω_n is the n-th subcarrier frequency. To make the average power of $X(t)$ equal to P_X, the amplitude must be $A = \sqrt{\frac{\sqrt{L}-1}{\sqrt{L}+1} \cdot \frac{6P_X}{N}}$. Substituting $P_{avg}/N = A$ into (5-34) yields [57][60]:

$$BER \approx Q\left(\sqrt{\frac{3\log_2 L}{2(L-1)}}\sqrt{\frac{P_X}{N_0 R_b}}\right). \tag{A-5}$$

The power requirement for N-L-QAM is therefore:

$$P_{X,req} = \frac{2(L-1)}{3\log_2 L} P_{X,2\text{-}PAM} . \tag{A-6}$$

Interestingly, the power penalty is independent of N, the number of subcarriers. In other words, there is no power penalty incurred by using multiple carriers on a conventional channel. This was not the case for the intensity-modulation channel.

L-PPM

Assume that the transmitted signal for L-PPM is given by:

$$X(t) = B + A \sum_{k=-\infty}^{\infty} a_k p_{l[k]}(t - kT) , \tag{A-7}$$

where $l[k] \in \{1, 2, \ldots, L\}$ denotes the position of the "on" chip during the k-th baud interval $t \in [(k-1)T, kT]$, and where $\{p_l(t)\}$ is the family of pulse shapes defined by (5-39). The offset B is chosen as $B = -A/L$ so that $X(t)$ has zero mean, an optimal condition [57]. Under this assumption, the signal amplitude must be $A = \sqrt{\frac{LP_X}{1-1/L}}$ to make the average power of $X(t)$ equal P_X. Substituting $LP_{avg} = A$ into (5-54) yields:

$$BER \approx Q\left(\sqrt{\frac{L \cdot \log_2 L}{2(L-1)}} \sqrt{\frac{P_X}{N_0 R_b}}\right) . \tag{A-8}$$

The power requirement for L-PPM is therefore:

$$P_{X,req} = \frac{2(L-1)}{L \cdot \log_2 L} P_{X,2\text{-}PAM} . \tag{A-9}$$

Comparing with (5-55), the increased power efficiency of L-PPM is not as dramatic for the conventional channel as it was for the intensity-modulation channel.