INCLUDES

NEWNES ONLINE MEMBERSHIP

WIRELESS NETWORKING

know it all

- **A comprehensive overview from our best-selling authors**
- **Explains the theory, concepts, design, and implementation of 802.11, 802.16, and 802.20**
- **Includes discussion of indoor networks, signal propagation, and network security**

Chandra • Dobkin • Bensky
Olexa • Lide • Dowla

Wireless Networking

Newnes Know It All Series

PIC Microcontrollers: Know It All
Lucio Di Jasio, Tim Wilmshurst, Dogan Ibrahim, John Morton,
Martin Bates, Jack Smith, D.W. Smith, and Chuck Hellebuyck
ISBN: 978-0-7506-8615-0

Embedded Software: Know It All
Jean Labrosse, Jack Ganssle, Tammy Noergaard, Robert Oshana, Colin Walls, Keith Curtis,
Jason Andrews, David J. Katz, Rick Gentile, Kamal Hyder, and Bob Perrin
ISBN: 978-0-7506-8583-2

Embedded Hardware: Know It All
Jack Ganssle, Tammy Noergaard, Fred Eady, Creed Huddleston, Lewin Edwards,
David J. Katz, Rick Gentile, Ken Arnold, Kamal Hyder, and Bob Perrin
ISBN: 978-0-7506-8584-9

Wireless Networking: Know It All
Praphul Chandra, Daniel M. Dobkin, Alan Bensky, Ron Olexa,
David A. Lide, and Farid Dowla
ISBN: 978-0-7506-8582-5

RF & Wireless Technologies: Know It All
Bruce Fette, Roberto Aiello, Praphul Chandra, Daniel M. Dobkin,
Alan Bensky, Douglas Miron, David A. Lide, Farid Dowla, and Ron Olexa
ISBN: 978-0-7506-8581-8

For more information on these and other Newnes titles visit: www.newnespress.com

Wireless Networking

Praphul Chandra
Daniel M. Dobkin
Alan Bensky
Ron Olexa
David A. Lide
Farid Dowla

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Newnes is an imprint of Elsevier

Cover image by iStockphoto
Newnes is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

Library of Congress Cataloging-in-Publication Data
A catalogue record for this book is available from the British Library.

British Library Cataloguing-in-Publication Data

Chandra, Praphul.
Wireless networking / Praphul Chandra, Ron Olexa, Alan Bensky.
p. cm. -- (The Newnes know it all series)
Includes index.
ISBN-13: 978-0-7506-8582-5 (pbk. : alk. paper) 1. Wireless communication systems. I. Olexa, Ron. II. Bensky, Alan, 1939- III. Title.
TK5103.2.C447 2007
621.384--dc22

2007029327

ISBN: 978-0-7506-8582-5

For information on all Newnes publications
visit our Web site at www.books.elsevier.com

07 08 09 10 10 9 8 7 6 5 4 3 2 1

Printed in the United States of America

Contents

This page intentionally left blank

About the Authors

Alan Bensky, MScEE (Chapters 4, 5, 6, and 15), is an electronics engineering consultant with over 25 years of experience in analog and digital design, management, and marketing. Specializing in wireless circuits and systems, Bensky has carried out projects for varied military and consumer applications. He is the author of *Short-range Wireless Communication, Second Edition*, published by Elsevier, 2004, and has written several articles in international and local publications. He has taught courses and gives lectures on radio engineering topics. Bensky is a senior member of IEEE.

Praphul Chandra (Chapters 10, and 11) works as a Research Scientist at HP Labs, India in the Access Devices Group. He joined HP Labs in April 2006. Prior to joining HP he was a senior design engineer at Texas Instruments (USA) where he worked on Voice over IP with specific focus on Wireless Local Area Networks. He is the author of two books – *Bulletproof Wireless Security* and *Wi-Fi Telephony: Challenges and Solutions for Voice over WLANs*. He is an Electrical Engineer by training, though his interest in social science and politics has prompted him to concurrently explore the field of Public Policy. He maintains his personal website at www.thecofi.net.

Daniel M. Dobkin (Chapter 1, 2, 3, and 9) is the author of *RF Engineering for Wireless Networks*. He has been involved in the design, fabrication, and characterization of devices and systems used in microwave wireless communications for over two decades. He is currently an independent consultant involved in research and teaching related to RFID and other fields in communications. He has taught numerous introductory short courses in RFID technology in the US and Singapore. Dr. Dobkin received his Ph.D. degree from Stanford University in 1985 and his B.S. from the California Institute of Technology in 1976. He is the author of about 30 technical publications, inventor or co-inventor of six U.S. patents, and has written two technical books: Principles of Chemical Vapor Deposition with Michael Zuraw and RF Engineering for Wireless Networks.

Farid Dowla (Chapters 12, 13, and 14) is the editor of *Handbook of RF & Wireless Technologies.* Dowla received his BS, MS, and PhD in electrical engineering from the Massachusetts Institute of Technology. He joined Lawrence Livermore National Laboratory

shortly after receiving his doctorate in 1985. His research interests include adaptive filters, signal processing, wireless communication systems, and RF/mobile communication. He currently directs a research team focused on ultra-widebandRFradar and communication systems. Dowla is also an adjunct associate professor of electrical engineering at the University of California at Davis. He is a member of the Institute of Electrical and Electronic Engineers (IEEE) and Sigma Xi. He holds three patents in signal processing area, has authored a book on neural networks for the U.S. Department of Defense, and has edited a book on geophysical signal processing. He contributes to numerous IEEE and professional journals and is a frequent seminar participant at professional conferences.

David A. Lide (Chapter 11) is the author of *Wi-Fi Telephony*. He currently is a Senior Member of the Technical Staff at Texas Instruments and has worked on various aspects of Voice over IP for the past nine years. Prior to that, he has worked on Cable Modem design and on weather satellite ground systems. He lives with his family in Rockville, Maryland.

Michael R. Moore (Chapter 13) was a contributor to *Handbook of RF & Wireless Technologies*. He is a research and development engineer in the Engineering and Science Technology Division at Oak Ridge National Laboratory. He holds a BS and MS in electrical engineering from Mississippi State University in Starkville, Miss. His current research expertise includes 16 years in RF instrumentation, health effects, and communications. He has several years of experience in shielding, generating, and modeling electromagnetic fields and their effects. He is an active member of IEEE SCC28 committee on the biological effects of RF and the IEEE 1451 committee on sensor networking. He currently directs several projects dealing with software radio technologies, specializing in spread-spectrum receivers, and is a communications analyst for the Army's Future Combat Systems (FCS) network, focusing on system issues, network vulnerability, and combat identification. He has several patents and patents pending in the area of wireless communications.

Asis Nasipuri (Chapter 12) was a contributor to *Handbook of RF & Wireless Technologies*. He is a professor in the department of electrical and computer engineering at the University of North Carolina at Charlotte. He received his BS in electronics and electrical communication engineering from the Indian Institute of Technology in Kharagpur, India in 1987 and his MS and PhD in electrical a computer engineering from the University of Massachusetts at Amherst in 1990 and 1993, respectively. He then joined the Indian Institute of Technology at Kharagpur, India as a faculty member in the Department of Electronics and Electrical Communication Engineering. From 1998 to 2000, he served as a visiting researcher in the Department of Computer Science at the University of Texas at San Antonio. Since 2000, he has been at UNC-Charlotte as an assistant professor of electrical and computer engineering. Nasipuri's research interests include mobile ad hoc and sensor networks, wireless communications, and statistical signal processing. He has published more than 20 research articles on these topics.

Ron Olexa (Chapters 7, 8, and 16) is the author of *Implementing 802.11, 802.16, and 802.20 Wireless Networks*. He is currently President of Horizon Wi-Com, a wireless carrier providing WiMax service to major markets in the Northeast US. He is also the owner of Wireless Implementation LLC, a consulting company that has provided technical support and business planning guidance to project as diverse at satellite communications systems, Cellular network deployments, WiMax and 802.11 hotspot and hotzone implementations. He has previously been CTO at Advanced Radio Telecom and Dialcall, COO of Superconducting Core Technologies, and has held various senior management positions in large wireless communications companies over his 30 year career.

Robert Poor (Chapter 14) was a contributor to *Handbook of RF & Wireless Technologies*. He is chief technology officer for Ember Corporation in Boston.

This page intentionally left blank

CHAPTER 1

Basics of Wireless Communications

Daniel M. Dobkin

1.1 Harmonic Signals and Exponentials

Before we begin to talk about wireless, we briefly remind the reader of a previous acquaintance with three concepts that are ubiquitous in radio engineering: sinusoidal signals, complex numbers, and imaginary exponentials. The reader who is familiar with such matters can skip this section without harm.

Almost everything in radio is done by making tiny changes—modulations—of a signal that is periodic in time. The archetype of a smooth periodic signal is the sinusoid (Figure 1.1), typically written as the product of the angular frequency ω and time t.

Both of these functions alternate between a maximum value of 1 and minimum value of -1; cosine starts at $+1$, and sine starts at 0, when the argument is zero. We can see that cosines and sines are identical except for an offset in the argument (the *phase*):

$$\cos(\omega t) = \sin\left(\omega t + \frac{\pi}{2}\right) \tag{1.1}$$

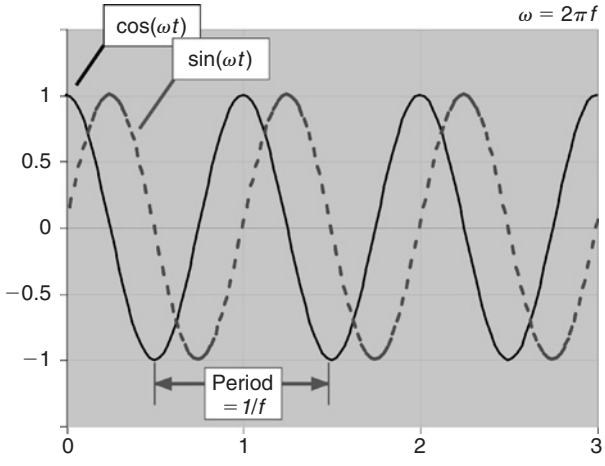


Figure 1.1: Cosine and Sine Functions

We say that the sine lags the cosine by 90 degrees. (Note that here, following common practice, we write angles in radians but often speak of them in degrees.) The cosine and sine are periodic with a period $= (1/f)$, where $f = \omega/2\pi$ is the frequency in cycles per second or hertz.

Let us now digress briefly to discuss complex numbers, for reasons that will become clear in a page or two. Imaginary numbers, the reader will recall, are introduced to provide square roots of negative reals; the unit is $i = \sqrt{(-1)}$. A complex number is the sum of a real number and an imaginary number, often written as, for example, $z = a + bi$. Electrical engineers often use j instead of i, so as to use i to represent an AC; we shall, however, adhere to the convention used in physics and mathematics. The complex conjugate z^* is found by changing the sign of the imaginary part: $z^* = a - bi$.

Complex numbers can be depicted in a plane by using the real part as the coordinate on the x- (real) axis, and the imaginary part for the y- (imaginary) axis (Figure 1.2). Operations on complex numbers proceed more or less the same way as they do in algebra, save that one must remember to keep track of the real and imaginary parts. Thus, the sum of two complex numbers can be constructed algebraically by

$$(a + bi) + (c + di) = [a + c] + [b + d]i \tag{1.2}$$

and geometrically by regarding the two numbers as vectors forming two sides of a parallelogram, the diagonal of which is their sum (Figure 1.3).

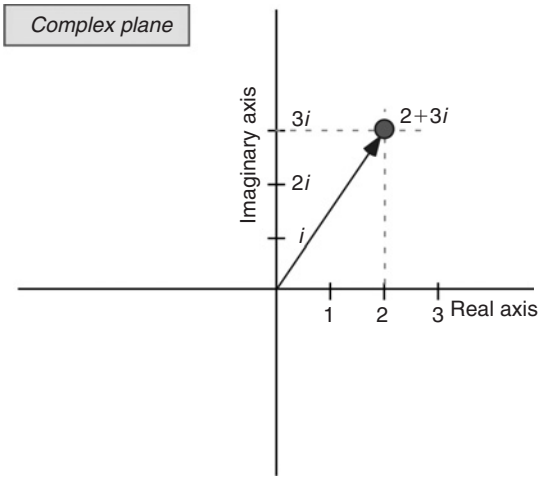


Figure 1.2: Complex Number Depicted as a Vector in the Plane

Multiplication can be treated in a similar fashion, but it is much simpler to envision if we first define the length (also known as the *modulus*) and angle of a complex number. We define a complex number of length 1 and angle θ to be equal to an exponential with an imaginary

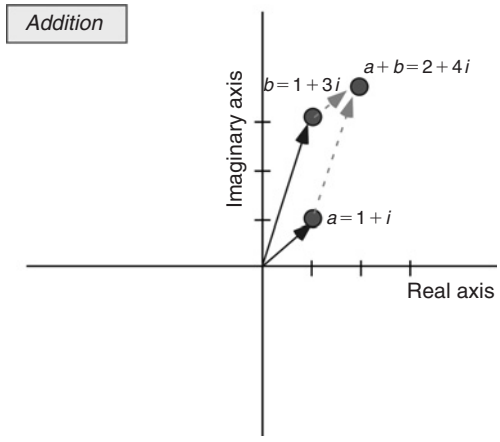


Figure 1.3: Addition of Complex Numbers

argument equal to the angle (Figure 1.4). Any complex number (e.g., *b* in Figure 1.4) can then be represented as the product of the modulus and an imaginary exponential whose argument is equal to the angle of the complex number in radians.

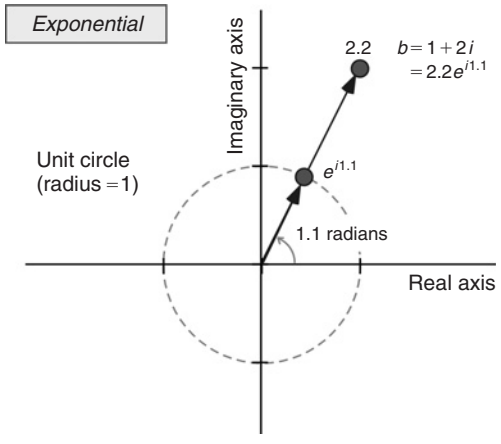


Figure 1.4: Imaginary Exponentials and Complex Numbers

By writing a complex number as an exponential, multiplication of complex numbers becomes simple, once we recall that the product of two exponentials is an exponential with the sum of the arguments:

$$(e^a) \cdot (e^b) = e^{[a+b]} \tag{1.3}$$

The product of two complex numbers is then constructed by multiplying their moduli and adding their angles (Figure 1.5).

$$\left(\rho_1 e^{i\theta_1}\right) \cdot \left(\rho_2 e^{i\theta_2}\right) = \left[\rho_1 \rho_2\right] e^{i[\theta_1+\theta_2]} \tag{1.4}$$

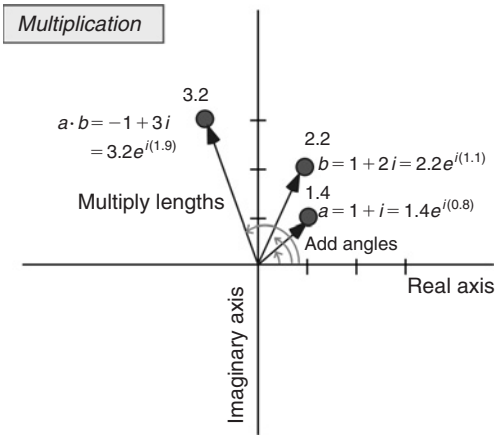


Figure 1.5: Multiplication of Complex Numbers

We took the trouble to introduce all these unreal quantities because they provide a particularly convenient way to represent harmonic signals. Because the *x*- and *y*-components of a unit vector at angle θ are just the cosine and sine, respectively, of the angle, our definition of an exponential with imaginary argument implies

$$e^{i\theta} = \cos(\theta) + i\sin(\theta) \tag{1.5}$$

Thus, if we use for the angle a linear function of time, we obtain a very general but simultaneously compact expression for a harmonic signal:

$$\begin{aligned} e^{i(\omega t+\phi)} &= \cos(\omega t + \phi) + i\sin(\omega t + \phi) \\ &= [\cos(\omega t) + i\sin(\omega t)] \cdot [\cos(\phi) + i\sin(\phi)] \end{aligned} \tag{1.6}$$

In this notation, the signal may be imagined as a vector of constant length rotating in time, with its projections on the real and imaginary axes forming the familiar sines and cosines (Figure 1.6). The phase offset ϕ represents the angle of the vector at $t = 0$.

In some cases we wish to use an exponential as an intermediate calculation tool to simplify phase shifts and other operations, converting to a real-valued function at the end by either simply taking only the real part or adding together exponentials of positive and negative frequency. (The reader may wish to verify, using equations [1.5] and [1.6], that the sum of exponentials of positive and negative frequencies forms a purely real or purely imaginary sinusoid.) However, in radio practice, a real harmonic signal $\cos(\omega t + \phi)$ may also be regarded as being the product of a real carrier $\cos(\omega t)$ and a complex number $\mathrm{I} + i\mathrm{Q} = [\cos(\phi) - i\sin(\phi)]/2$, where the imaginary part is obtained through multiplication with $\sin(\omega t)$ followed by filtering. (Here I and Q denote "in-phase" and "quadrature," that is, 90 degrees out of phase, respectively.) We'll have more to say about the uses of such decompositions when we discuss radios in Chapter 3.

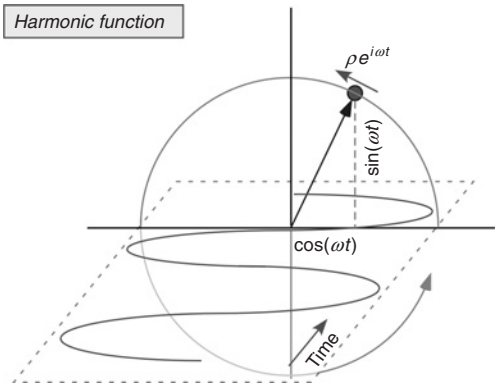


Figure 1.6: An Imaginary Exponential Can Represent Sinusoidal Voltages or Currents

Finally, we note one other uniquely convenient feature of exponentials: differentiation and integration of an exponential with a linear argument simply multiply the original function by the constant slope of the argument:

$$\frac{d}{dx}(e^{ax}) = ae^{ax} \qquad \int e^{ax}dx = \frac{1}{a}e^{ax} \tag{1.7}$$

1.2 Electromagnetic Waves and Multiplexing

Now that we are armed with the requisite tools, let us turn our attention to the main topic of our discussion: the use of electromagnetic waves to carry information. An electric current element **J** at some location [1] induces a potential **A** at other remote locations, such as [2]. If the current is harmonic in time, the induced potential is as well. The situation is depicted in Figure 1.7.

The magnitude of the induced potential falls inversely as the distance and shifts in phase relative to the phase of the current. (The reader may wish to verify that the time dependence of **A** is equivalent to a delay by r/c.) The induced potential in turn may affect the flow of electric current at position [2], so that by changing a current **J**[1] we create a delayed and attenuated but still detectable change in current **J**[2]: we can potentially communicate between remote locations by using the effects of the electromagnetic disturbance **A**.

In principle, every current induces a potential at every location. It is this universality of electromagnetic induction that leads to a major problem in using electromagnetic waves in communications. The potential at our receiver, **A**, can be regarded as a medium of communications that is shared by every possible transmitter **J**. How do we detect only the signal we are interested in?

The sharing of a communications channel by multiple users is known as *multiplexing*. There are a number of methods to successfully locate the signals we wish to receive and reject others.

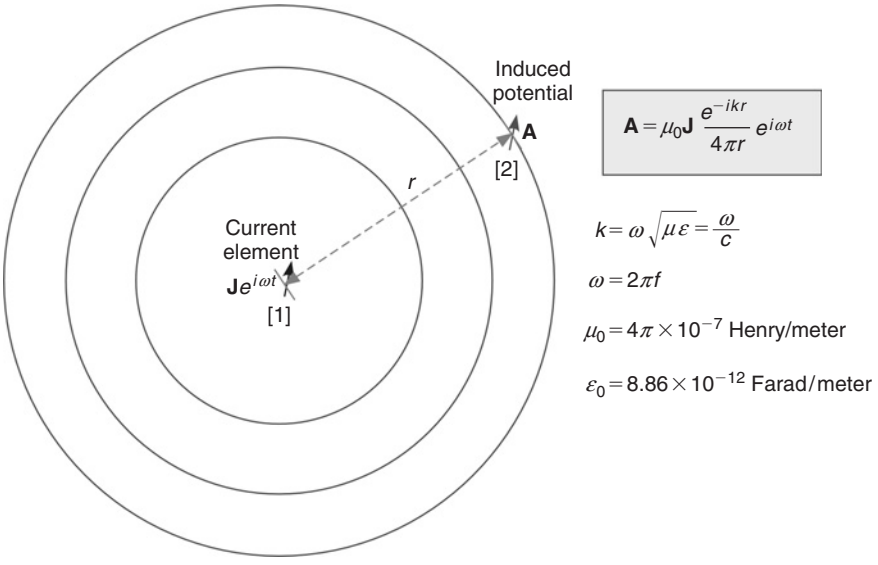


Figure 1.7: A Harmonic Current at [1] Induces a Harmonic Potential at [2]

A few important examples are the following:

- *Frequency-division multiplexing*: only receive signals with a given periodicity and shape (sinusoidal, of course).
- *Spatial multiplexing*: limit signals to a specific geographical area. Recall that induced potentials fall off as (1/distance) in the ideal case, and in practice attenuation of a signal with distance is often more rapid due to obstacles of various kinds. Thus, by appropriate choice of signal power, location, and sensitivity, one can arrange to receive only nearby signals.
- *Time-division multiplexing*: limit signals to a specific set of time slots. By appropriate coordination of transmitter and receiver, only the contents of the desired time slot will be received.
- *Directional multiplexing*: only listen to signals arriving from a specific angle. This trick may be managed with the aid of antennas of high directivity.
- *Code-division multiplexing*: only listen to signals multiplied by specific code. Rather in the fashion that we can listen to a friend's remarks even in a crowded and noisy room, in code-division multiplexing we select a signal by the pattern it obeys. In practice, just as in conversation, to play such a trick it is necessary that the desired signal is at least approximately equal to other undesired signals in amplitude or power, so that it is not drowned out before we have a chance to apply our pattern-matching template.

In real communications systems, some or all of these techniques may be simultaneously used, but almost every modern wireless system begins with frequency-division multiplexing by transmitting its signals only within a certain frequency band. (We briefly examine the major exception to this rule, ultrawideband communications, in section 1.5.) We are so accustomed to this approach that we often forget how remarkable it is: the radio antenna that provides us with music or sports commentary at 105 MHz is also exposed to AM signals at hundreds to around a thousand kHz, broadcast television at various frequencies between 50 and 800 MHz, aeronautical communications at 108–136 MHz, public safety communications at 450 MHz, cellular telephony at 880 and 1940 MHz, and cordless telephones, wireless local area networks (WLANs), and microwave ovens in the 2400-MHz band, to name just a few.

All these signals can coexist harmoniously because different frequencies are *orthogonal*. That is, let us choose a particular frequency, say ω_c, that we wish to receive. To extract only the part of an incoming signal that is at the desired frequency, we multiply the incoming unknown signal $s(t)$ by a sine or cosine (or more generally by an exponential) at the *wanted* frequency ω_c and add up the result for some time—that is, we integrate over a time interval T, presumed long compared with the periodicity $1/f$ (equation [1.8]). The reader may recognize in equation [1.8] the *Fourier cosine transform* of the signal s over a finite domain. A similar equation may be written for the sine, or the two can be combined using an imaginary exponential.

$$\tilde{S}(\omega_c) = \frac{1}{T}\int_0^T s(t)\cos(\omega_c t)dt \tag{1.8}$$

If $s(t)$ is another signal at the same frequency, the integral will wiggle a bit over each cycle but accumulate over time (Figure 1.8).

On the other hand, if the unknown signal is at a different frequency, say $(\omega_c + \delta)$, the test and unknown signals may initially be in phase, producing a positive product, but over the course

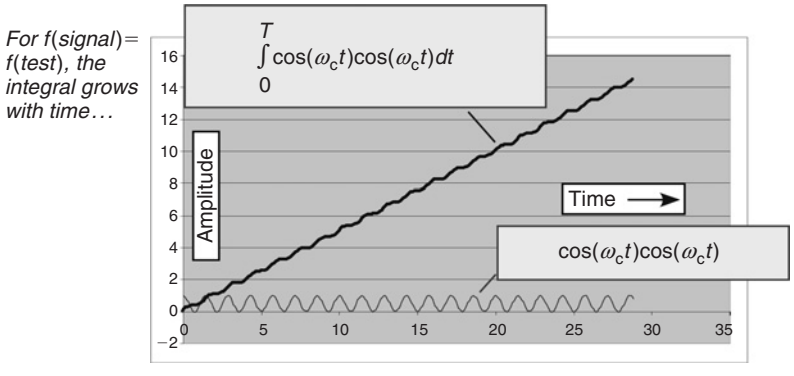


Figure 1.8: Unknown Signal at the Same Frequency as Wanted Signal

of some time they will drift out of phase, and the product will change signs (Figure 1.9). Thus, the integral will no longer accumulate monotonically, at least over times long compared with the difference period ($1/\delta$) (Figure 1.10); when we divide by T and allow T to become large, the value of $S(\omega_c)$ will approach zero.

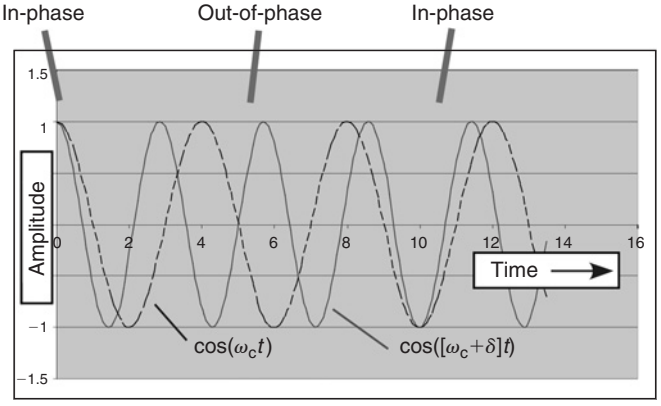


Figure 1.9: Two Signals at Different Frequencies Do Not Remain in Phase

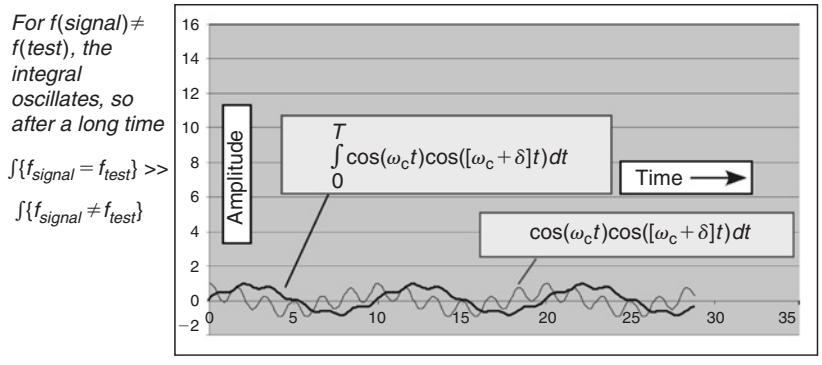


Figure 1.10: Unknown Signal at a Different Frequency from Wanted Signal

Any signal that is periodic in time can be regarded as being composed of sinusoids of differing frequencies: in more formal terms we can describe a signal either as a function of time or as a function of frequency by taking its Fourier transform (i.e., by performing the integration [1.8] for each frequency ω_c of interest.) The orthogonality of those differing frequencies makes it possible to extract the signal we want from a complex mess, even when the wanted signal is small compared with the other stuff. This operation is known generally as *filtering*. A simple example is shown in Figure 1.11. It is generally very easy when the frequencies are

widely separated, as in Figure 1.11, but becomes more difficult when frequencies close to the wanted frequency must be rejected. We examine some of the means to accomplish this task for WLAN radios in Chapter 3.

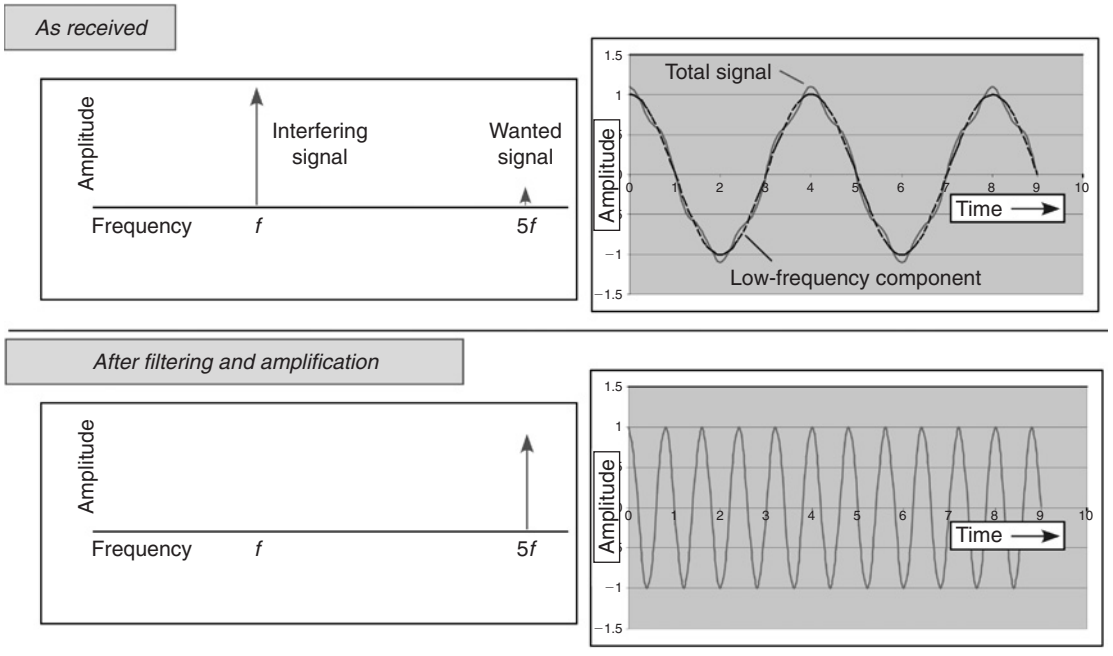


Figure 1.11: Extraction of a Wanted Signal in the Presence of a Large Unwanted Signal

1.3 Modulation and Bandwidth

1.3.1 Simple Modulations

So far the situation appears to be quite rosy. It would appear that one could communicate successfully in the presence of an unlimited number of other signals merely by choosing the appropriate frequency. Not surprisingly, things are not so simple: a single-frequency signal that is always on at the same phase and amplitude conveys no information. To actually transmit data, some aspect of our sinusoidal signal must change with time: the signal must be *modulated*. We can often treat the modulation as a slowly varying function of time (slow being measured relative to the carrier frequency) multiplying the original signal.

"slowly" varying modulation function — sinusoidal vibration at carrier frequency

$$f(t) = m(t)\cos(\omega_c t) \tag{1.9}$$

A simple example of a modulated signal may be obtained by turning the carrier on and off to denote, for example, 1 and 0, respectively: that is, $m(t) = 1$ or 0. This approach is known as *on–off keying* or OOK (Figure 1.12). OOK is no longer widely used in wireless communications, but this simple modulation technique is still common in fiber optic signaling.

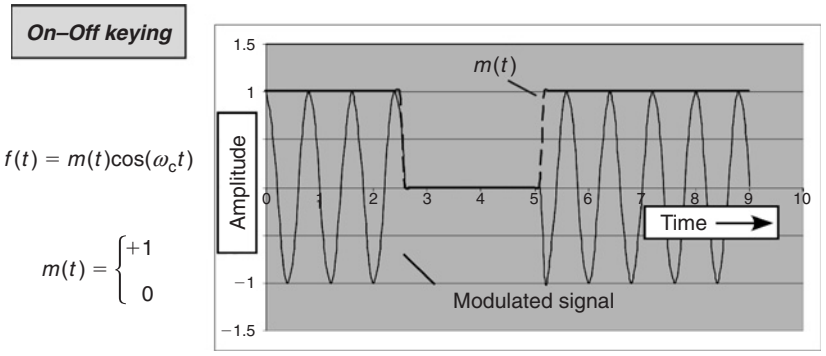


Figure 1.12: Modulation by Turning the Carrier On or Off

A key consequence of imposing modulation on a signal at a frequency ω_c is the inevitable appearance of components of the signal at *different frequencies* from that of the original carrier. The perfect orthogonality of every unique frequency present in the case of unmodulated signals is lost when we actually transmit data. Let us examine how this comes about for the particularly simple case of a sinusoidal modulation, $m = \cos(\omega_m t)$. Recall that the orthogonality of two different frequencies arose because contributions to the average from periods when the two signals are in phase are canceled by the periods when the signals are out of phase (Figure 1.9). However, the modulated signal is turned off during the periods when it is out of phase with the test signal at the different frequency $(\omega_c + \delta)$ so the contribution from these periods no longer cancels the in-phase part (Figure 1.13). The modulated carrier at (ω_c) is now detected by a filter at frequency $(\omega_c + \delta)$.

The astute reader will have observed that this frustration of cancellation will only occur when the frequency offset δ is chosen so as to ensure that only the out-of-phase periods are suppressed. In the case of a periodic modulation, the offset must obviously be chosen to coincide with the frequency of the modulation: $|\delta| = \omega_m$. In frequency space, a modulated carrier at frequency f_c acquires power at sidebands displaced from the carrier by the frequency of the modulation (Figure 1.14).

In the case of a general modulating signal $m(t)$, with Fourier transform $M(\omega)$, it can be shown that the effect of modulation is to translate the spectrum of the modulating or *baseband* signal up to the carrier frequency (Figure 1.15).

We can now see that data-carrying signals have a finite bandwidth around their nominal carrier frequency. It is apparent that to pursue our program of frequency-division multiplexing of

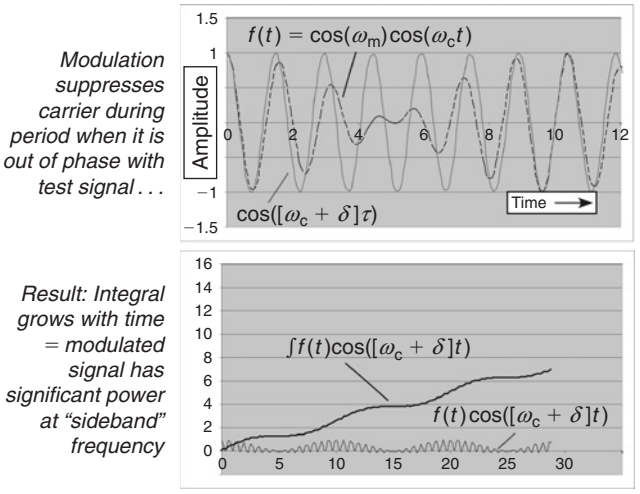


Figure 1.13: A Modulated Signal Is No Longer Orthogonal to All Other Frequencies

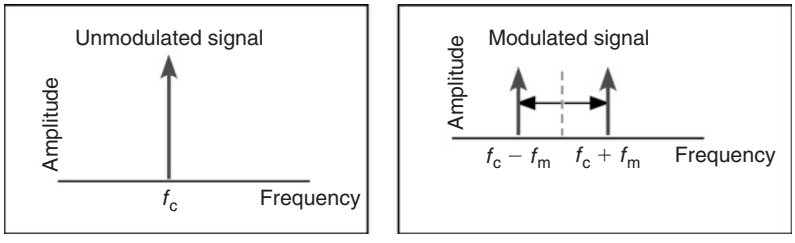


Figure 1.14: Modulation Displaces Power From the Carrier to Sidebands

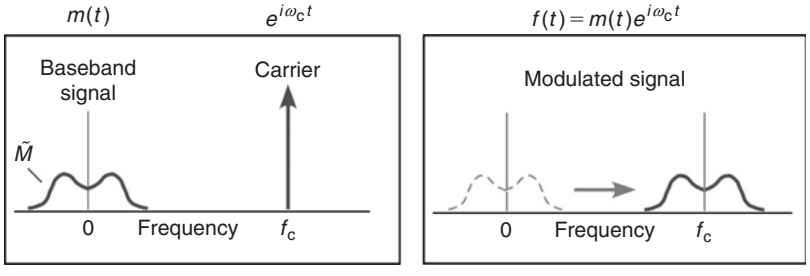


Figure 1.15: The Spectrum of a Carrier Modulated by a General Signal *m*(*t*)

signals, we shall need to allocate bands of spectrum to signals in proportion to the bandwidth those signals consume. Although the spectrum of a random sequence of bits might be rather more complex than that of a simple sinusoid, Figure 1.14 nevertheless leads us to suspect that the faster we modulate the carrier, the more bandwidth we will require to contain the resulting sidebands. More data require more bandwidth (Figure 1.16).

It would seem at first glance that the bandwidth required to transmit is proportional to the data rate we wish to transmit and that faster links always require more bandwidth. However, note

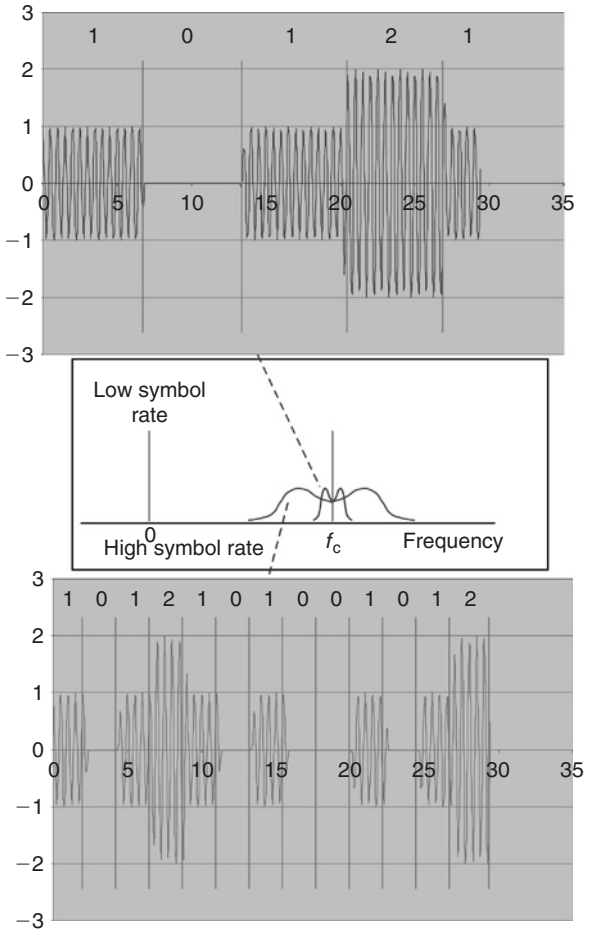


Figure 1.16: Faster Symbol Rate = More Bandwidth

that in Figure 1.16 we refer not to the bit rate but to the *symbol* rate of the transmitted signal. In the case shown, a symbol is one of three possible amplitudes, corresponding to a data value of 0, 1, or 2: this is an example of amplitude-shift keying (ASK), a generalization of OOK. (Note that in this and other examples we show abrupt transitions between different states of the carrier; in practice, the transitions are smoothed to minimize the added sidebands.) Each symbol might be said to contain 3/2 bit. The bit rate is thus 1.5 (symbol rate). More generally, we can envision a number of approaches to sending many bits in a single symbol. For example, we could use more amplitudes: if 8 amplitudes were allowed, one could transmit 3 bits in each symbol. Because the width of the spectrum of the modulating signal is mainly dependent on the rate at which transitions (symbols) occur rather than exactly what the transition is, it is clear that by varying the modulation scheme, we could send higher data rates without necessarily expanding the bandwidth consumed.

We can nevertheless guess that a trade-off might be involved. For example, the use of 8 distinct amplitudes means that the difference between (say) a "3" and a "4" is smaller than the difference between an OOK "1" and "0" for the same overall signal power. It seems likely that the more bits we try to squeeze into a symbol, the more vulnerable to noise our signal will become.

With these possibilities in mind, let us examine some of the modulation schemes commonly used in data communications. The first example, in Figure 1.17, is our familiar friend OOK. Here, in addition to showing the time-dependent signal, we have shown the allowed symbols as points in the phase/amplitude plane defined by the instantaneous phase and amplitude of the signal during a symbol. The error margin shows how much noise the receiver can tolerate before mistaking a 1 for a 0 (or vice versa).

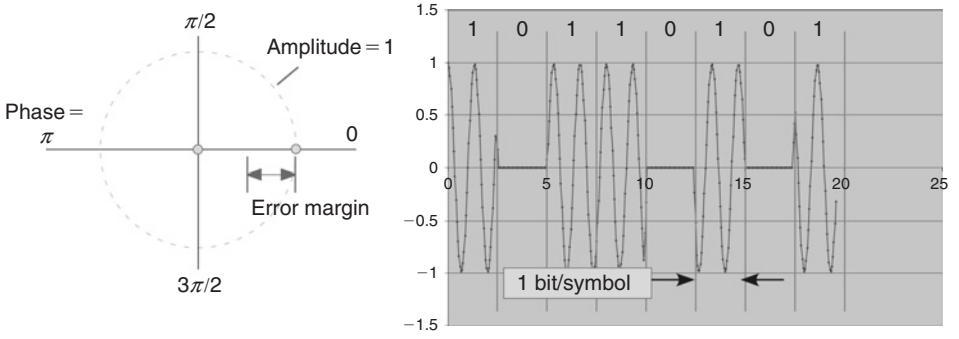


Figure 1.17: On–Off Keying (OOK)

Note that although we have shown the 1 symbol as a single point at a phase of 0 and amplitude 1, a symbol at any other phase—that is, any point on the circle *amplitude* =1—would do as well. OOK is relatively easy to implement because the transmitter doesn't need to maintain a constant phase but merely a constant power when transmitting a 1 and the receiver needs merely to detect the signal power, not the signal phase. On the down side, OOK only sends one bit with each symbol, so an OOK-modulated signal will consume a lot of bandwidth to transmit signals at a high rate.

As we mentioned previously, we might add more amplitudes to get more data: ASK (Figure 1.18). The particular example in Figure 1.18 has four allowed amplitudes and is denoted 4ASK. Once again we have collapsed the allowed states onto points for clarity but with the understanding that any point on, for example, the 2/3 circle will be received as (10), etc. 4ASK allows us to transmit 2 bits per symbol and would be expected to provide twice the data rate of OOK with the same bandwidth (or the same data rate at half the bandwidth). However, the margin available before errors in determining what symbol has been received (i.e., before symbol errors occur) is obviously much smaller than in the case of OOK. 4ASK cannot tolerate as much noise for a given signal power as can OOK.

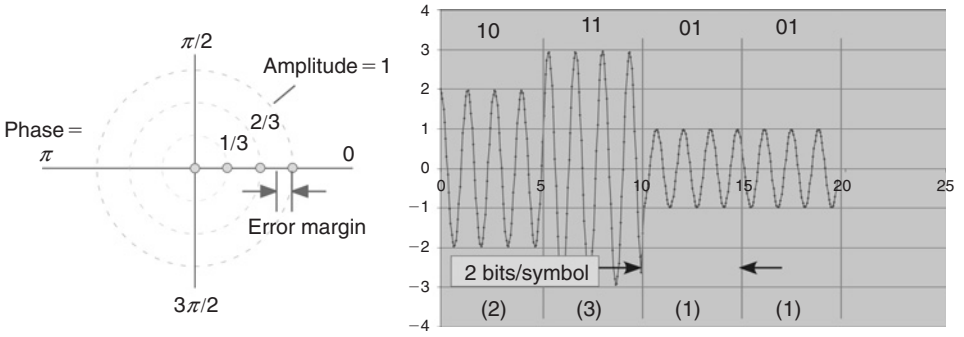


Figure 1.18: 4-Amplitude-Shift Keying (4ASK)

Although it is obviously possible to keep adding amplitudes states to send more bits per symbol, it is equally apparent that the margin for error will decrease in inverse proportion to the number of amplitude states. A different approach to increasing the number of states per symbol might be useful: why not keep track of the phase of the signal?

The simplest modulation in which phase is used to distinguish symbols, binary phase-shift keying (BPSK), is depicted in Figure 1.19. The dots in the figure below the binary symbol values are placed at constant intervals; a 1 is transmitted with the signal peaks coincident with the dots, whereas a 0 has its peaks between dots: 180 degrees or π radians out of phase. In phase-shift keying, the nominal symbols are points in the phase plane rather than circles: the group of points is known as a signal constellation. However, as long as the signal is large enough for its phase to be determined, the signal amplitude has no effect: that is, any received signal on the right half of the phase-amplitude plane is interpreted as a 1 and any signal on the left half is interpreted as 0. The error margin is thus equal to the symbol amplitude and is twice as large as the error margin in OOK for the same peak power. BPSK is a robust modulation, resistant to noise and interference; it is used in the lowest rate longest range states of 802-11 networks.

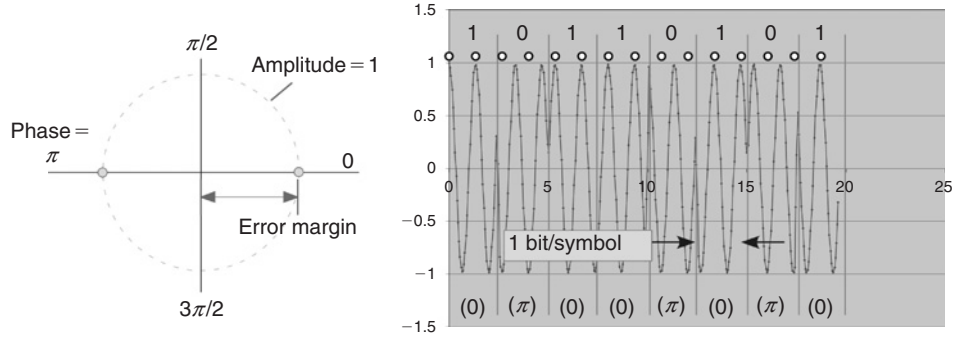


Figure 1.19: Binary Phase-Shift Keying (BPSK)

It is apparent that to get the best error margin for a given peak power, the constellation points ought to be spaced as far from one another as possible. To get 2 bits in one symbol, we ought to use four phases spaced uniformly around the amplitude = 1 circle: quaternary phase-shift keying (QPSK), shown in Figure 1.20. QPSK sends 2 bits per symbol while providing a noise margin larger than that of OOK at 1 bit per symbol for the same peak power. It is a very popular modulation scheme; we will encounter a variant of it in 802-11 WLANs.

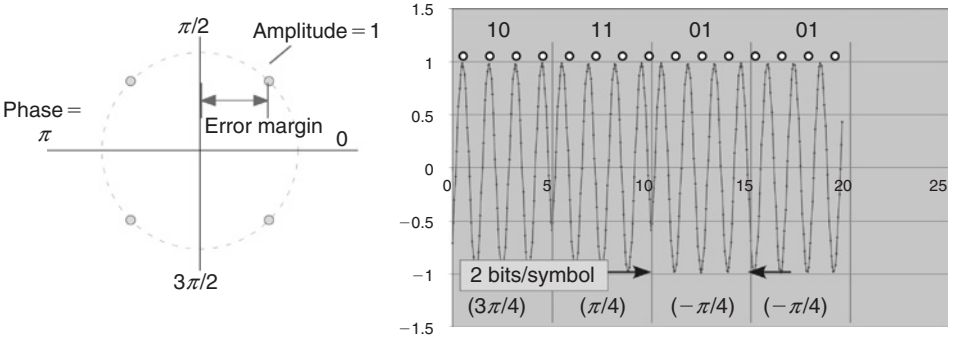


Figure 1.20: Quaternary Phase-Shift Keying (QPSK)

Again, we could continue along this path, adding more phase states to provide more bits per symbol, in a sequence like 8PSK, 16PSK, and so on. However, as before such a progression sacrifices error margin. An alternative approach is to combine the two schemes that we have heretofore held separate: that is, to allow both amplitude and phase to vary. Such modulation schemes are known as *quadrature-amplitude-modulation* or QAM. An example, 16 QAM, is depicted in Figure 1.21. Four bits can be sent with a single symbol, and the noise margin is still superior to that of 4ASK at 2 bits per symbol. More points could be added to the signal constellation; doubling the number of states in each axis for each step, we obtain 64 QAM, 256 QAM, and 1024 QAM (the latter perhaps more aptly described as a clear night sky in the Sierra than a mere constellation).

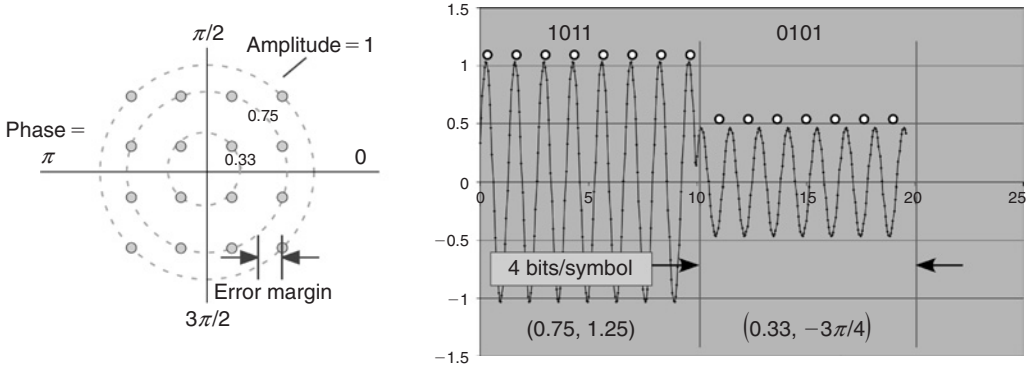


Figure 1.21: 16-Quadrature-Amplitude-Modulation (16QAM)

The various modulation schemes we examined are summarized in Table 1.1. Here the error margin is quoted in units where the peak signal amplitude is defined as 1. It is apparent that the choice of modulation is a trade-off between the target data rate, the bandwidth available to transmit it in, the noise immunity and therefore range of the transmission, and the complexity and cost of the resulting transmitter and receiver.

Although it seems likely that the noise susceptibility of a modulation scheme is increased when its error margin is reduced, it would clearly be helpful to make some quantitative statement about the relationship between a measure of error, such as the symbol or bit error rate, and a measure of noise, such as the ratio of the signal power to the noise power, the *signal-to-noise ratio* (*S*/*N*). Such relationships can be constructed, assuming the noise to act as a Gaussian (normally distributed) random variable in the phase plane, but the calculations are rather laborious. We can achieve almost the same result in a much simpler manner by exploiting the rule of thumb that almost all the area of a Gaussian distribution is within three standard deviations of the mean value. Assume the noise voltage is a Gaussian with standard deviation σ and thus average power proportional to σ^2. If the error margin of a modulation, measured as a fraction of the signal as in Table 1.1, is larger than 3σ, then the symbol error rate is likely to be very small (Figure 1.22).

Table 1.1: Summary of Carrier Modulation Approaches

Modulation	Bits/Symbol	Error Margin		Complexity
OOK	1	1/2	0.5	Low
4ASK	2	1/6	0.17	Low
BPSK	1	1	1	Medium
QPSK	2	$1/\sqrt{2}$	0.71	Medium
16QAM	4	$\sqrt{2}/6$	0.23	High
64QAM	6	$\sqrt{2}/14$	0.1	High

Using this approach, we can obtain a value of (*S*/*N*) that would be expected to provide reasonably error-free communications using each example modulation scheme. The noise standard deviation is allowed to be equal to 1/3 of the error margin, and the signal amplitude is averaged over the constellation points to obtain an estimate of average signal amplitude; the ratio of the squares provides an estimate of the ratio of the signal power to the allowed noise power. However, one might complain that such a comparison does an injustice to the schemes using larger numbers of bits per symbol, because these deliver more information per symbol. If we divide the (*S*/*N*) by the number of bits in the symbol, we obtain a measure of the amount of signal required *per bit* of data, which seems a more appropriate way to compare differing modulation approaches, because it is after all bits, not symbols, that are the ultimate currency

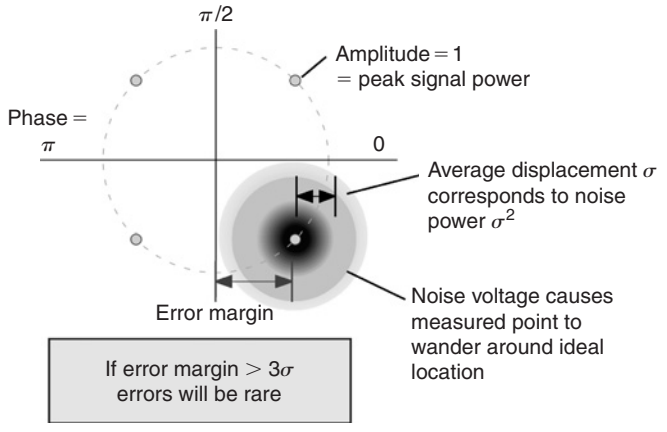


Figure 1.22: Approach for Estimation of *S/N* Requirement

of data transfer. The quantity (*S*/*N*)/(bits/symbol) can be reexpressed in a fashion that has an appealing interpretation:

$$\left\{\frac{S}{N}\right\}\Big/\left\{\frac{bits}{symbol}\right\} = \frac{S}{N_o(BW)}\frac{symbols}{bit} = \frac{S}{N_o(1/T_{symbol})}\frac{symbols}{bit}$$
$$= \frac{ST_{symbol}}{N_o}\frac{symbols}{bit} = \frac{ST_{bit}}{N_o} = \boxed{\frac{E_{bit}}{N_o}} \tag{1.10}$$

In equation [1.10], we defined a quantity, N_o, that is the total noise divided by the signal bandwidth or the noise power in, for example, watts per hertz. We also assumed that the bandwidth is inversely proportional to the symbol time (Figure 1.16). The product of symbol time and (symbols/bit) becomes an effective bit time; the product of bit time and signal power is the *energy per bit*, E_{bit}, often abbreviated E_b. The quantity (E_b/N_o) is thus a measure of the signal energy available per bit of data to be transmitted, relative to the fundamental amount of noise in the channel, and is a more appropriate basis for comparison of modulation approaches than raw (*S*/*N*).

The resulting extended comparison of modulation noise immunity is summarized in Table 1.2. The ratios (*S*/*N*) and (E_b/N_o) vary over a wide range and are conveniently expressed in decibels (dB): $(S/N)_{dB} = 10 \log[(S/N)]$. The noise amplitude is simply 1/3 of the error margin. <Signal> is the average of the signal over the constellation points, normalized to the peak amplitude. This is a reasonable first estimate of the average signal power, though as we'll see in Chapter 3, it must also be corrected for the path the signal takes from one constellation point to another. The results are within 0.4 dB of the values of (E_b/N_o) required to achieve a bit error rate of 10^{-5} by a more accurate but much more laborious examination of the probability of error within the constellation point by point.

The results, particularly when expressed as the normalized quantity (E_b/N_o), confirm the initial impression that PSK and QAM modulations do a better job of delivering data in the presence of noise than corresponding ASK approaches. It is interesting to note that BPSK and QPSK have the same (E_b/N_o), even though QPSK delivers twice as many bits per symbol. QPSK is an excellent compromise between complexity, data rate, and noise immunity, and we shall find it and its variants to be quite popular in wireless networking. Table 1.2 also reemphasizes the fact that higher rates require a higher *S*/*N*: faster means shorter range for the same bandwidth and power.

Table 1.2: Noise Immunity of Various Modulation Schemes

Modulation	Noise Amplitude	<Signal>	(S/N) (dB)	(E_b/N_o) (dB)
OOK	0.167	0.71	12.6	12.6
4ASK	0.053	0.62	21.4	18.3
BPSK	0.333	1.00	9.5	9.5
QPSK	0.237	1.00	12.5	9.5
16QAM	0.077	0.74	19.7	13.7
64QAM	0.033	0.65	25.8	18.1

One more minor but by no means negligible aspect of modulating a signal must still be examined. The average value of the signal, normalized to a peak value of 1 for the constellation point most distant from the origin, is seen to decrease for more complex modulations: equivalently, the ratio of the peak to average power increases as more bits/symbol are transmitted. Other effects, having to do with both the details of implementation of the modulation and more subtle statistical issues for complex modulation schemes, give rise to further increases in peak-to-average power ratios. High peak-to-average ratios require that the transmitter and receiver must be designed for much higher instantaneous power levels than the average signal level would indicate, adding cost and complexity. A more detailed examination of the peak-to-average ratio is presented in the discussion of radios in Chapter 3.

We examined a number of common modulation schemes, exposing the trade-off between data rate, bandwidth, and noise. Any number of variants on these approaches could be imagined, including adjustments in the exact location of the constellation points, differing numbers of points, and different conventions on how to determine what value had been detected. Yet there certainly seems to be a trend, as shown in Table 1.1, that the more bits one transmits per symbol, the less noise can be tolerated. This observation suggests that perhaps the path of increasing symbol complexity cannot be continued indefinitely and that some upper limit might exist on the amount of data that can be sent in a given bandwidth in the presence of noise, no matter how much ingenuity is devoted to modulation and detection. In a series of seminal publications in the late 1940s, Claude Shannon of Bell Laboratories demonstrated that

any communications channel has a finite data capacity determined by the bandwidth [BW], signal power *S*, and noise power *N* in the channel:

$$\left\{\frac{bits}{s}\right\} \leq [BW]\log_2\left(1+\frac{S}{N}\right) \tag{1.11}$$

The ultimate capacity of a radio link, like any other channel, is determined by its bandwidth and the (*S*/*N*). To clarify the relevance of this limitation, let us look at an example. As explained in detail in Chapter 2, most of the 802.11-based (Wi-Fi) local area network protocols use a channel that is roughly 16 MHz wide. The (*S*/*N*) will vary widely depending on the separation between transmitter and receiver and many other variables, but for the present purpose let us adopt a modest and computationally convenient value of (*S*/*N*) = 7:1. We then have

$$\left\{\frac{bits}{s}\right\} \leq [16\times10^6]\log_2(8) = 4.8\times10^7 \tag{1.12}$$

The capacity of an 802.11 channel is about 48 Mbps (megabits/second) at what we will find corresponds to a quite modest signal-to-noise requirement. If we assume that the channel bandwidth is approximately equal to the inverse of the symbol rate, this upper bound corresponds to about 3 bits/symbol. (The direct-sequence version of 802.11, and 802.11b, actually uses a symbol rate of 11 megasamples/second (Msps), so that 4 bits/symbol would be available.) We can thus infer that in this case, little or no advantage would result from using a modulation such as 64QAM (6 bits/symbol): the noise-created symbol errors could presumably be corrected by coding, but the overhead associated with correcting the mistakes would exceed the benefit gained. Note that if a higher *S*/*N* could be obtained, by turning up the transmit power, reducing the transmit-to-receive distance, or other means, the capacity of the channel would be increased, allowing for exploitation of more complex symbols, though the increase is logarithmic: an eightfold increase in the signal is required to achieve a doubling of the data rate. On the other hand, it is also apparent that there's a lot of room in the 802.11 channel for improvement over the 802.11b maximum data rate of 11 Mbps, even at quite modest (*S*/*N*).

A broader examination of the performance of some important communications channels relative to the Shannon limit is provided in Table 1.3. It is apparent that channels cannot normally make use of the whole theoretical capacity available and that in many cases less than half of the upper limit is achieved. Wi-Fi stands out as a notably inefficient user of its allotted bandwidth, which is a hint that more efficient modulations could be used (which has been done with the advent of 802.11 g; see Chapter 2). Note also that a cable modem connection, being made over a wired link, can deliver a much larger (*S*/*N*) than is normally practical for a wireless link. A cable modem using 6 MHz can support actual data rates of 30 Mbps, notably larger than 802.11's 11 Mbps in 16 MHz, albeit a rather modest improvement given the huge increase in (*S*/*N*) necessary to achieve it.

Table 1.3: Actual and Theoretical Capacity of Some Communications Links

	EV-DO CDMA (cellphone)	Cable Modem	Wi-Fi (802.11b)
Bandwidth (MHz)	1.25	6	16
Configuration	Mobile, 2 km, LOS	Minimum FCC *S/N*	Indoors, 30 m, NLOS
(*S/N*)	6:1	2000:1	7:1
Maximum rate (Mbps)	3.6	65.8	48
Actual rate (Mbps)	2.5	30.3	11
Percent of maximum rate	70	46	23
LOS, line-of-sight from transmitter to receiver; NLOS, non-line-of-sight from transmitter to receiver.			

1.3.2 Orthogonal Frequency-Division Multiplexing

The modulation approaches we discussed so far are called *single-carrier* modulations: at any given moment, a carrier with a particular phase and amplitude can be said to exist. Single-carrier modulations are versatile and relatively simple to implement. However, they have some fundamental limitations when used in a wireless link. To understand why, we need to introduce the concept of *multipath*.

Consider a typical real-world transmission between a transmitting and receiving antenna (Figure 1.23). Various sorts of obstacles may exist that reflect the signal, providing alternatives to the direct path. The time required to propagate along these alternative paths, t_2 and t_3, will in general differ from the time taken by the direct path, t_1.

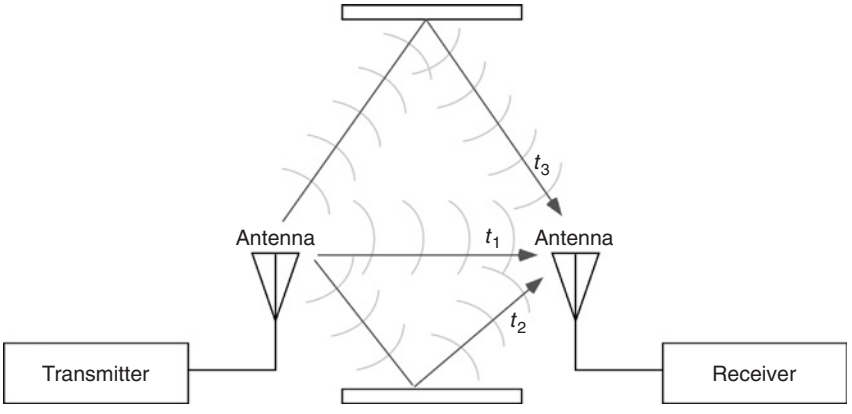


Figure 1.23: Multiple Paths May Exist Between Transmitter and Receiver

If the delay associated with the other times is comparable with the symbol time and the signal strengths don't differ by too much, a serious problem is encountered at the receiver: the sum of all the received signals may not match the transmitted signal. In Figure 1.24, we depict the two delayed and attenuated replicas of the signal added to the directly transmitted version. The

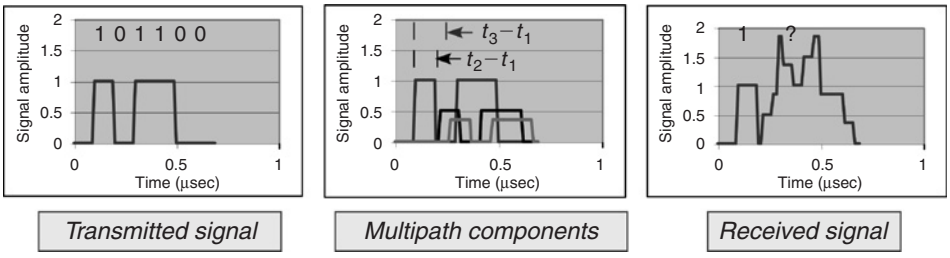


Figure 1.24: Multipath Results in Garbled Signals

sum is garbled after the first bit, and there is no easy way to extract the transmitted symbols: intersymbol interference has wiped out the data even though there is ample signal power available.

We should note that in Figure 1.24, for simplicity we show the three signals as adding after demodulation, that is, after the carrier frequency has been removed. In fact, the signals add at the carrier frequency, which may give rise to the related but distinct problem of fading.

When is this sort of distortion likely to be a problem? In general, multipath is important when propagation delays are comparable with symbol times. Using the convenient approximation that the speed of light is about 3.3 nsec/m, Table 1.4 shows the typical regions that one can expect to cover at a given data rate before multipath rears its ugly head. We see that one can transmit a respectable 10 Msps over a sizable building with single-carrier modulations, but if we wish to achieve higher data rates over larger regions, some other approach is required.

Table 1.4: Data Rate vs. Multipath-Free Region Size

Symbol Rate (Msps)	Symbol Time (msec)	Path Distance (m)	Path Description
0.1	10	3300	City
1	1	330	Campus
10	0.1	33	Building
100	0.01	3	Room

An increasingly popular means of tackling multipath to allow higher rates over larger areas is the use of a specialized modulation, orthogonal frequency-division multiplexing (OFDM). OFDM is used in the WLAN standards 802.11a, 802.11 g, and HiperLAN, as well as digital broadcast television in Europe and Asia (known as COFDM). OFDM has also been proposed for very-high-rate personal area networks to deliver high-resolution video within the home. In this section we'll take a general look at how OFDM works, in preparation for examining the specific OFDM modulations used in the WLAN standards described in Chapter 2.

OFDM uses three basic tricks to overcome multipath. The first is parallelism: sending one high-speed signal by splitting it into a number of lower speed signals sent in parallel

(Figure 1.25). Serial-to-parallel conversion is very well known in the wired world: open up any desktop computer and you'll see numerous ribbon cables, composed of a large number of inexpensive wires carrying many simultaneous signals at a low rate, emulating the function of a more expensive and delicate coaxial cable carrying a much higher rate serial signal. In the case shown in Figure 1.25, the signal path is split into five parallel paths, each carrying 2 bits/symbol and thus running at 1/10th of the rate of the original serial path, allowing a 10-fold increase in the transmission delay allowed before serious multipath distortion occurs. A serial symbol time of 0:1 msec becomes a parallel symbol time of 1 msec.

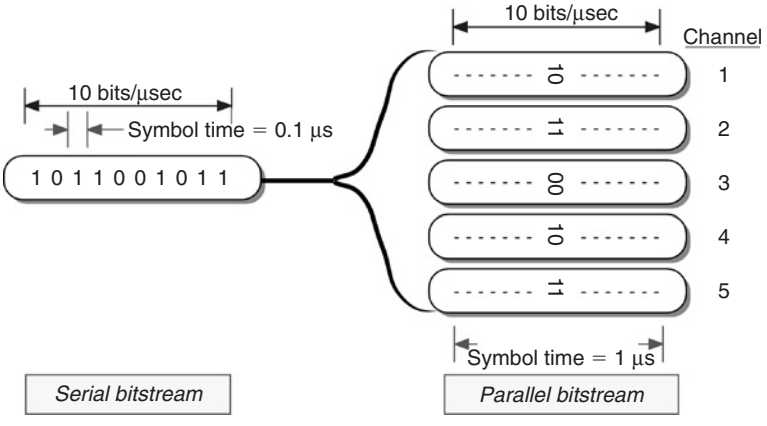


Figure 1.25: Split a High-Speed Serial Signal into Multiple Slow Parallel Signals

Where do we get these parallel channels? The obvious answer is to use separate carrier frequencies, one for each channel. Such an arrangement is shown in Figure 1.26. If implemented in a conventional fashion using separate transmitters and receivers with filters for each subcarrier, there are two problems: the cost of the equipment will increase linearly with the number of parallel channels or subcarriers, and the spectrum will be used wastefully due to the need for guard bands between the subcarriers to allow the receivers to filter their assigned subcarrier out of the mess.

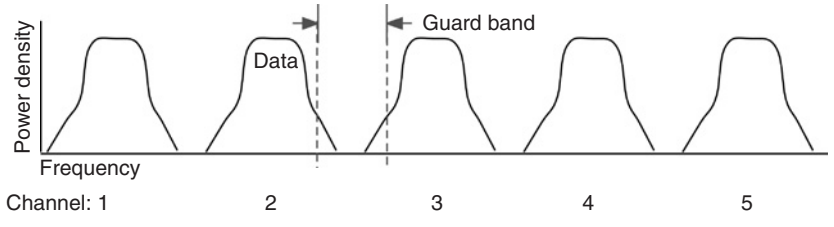


Figure 1.26: Conventional Frequency-Division Multiplexing for Subcarriers

To surmount these two obstacles, the two additional tricks at the heart of OFDM are necessary. Trick 2 is to exploit the orthogonality of different frequencies in a more subtle fashion than

we have done heretofore. We already know that differing frequencies are orthogonal when integrated over any long time period. However, we can be much more specific: two different frequencies are exactly orthogonal if the integration time contains an integral number of cycles of each frequency:

$$\int_0^T \cos\left(2\pi\left[\frac{n}{T}\right]t\right)\cos\left(2\pi\left[\frac{m}{T}\right]t\right)dt = 0 \quad for\ m \neq n \tag{1.13}$$

In equation [1.13], we can see that the two cosines undergo, respectively, $(n/T) \times T = n$ and $(m/T) \times T = m$ full cycles of oscillation in the time T. If we choose T as our symbol time, we can send symbols consisting of an amplitude and phase for each frequency in parallel at frequencies spaced by integers and still have every frequency orthogonal to every other frequency, as long as we are careful to integrate over the correct time interval. The situation is depicted graphically in Figure 1.27.

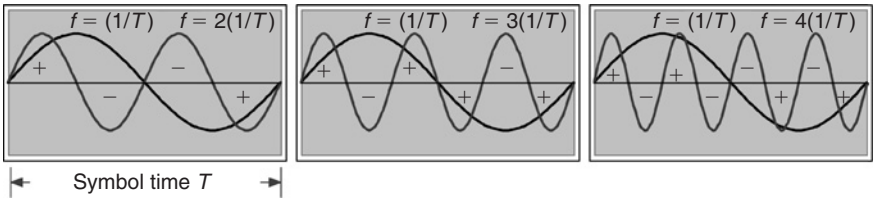


Figure 1.27: Products of Integer Frequencies Integrate to 0 Over an Integer Number of Cycles

The result of such a scheme, as shown in Figure 1.28, is the elimination of the guard bands otherwise required between subcarriers, resulting in much more efficient use of a given slice of spectrum (contrast this image with that of Figure 1.26).

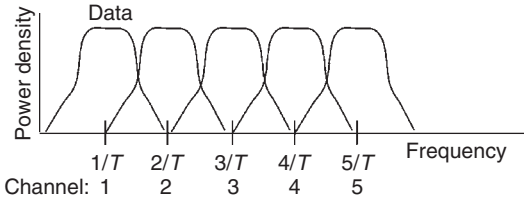


Figure 1.28: Orthogonal Subcarriers Use Spectrum Efficiently

By thus using closely spaced orthogonal subcarriers, we can split a serial symbol stream into a number of slower parallel symbol streams. If we imagine that multipath delay was equal to the serial symbol time, the same delay will now constitute 1/5th of the symbol time of our parallel symbols. Although this is obviously an improvement over having the whole symbol garbled, we can do even better at the cost of some slight decrease in data rate. If the symbol time is extended by (in this case) an extra 20% and then we simply eliminate the beginning part where intersymbol interference takes place, we can perform our integration over the remainder

of the symbol time where the received symbol is essentially ideal. Such a scheme is depicted schematically in Figure 1.29. The extra symbol time is called the *guard interval* and represents signal energy intentionally wasted to improve multipath resistance.

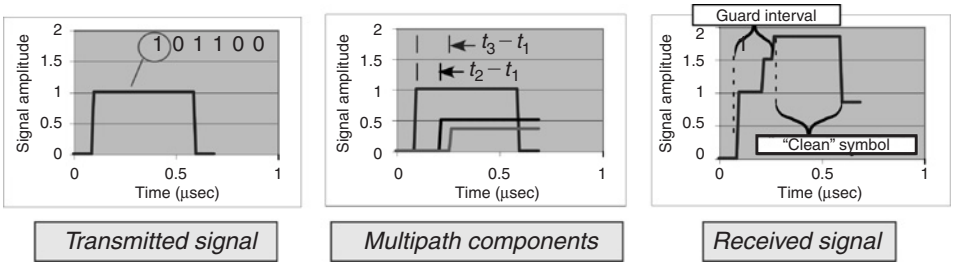


Figure 1.29: Use of a Guard Interval to Remove Intersymbol Interference

If one knew in advance exactly where the multipath boundary was, the guard interval could be implemented by simply turning the signal off during this time. However, in this case any error in determining the edge of the guard interval would result in integration of the signal over partial cycles: the orthogonality of the different subcarriers would be lost and interference between the parallel symbols would result. Instead, the guard interval is normally implemented by a *cyclic extension* of the original symbol, as shown in Figure 1.30. The portion of each subcarrier that occurs during the time period corresponding to the guard interval is simply repeated at the end of the base interval. Because all the subcarriers are periodic in *T*, this is equivalent to performing the same operation on the final symbol, which is the sum of all the subcarriers. The resulting cyclically extended symbol preserves the orthogonality of its subcarriers over the interval T even if the start of the interval is displaced slightly, because the portion that is removed from the beginning is just added at the end.

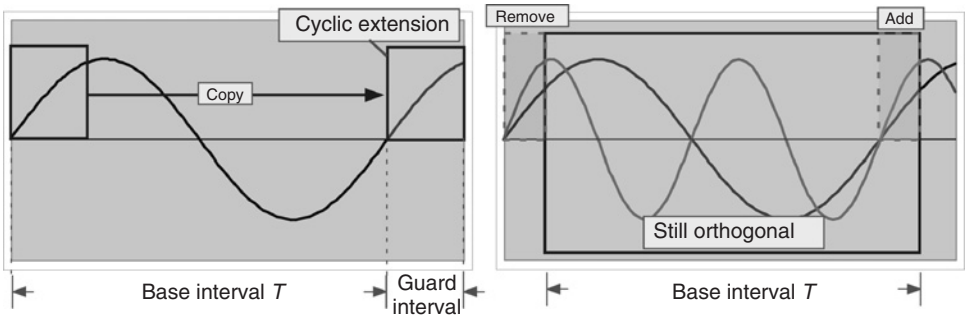


Figure 1.30: Cyclic Extension of an OFDM Symbol Makes Integration Insensitive to Small Time Offsets

The combination of parallelism implemented with OFDM and a cyclically extended symbol to provide a guard interval gives us a method of transmitting a clean undistorted set of symbols

at a high overall data rate in the presence of multipath distortion. Because the subcarriers are separated by an integration over a specific time interval rather than an analog filtering operation, only one radio is required to transmit and receive the signals. However, our putative radio still faces a significant challenge: to get a big improvement in multipath resistance, we need to use a lot of subcarriers. To extract each subcarrier, we must integrate the product of the received symbol and the desired frequency, and the number of points we must use in the integration is set by the highest frequency subcarrier we transmit. For example, in our five-subcarrier example, to extract the portion of a symbol at the lowest frequency $n = 1$, we must use enough points in our integration to accurately remove the part at the highest frequency, $n = 5$ ($4 \times 5 = 20$ sample points in time, or 10 if we treat the data as complex, will do). For N subcarriers, it appears that we have to do on the order of N integrations with N points each, or roughly N^2 multiplications and additions. For small N, that's no big deal, but we get little benefit from using small values of N.

Let us consider the example of 802.11a/g, to be discussed in more detail in Chapter 2. In this case 64 subcarriers are defined (though not all are used): $N = 64$. To extract all the subcarriers, we must perform approximately 4096 multiplications and additions during each symbol time of 4 msec. Each operation must be performed with enough resolution to preserve the original data: imagine that 8 bits is sufficient. If performed serially, this requires an operation every 49 nsec, or about 2 billion 8-bit adds and multiplies per second. Such an accomplishment is by no means impossible with modern digital hardware, but it is hardly inexpensive when one considers that the networking hardware is to be a small proportion of the cost of the device that makes use of it.

Fortunately, there is one trick left in the bag that immensely improves the situation: the *fast Fourier transform* (FFT). FFT algorithms achieve the necessary integration in roughly $N \log N$ operations. For large N, this represents a huge improvement over N^2: for 802.11a the problem is reduced to roughly 400 operations instead of 4000, and for schemes such as COFDM, which uses up to $N = 1024$, the improvement is on the order of 100-fold. We provide a very brief discussion of this important class of algorithms.

To see how the FFT works, we first examine in more detail the problem of extracting an approximation to the Fourier transform of a time signal sampled at discrete intervals τ. For simplicity we'll limit ourselves to an even number of points. The operation we wish to perform is shown schematically in Figure 1.31: we wish to convert N samples ($N = 16$ here) in time into an estimate of the Fourier transform of the signal at points $k\delta$, where $\delta = 1/N\tau$; k ranges from $-(N/2)$ to $(N/2)$.

It is easy to see why the number of frequencies is limited to the number of samples. A sine or cosine with a frequency of $(N + 1)\delta$ has the same values at the sample points as one with frequency δ: $\cos[2\pi(N + 1)\delta\tau] = \cos[2\pi(N + 1)\tau/N\tau] = \cos[2\pi(1 + 1/N)] = \cos[2\pi + 2\pi\delta\tau] = \cos[2\pi\delta\tau]$. This phenomenon is known as *aliasing*. By the same argument, the values are the

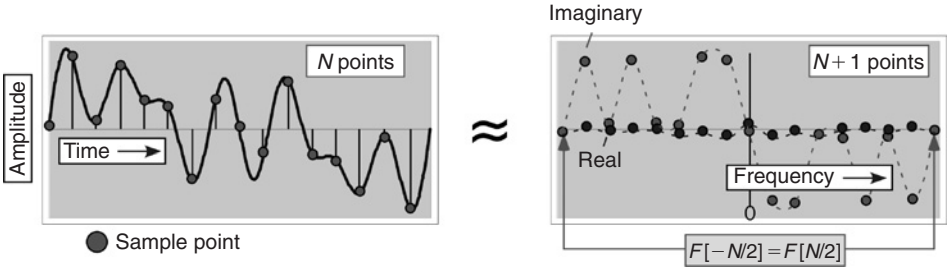


Figure 1.31: Discrete Fourier Transform of a Signal Sampled in Time

same at $k = \pm(N/2)$. In general, the signals could have complex values: in our example the time data are real, forcing the values in frequency space to be complex conjugates with respect to $f = 0$: that is, $F(-k\delta) = F^*(k\delta)$.

The basic mathematics are shown in equation [1.14]. We approximate the Fourier transform integral as a sum, where at the nth sample point the sampled value of the signal, $f(n)$, is multiplied by the value of the exponential for the frequency of interest, $k\delta$. Substituting for the frequency increment, we find that the sum is actually independent of both the sample increment τ and frequency increment δ and is determined only by the value of N and f for a given value of k. The part in brackets is often called the discrete Fourier transform, with the factor of τ added in later to set the time and frequency scales.

$$\begin{aligned} F[k\delta] &= \tau\sum_n e^{-2\pi ink\delta\tau} f(n) = \tau[f(0) + e^{-2\pi ik\delta\tau} f(1) + e^{-2(\pi)2ik\delta\tau} f(2) + \cdots] \\ &= \tau\sum_n e^{-2\pi ink\left(\frac{1}{N\tau}\right)\tau} f(n) = \tau\sum_n e^{-2\pi ik\left(\frac{n}{N}\right)} f(n) \\ &= \tau\left[f(0) + e^{-2\pi ik\left(\frac{1}{N}\right)} f(1) + e^{-2\pi ik\left(\frac{2}{N}\right)} f(2) + \cdots\right] \end{aligned} \tag{1.14}$$

To clarify this perhaps confusing expression, we depict the sums in tabular form in Figure 1.32 for the simple case of $N = 4$. A transform at normalized frequency k is obtained by summing all the terms in the kth row of the table.

The index k is conventionally taken to vary from 0 to $(N - 1)$; because of aliasing as noted above, the $k = 2$ row is the same as $k = -2$ and $k = 3$ is the same as $k = -1$.

Note that the sum of each row can be rewritten slightly by grouping the terms into those for which n is even and odd, respectively. An example of such a rearrangement is shown in Figure 1.33. Although so far we haven't actually saved any calculations, note that after the regrouping, both terms in brackets look like transforms with $N = 2$.

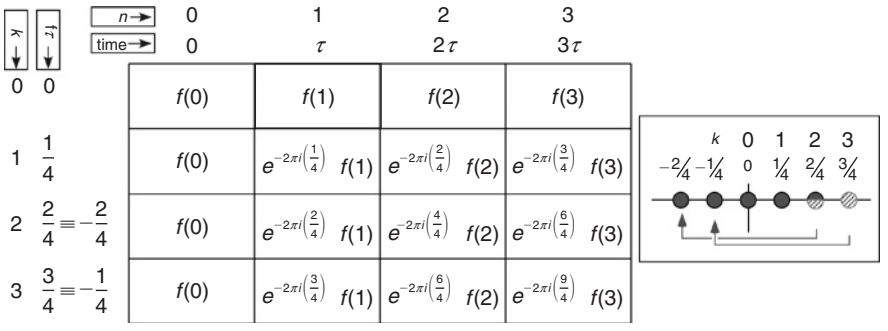


Figure 1.32: Calculation of the Discrete Fourier Transform for $N = 4$

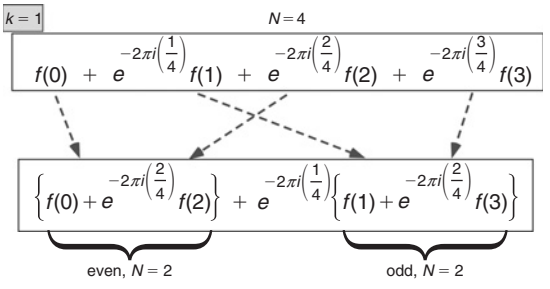


Figure 1.33: Regrouping the Sum Along a Row

When we now view the whole table in this light, we can see the utility of this rearrangement. In Figure 1.34 we show the sum for the odd rows $k = 1, 3$. In Figure 1.34(a), the sum of each row is written in recursive fashion as the sum of two transforms of order $N = 2$, denoted as 2E or 2O for even or odd, respectively. Each of these can be further regarded as the sum of two transforms of order $N = 1$ (which are just the time samples themselves). But we note that the transform $F_{2E,k=1}$ is exactly the same as the transform $F_{2E,k=3}$: only the multiplying factor is different. The same is true for the pair of odd transforms. Thus, as shown in Figure 1.34(b), only half as many calculations as we would naively suspect are actually needed. The same argument naturally can be used for the $k = 0, 2$ rows. To summarize, at first glance one would believe that to evaluate the table of Figure 1.32 would require nine multiplications (not counting the "$\times 1$" terms) and 16 additions, or 25 operations. When regarded recursively, only five multiplications and eight additions, or 13 operations, are required.

In general, because the approach is recursive, we see that an $N = 8$ transform will be constructed from two $N = 4$ algorithms, so that an additional addition and multiplication will be required for each final row (i.e., eight more of each operation, though some will be trivial), rather than $8^2 - 4^2$ additional operations. Each time we double the number of points, we incur an additional N operations: the complexity of the whole process is approximately $N \log_2 N$ instead of N^2.

A closer inspection will also discover that many of the multiplications are actually mere changes in sign: for example, in the $k = 1$ and $k = 3$ cases in Figure 1.34(b), the multiplying factors differ by $e^{-\pi i} = -1$. By some additional ingenuity, the multiplications can be reduced to a small number of nontrivial factors for any given N. More detailed treatments of the FFT are available in any number of texts and web locations; some are provided in the suggested reading at the end of the chapter.

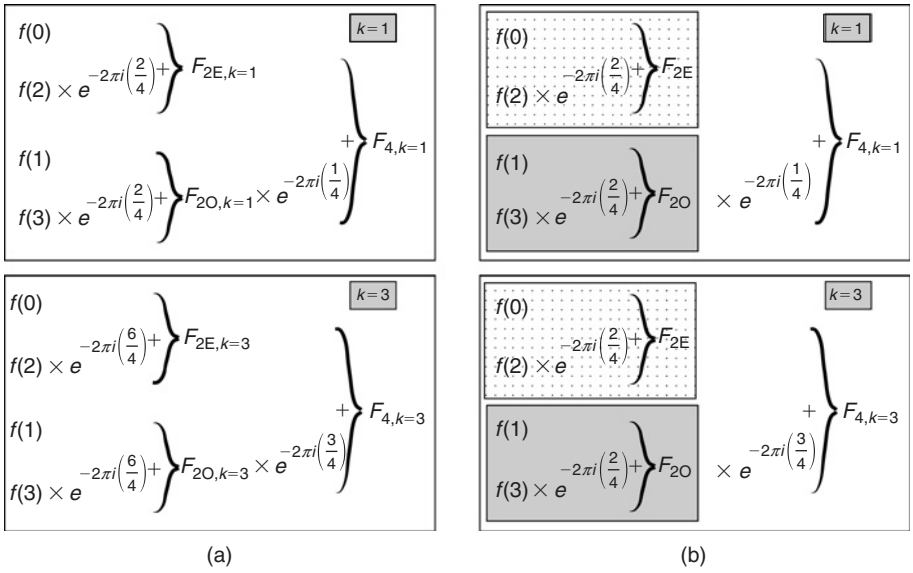


Figure 1.34: Recursive View of Discrete Fourier Transform

We can now summarize the construction of an OFDM symbol, as depicted in Figure 1.35:

- We start with a serial data set, grouped as appropriate for the modulation to be used. For example, if QPSK were to be used on each subcarrier, the input data would be grouped two bits at a time.
- Each data set is converted into a complex number describing the amplitude and phase of the subcarrier (see Figure 1.20 for the QPSK constellation diagram) and that complex number becomes the complex amplitude of the corresponding subcarrier.
- We then take an inverse FFT to convert the frequency spectrum into a sequence of time samples. (The inverse FFT is essentially the same as the FFT except for a normalization factor.) This set of numbers is read out serially and assigned to successive time slots.
- The resulting complex numbers for signal versus time are converted into a pair of voltages by an analog-to-digital converter (ADC); the real part determines the inphase or I channel and the imaginary part determines the quadrature or Q channel.

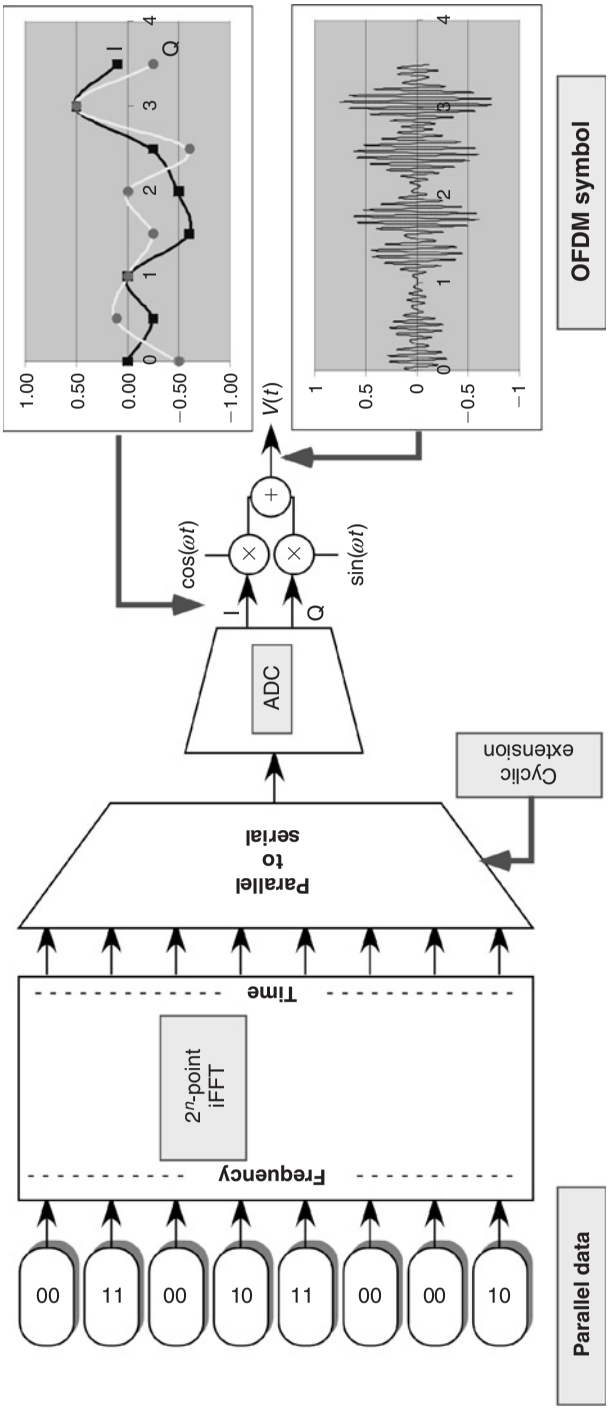


Figure 1.35: Schematic Depiction of OFDM Symbol Assembly for $N = 8$ Subcarriers

- The I and Q voltages are multiplied, respectively, by a cosine and sine at the carrier frequency, and the result is added (after filtering) to produce a real voltage versus time centered around the carrier frequency.

The received symbol is demodulated in a similar fashion: after extraction of the I and Q parts by multiplication by a cosine and sine, a digital-to-analog converter produces a serial sequence of complex time samples, which after removal of the guard interval are converted by an FFT to (hopefully) the original complex amplitude at each frequency.

Notice in Figure 1.35 that the behavior of the resulting OFDM symbol is quite complex, even for a modest number of subcarriers. In particular, large excursions in voltage occur in a superficially unpredictable fashion: the ratio of peak power to average power is much larger than for, for example, the QPSK signal contained in each subcarrier (refer to the right-hand side of Figure 1.20). The absolute peak power relative to the average power grows linearly with the number of subcarriers, because the peak power is the square of the sum of the voltages (N^2), whereas the average power is the sum of the individual average powers (N). Fortunately, as the number of subcarriers grows large, the few symbols with all the subcarriers in phase grow increasingly rare as a percentage of the possible symbols, but for practical signals, such as those encountered in 802.11a/g, peak-to-average power ratios nearing 10 dB are encountered. We discuss the consequences of high peak-to-average ratios in more detail in Chapter 3.

In Figure 1.35, almost all the heavy lifting associated with this specialized modulation is performed in the digital domain: analog processing is very similar to that used in a more conventional radio, though as one might guess from the discussion of the previous paragraph, the radio specifications are more demanding in some respects. This is characteristic of advanced communications systems: because digital functions continue to decrease in both power and cost as long as the integrated circuit industry can keep scaling its products according to Moore's law, whereas analog component cost and size do not scale with feature size, it is generally advantageous to add as much value as possible in digital signal processing. It is interesting to note that significant obstacles are being encountered in scaling digital circuits as metal-oxide-semiconductor field-ving force in electronic design until a new generation of clever techeffect transistor (MOSFET) gate oxides become comparable in thickness with a single monolayer of SiO_2. It remains to be seen whether a truly viable substitute for thermal silicon dioxide can be found or whether Moore's law will slowly fade as a drinologies comes about.

1.3.3 Ultrawideband: A License to Interfere (Sort of)

In section 1.2 we learned that to share the electromagnetic medium, simultaneous users agree to occupy different frequencies. In section 1.3 we found that the need to modulate forces

those users to take up not infinitesimal slices but sizable swathes of spectrum, the extent being proportional to the amount of data they wish to transfer, subject to Shannon's law. The combination of these two facts means that there is a finite amount of data that can be sent in a given geographical region with a given chunk of usable spectrum without having the various users interfere with one another so that nothing gets through. This fact has been recognized in a general way since the early twentieth century, when such incidents as the sinking of the ocean liner *Titanic* dramatized the difficulties of uncoordinated sharing of the radio medium. The solution has traditionally been for regulatory bodies in each nation or supranational region to parcel out the available spectrum to specific users and uses. As one might expect, after a long time (nearly a century) under such a regimen, it becomes difficult to obtain large chunks of "new" spectrum for "new" uses.

Various innovative responses to this conundrum have arisen. Cellular telephony uses conventional licensed spectrum allocations, but the service providers limit the power and range of individual transmitters so that the same spectrum can be reused at a number of locations, increasing the number of total users on the system: this is an example of spatial multiplexing. WLANs and other users of unlicensed spectrum go further, by limiting transmit power and using interference-minimizing protocols to allow a sort of chaotic reuse of the same frequency bands with at least graceful degradation of performance. In the last 10 years or so, a more radical approach has been advocated: that new users simply reuse whole swathes of spectrum already allocated for other purposes but with appropriate precautions to minimize the disruption of existing users while providing new data communications and other services. This approach has come to be known as *ultrawideband* (UWB) communications.

The basic approach to making UWB signals tolerable to other users is to keep the transmitted power in any given conventional band sufficiently small. As we discuss in more detail in section 4, any radio at a finite temperature has a certain noise threshold due to both intrinsic thermal noise and limitations of the radio's components: signals smaller than this noise level will have essentially no effect. The approach taken by UWB is to transmit a modest but appreciable power level but to spread it so thinly over the conventional bands that any conventional receiver reasonably far from the transmitter sees only a negligible addition to normal noise levels. The user of a special UWB receiver combines all the UWB power back into a single signal, which is larger than the background noise and thus detectable. There are several approaches to performing this nifty trick, but in this section we take a brief look at *pulse UWB* radios. Pulse radios are quite different from the modulation approaches we discussed so far and harken back to the original spark-gap transmitters, like that of the *Titanic* and its contemporaries, used in the pre–Federal Communications Commission (FCC) era of radio communication.

A pulse radio, instead of using a modulated carrier of a single frequency, send a series of brief pulses in which each pulse consists of one or at most a few excursions of the signal voltage

away from 0 (Figure 1.36). It is important to note that what is depicted in Figure 1.36 is not a baseband signal before modulation but the actual transmitted voltage. Typical pulses lengths are less than a nanosecond.

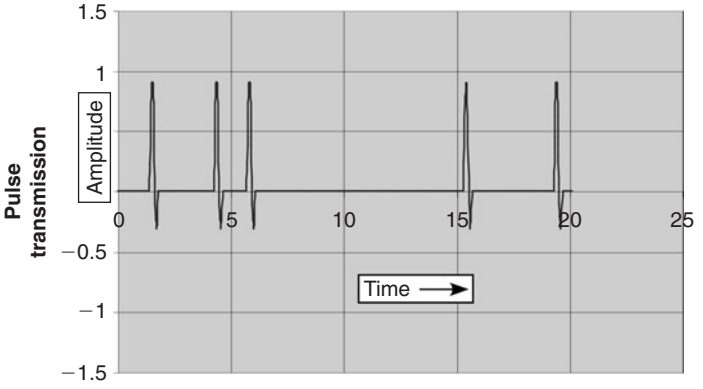


Figure 1.36: Pulse Transmission Voltage vs. Time

To see why such a signal can be considered as UWB, let's try to take the Fourier transform of a pulse following the procedure we used in section 2: we multiply the pulse by a sinusoid at the frequency of interest and integrate. This operation is shown in Figure 1.37.

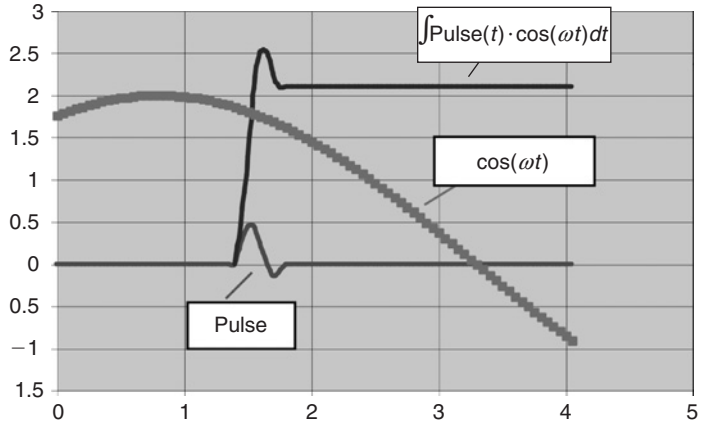


Figure 1.37: What Is the Component of a Short Pulse at Frequency ω?

It is easy to see that as long as the period of the sinusoid is long compared with the pulse, the integral of their product will be rather insensitive to ω. During the pulse, $\cos(\omega t)$ a constant value or perhaps at most a linearly increasing or decreasing voltage. If the pulse is, for example, 0.3 nsec long, one would expect that the resulting frequency spectrum would be roughly flat up to a frequency of about $(1/0.3 \times 10^9)$ or 3.3 GHz (Figure 1.38). Because the total energy of the pulse is distributed over this broad band of frequencies, the amount of

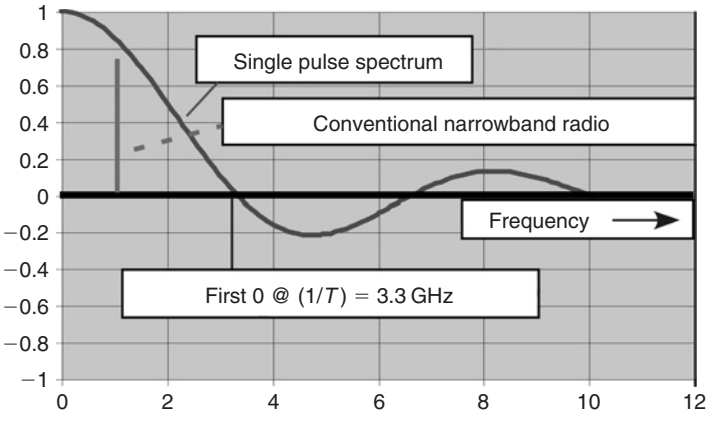


Figure 1.38: Representative Spectrum of a Short Pulse

energy collected by a conventional narrowband receiver would be a small fraction of the total energy transmitted. The exact shape of the spectrum will depend on the detailed shape of the pulse, but the general features will be similar for any sufficiently short pulse.

We can easily envision a UWB communications scheme based on short pulses. A pulse could represent a 1 and the absence of a pulse 0; by implication, one would need to have defined time slots in which to look for the pulses. Such a scheme is known as *pulse-code modulation* and was used in what one could regard as the archetype of all digital communications systems, wired telegraphy using Morse code. A more sophisticated approach might be to displace successive pulses in time in a direction corresponding to the value of the applicable data bit: *pulse-position modulation*. However, any simplistic scheme of this nature runs into a serious problem when applied within the constraints of the UWB environment, that the power in any conventional radio should be small. Let us examine why.

Consider a periodic pulse train, that is, a series of identical pulses evenly spaced in time. If we multiply by a sinusoidal signal whose period is the same as that of the pulse train, successive pulses will contribute exactly the same amount to the integral (Figure 1.39(a)); if we multiply by some other sinusoid with a random incommensurate period, the contributions from the various pulses will vary randomly and sum to some small value. The result, depicted in Figure 1.39(b), is that the spectrum of a periodic pulse train is strongly peaked at frequencies whose periodicity is an integer multiple of the pulse train. Because any of these peaks may lie within a conventional radio's bandwidth, the overall pulse power must be reduced, or *backed off*, by the height of the peak, decreasing the range of the UWB radio.

Although a pulse-code-modulated stream is not exactly like the simple periodic pulse stream, because the presence or absence of a pulse varies randomly with the incoming data, it will still have a peaked frequency spectrum, limiting the transmit power of the UWB radio. Fortunately, there are several approaches to smoothing the spectrum to resemble that of a single pulse.

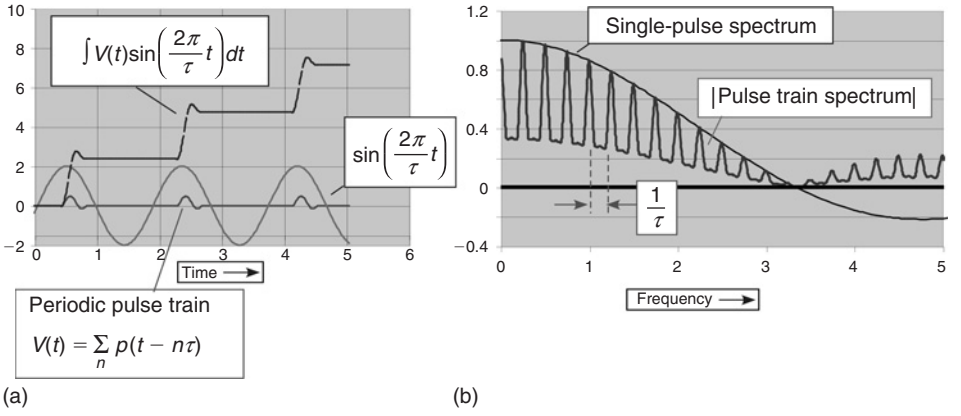


Figure 1.39: Spectrum of a Periodic Pulse Train

One simple method is called *biphase modulation*. In this approach, 1s and 0s are encoded with positive or negative pulses, respectively. It is easy to see that in the case of random or randomized data, the integral in Figure 1.39(a) would become 0 instead of accumulating, because the contribution of each pulse would cancel that of its predecessor. In general, if the data sequence is effectively random, the spectrum of the pulse train is found to be very smooth, approaching an ideal single-pulse spectrum in the limit of a perfectly random data stream. Biphase modulation is in some sense a generalization of BPSK and shares its nice properties of simplicity and noise robustness.

Biphase modulation can be regarded as the multiplication of each data bit by a simple randomizing sequence of either $+1$ or -1. A generalization of biphase modulation is the pseudo-noise encoded biphase modulated wavelet. This elaborate name merely means that instead of multiplying a data bit by a single value, each bit is multiplied by a longer sequence of pulses, where the sequences are chosen to "look" random, that is, they have the spectral characteristics of random noise even though, being found by some deterministic algorithm, they are not truly random. Such a scheme avoids the possibility of spectral lines resulting from inconvenient data (essentially long strings of 1s or 0s), while preserving the benefits of biphase modulation.

A different approach to eliminating spectral lines is to (pseudo-)randomly shift the position of the pulses with respect to their nominal periodic appearance, an approach known as dithering. As the magnitude of the average displacement increases, a dithered pulse stream's spectrum becomes smoother, approaching that of an ideal single pulse. However, as the amount of dithering increases, it also becomes more difficult to synchronize the receiver: how does the receiver figure out what the nominal pulse position is so that it can subtract the actual position from this nominal value to get the displacement?

By using very long sequence lengths either for pseudo-noise encoded biphase modulation or pseudo-random dithered sequences, extremely low signal powers can be demodulated. In formal terms, such techniques use a spread-spectrum approach and exploit spreading gain in the receiver. At heart, these are merely fancy averaging techniques, pseudo-random equivalents of sending the same signal again and again and taking the average of the result to improve the *S*/*N*. The benefit of such an approach—the spreading gain—is essentially proportional to the ratio of the bandwidth used by the transmission to the bandwidth actually required by the data. If we use a UWB radio, with a bandwidth of, for example, 1 GHz, to send a 10 kbps data stream, we have a spreading gain of $(10^9)/(10^4) = 100{,}000{:}1$. This is equivalent to saying that we get to transmit the same signal 100,000 times and take the average: noise and interference average to 0 and the wanted signal grows. By using spreading gain we can use power levels that are imperceptible to other radios and still obtain ample *S*/*N*s.

However, note that if we increase the data rate, the spreading gain decreases; in the limit where the data rate is 1 Gbps, there is no spreading gain at all. Furthermore, just like conventional radios, a pulse radio can encounter multipath problems when the inverse of the data rate becomes comparable with the propagation delays in the environment. In an indoor environment, for example, where we expect propagation delays in the 10 s of nanoseconds (Table 1.4), it is clearly easy to demodulate a data stream consisting of pulses at intervals of around a microsecond. Just as in conventional radio, we won't be bothered by a few echoes at, for example, 5 or 10 or 20 nsec, because we know that none of them could be a true data pulse. In fact, we might be able to use the echoes to help us decode the pulse: in such an approach, known as a rake receiver, the total received signal is correlated with a series of delayed versions of the desired signal, each multiplied by an appropriate empirically determined coefficient, and the various delayed versions summed to get a best guess at the transmitted signal. However, if we attempted to increase the data rate to 100 Mbps, requiring data pulses at rates of one every 10 nsec, we would anticipate serious problems distinguishing the transmitted pulse from its echoes.

Receivers for UWB pulse radio signals are somewhat different from the conventional radio components to be described in some detail in Chapter 3. A UWB receiver may be constructed from a bank of correlators combined with a very-low-resolution (1 or 2 bit) ADC. Essentially, the job of the *k*th correlator is to attempt to determine whether a pulse was present at each possible time offset t_k without any need to determine the size or detailed characteristics of the constituent pulse. UWB receivers of this type lend themselves to implementations on complementary silicon metal-oxide semiconductor MOS (CMOS) circuitry, because the receiver is based on a large number of parallel circuit elements, with each element required to act rapidly but not very accurately. In conventional radio terms, the correlators do not need to be very linear, and if a large spreading gain is used, they may not need to be very sensitive either.

In summary, pulse radios represent an interesting new approach to wireless communications. Although no new principles are required to understand them, pulse radios invert the normal

relationships between frequency and time, which results in differences in optimal approaches for encoding, transmission, and detection.

It is also important to note that pulse radios are not the only possible approach to implementing UWB communications. As we discuss in more detail in Chapter 3, an alternative approach involves the use of very fast FFTs (FFFFFFTs?) and a wideband OFDM scheme.

1.4 Wireless Link Overview: Systems, Power, Noise, and Link Budgets

1.4.1 A Qualitative Look at a Link

We now know that a radio system must transmit a modulated signal centered around a known frequency and receive it with sufficient *S/N* to ensure that the demodulated signal is interpreted correctly most of the time. Let us use this knowledge to get a top-level view of how to design a link between the transmitter and receiver so that these requirements are met.

The basic elements of a radio link are depicted in Figure 1.40. A *transmitter* modulates a carrier according to the input signal and produces a sufficiently large output signal. The *transmitting antenna* radiates the output as a propagating wave, hopefully in the direction of the receiver. The *environment* then makes a complete mess of the nice transmitted signal.

The receive antenna converts the propagating wave, and much of the other junk in the environment, into a voltage, which the receiver filters, amplifies, and demodulates to recover something bearing a resemblance to the original input signal. Let us briefly discuss each

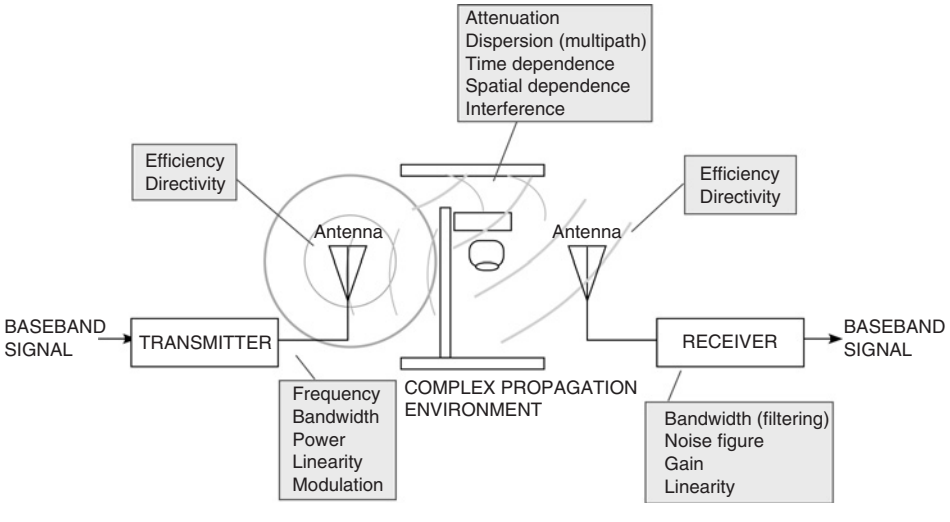


Figure 1.40: Element of a Radio Link

of these components of the link. The transmitter must be designed to operate at the desired frequency. High frequencies offer more absolute bandwidth (and thus more data) for the same percentage bandwidth and can use smaller antennas. However, more advanced and expensive components are needed. The choice of frequency is usually strongly constrained by regulatory requirements, cost, and mechanical limitations. The bandwidth of the transmitted signal must be large enough, in combination with the expected (*S*/*N*), to provide the desired data rate, but one is again typically constrained by the requirements of regulatory bodies and by the choices made in separating the available band into chunks available for specific users: *channelization*. For example, 802.11b operates within the ISM band from 2.4 to 2.483 GHz. The total bandwidth is thus 83 MHz. However, the available bandwidth is partitioned into three noninterfering 25-MHz slots to allow three multiple noninterfering networks in the same place. Within each channel, the actual signal energy is mostly contained in a region about 16 MHz wide and can carry up to 11 Mbps. The transmitter can be designed with various architectures, denoted by the nature of the *frequency conversion* operations used to get the data onto the carrier. The modulation of the carrier by the data can take place in a single step (a *direct conversion* transmitter) or multiple steps (a *superheterodyne* transmitter). The modulated signal must then be amplified to the desired transmit power. More power provides better (*S*/*N*) and thus higher data rates or longer range (or both), but the components are more expensive and consume more DC power, which impacts battery life of portable devices. Total output power is also often constrained by regulation. The transmitter must also transmit a reasonably undistorted version of the modulated signal: it needs good *linearity*. Distorted signals will contain components at undesired frequencies, possibly outside the channel or band assigned to the transmitter, resulting in interference with other radios. The choice of modulation is a trade-off between data rate, linearity, and tolerance to multipath.

After spending the power and money to create a transmitted signal, it is obviously desirable to have that signal converted into a radio wave with good *efficiency*. Antenna efficiency is generally close to 1 for large antennas but degrades for antennas that are very small compared with a wavelength of the transmitted or received signal. However, the size of the antenna may be constrained by the requirements of the application, with little respect for the wavelength: for example, antennas for a self-contained PC-card radio must fit within the PC-card form factor even if larger antennas would work better.

Antennas do not radiate equally in all directions: they have directivity. If you don't know what direction the receiver is in, a nearly isotropic antenna is best. If you do know exactly where you need the waves to go, a highly directional antenna can be used—but directional antennas are physically large and may be inconvenient. Adaptive or "smart" antennas can exploit directivity to minimize the power needed to reach the receiver at the cost of additional complexity.

Antennas must accept signal voltages from the transmitter without reflecting the input power back into the connection: they need to be matched to the transmitter impedance. This matching

typically can only be achieved over a finite bandwidth, which must at least extend over the band the antenna is to be used for. Antennas need to be physically robust enough to survive in service: PC-card antennas will get whacked and banged, and outdoor installations must survive heat, cold, wind, rain, and snow. Finally, esthetics must also be considered: for example, Yagi antennas, although cheap and efficient, are ugly and are typically covered in plastic shrouds.

Radio performance is strongly influenced by the *environment* in which the transmitter and receiver are embedded. In empty space between the stars, the received power will decrease as the inverse square of the distance between transmitter and receiver, as it ought to conserve total energy. Thus, the power received 1 km from the transmitter is 1 million times smaller than that received at 1 m. However, on earth things are more complex. Indoors, the wave will encounter reflection and absorption by walls, floors, ceilings, ducts, pipes, and people as well as scattering by small objects and interference from other radios and unintentional emitters such as microwave ovens. Outdoors, in addition to reflection and absorption by buildings, the earth's often craggy surface, and lakes, rivers, or oceans, the wave can be scattered by cars and trucks and absorbed by foliage or heavy rain. The net result is that the received wave may have been severely attenuated, delayed, replicated, and mixed with other interfering signals, all of which need to be straightened out in the receiver.

The *receiver* begins its task by *filtering* out the band of interest, in the process rejecting most of the interfering signals that weren't rejected by the limited bandwidth of the antenna. The signal is then amplified; the initial stages of amplification are responsible for determining how much noise the receiver adds to whatever was present in the signal to begin with and thus the *noise figure* of the receiver. Receivers can also use single-step *direct conversion* or multiple-step *superheterodyne* architectures. The gain of the receiver—the ratio between the received voltage and the voltage delivered to the ADC—must be large enough to allow a faithful digital reproduction of the signal given the resolution of the ADC. Overall gain is generally inexpensive and does not limit receiver performance, but note that the incoming signal strength can vary over many orders of magnitude, so the gain must be adjustable to ensure that the wildly varying input is not too big or too small but appropriately matched to the *dynamic range* of the ADC.

All the tasks required by a successful radio link must be performed under severe real-world constraints. To keep cost down, the radio will be fabricated if at all possible on inexpensive circuit board material (the laminate FR4 is common). As many components as possible will be integrated into CMOS circuitry, constantly trading off total cost versus the use of large but cheap external discrete components (resistors, capacitors, inductors, etc.). Low-cost radios will have only enough linearity to meet specifications and mediocre noise figure relative to the ideal receiver. Filtering may provide only modest rejection of nearby interferers.

Intentional emitters must be certified by the FCC in the United States, or appropriate regulatory bodies in other jurisdictions, before they can be offered for sale. Regulations will normally

constrain the frequency of transmission, the total power, and the amount of radiation emitted outside the allowed frequency bands. In unlicensed bands, constraints are typically imposed on the modulation approaches and other etiquettes that must be used to minimize interference with other users of the band.

For many uses, size is highly constrained. Add-on radios for laptop computers must fit within a PC-card (PCMCIA) form factor, often at the cost of using inefficient antennas; radios for personal digital assistants are constrained to the still-smaller compact-flash size. Radios for consumers must be simple to use and safe. Portable devices require long battery life, limiting transmit power, and receiver sensitivity. Though there is no evidence that radio waves in the ordinary ambient represent a safety concern, near the transmitter power levels are higher, and consideration must be given to the health of users. Antenna designs that minimize power delivered to the user may be desirable. The resulting product must dissipate the heat generated by its operations and survive thermal, mechanical, and electrical stress to perform reliably for years in the field. All these constraints must be met while minimizing the cost of every element for consumer products.

A key to constructing such high-performance low-cost radio links is to use as much digital intelligence as possible. Wireless links are complex, time varying, and unreliable. The digital hardware and protocols use modulation, coding, error correction, and adaptive operation to compensate for the limitations of the medium, ideally delivering performance comparable with a wired link without the wires.

1.4.2 Putting the Numbers In

A system or radio designer needs to do more than exhort his or her components to be cheap and good: the designer must be able to calculate what will work and what won't. At the system level, the most fundamental calculation to be performed is an estimate of the *link budget:* the amount of loss allowed to propagation and inefficiency given the transmit power and receiver noise that will still meet the accuracy requirements of the protocol given the modulation used. To perform such a calculation, we first need to introduce a few definitions.

A single-frequency harmonic signal dissipates power P in a resistive load proportional to the square of the signal voltage, just as in the DC case, though the constant of proportionality is divided by 2.

$$V(t) = v_o \cos(\omega t) \rightarrow \langle P \rangle = \frac{v_o^2}{2R} \tag{1.15}$$

Some folks define a root-mean-square or RMS voltage to avoid the extra factor of 2.

$$v_{rms} = v_o/\sqrt{2} \rightarrow \langle P \rangle = v_{rms}^2/R \tag{1.16}$$

The nominal load resistance associated with a power level is usually 50 Ω in radio practice, and when the load impedance is not stated, one may safely use this value. Fifty ohms was historically chosen as a compromise value for the impedance of coaxial cables, about midway between the best power handling (roughly 30 Ω) and the lowest loss (about 75 Ω). Note, however, that in cable television systems power levels are low and loss in the cable is important, so cable television engineering has historically used 75 Ω impedances.

The power levels encountered in radio engineering vary over an absurdly huge range: the transmitted power could be several watts, whereas a received power might be as small as a few attowatts: that's 1/1,000,000,000,000,000 of a watt. Getting this tiny signal up to a power level appropriate for digital conversion may require system gains in the millions or billions. Logarithms aren't just nice, they are indispensable.

Gain is typically measured logarithmically in decibels or dB (the notation dates back to Alexander Graham Bell, inventor of the telephone):

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

Thus, 10 dB represent a gain of 10, and 3 dB is (very nearly) a factor of 2.

To measure power logarithmically, we must choose a reference power level to divide by, because the argument of a logarithm must be dimensionless. In radio practice it is very common to use a reference level of 1 mW. Logarithmic power referred to 1 mW is known as decibels from a milliwatt or dBm:

$$P(\text{dBM}) = 10 \cdot \log_{10}\left(\frac{P_2}{1\,\text{mW}}\right) \tag{1.18}$$

Other units are occasionally encountered: dBmV is dB from a millivolt but is a measure of power not voltage; dBμv (referenced to a microvolt) is also used. One may convert from dBmV to dBm by subtracting 49 dB.

It is a (in the present author's view unfortunate) convention that voltages and voltage gains are defined with the respective logarithms multiplied by 20 instead of by 10. This definition allows a gain to be constant in decibels whether voltages or power are entered into the equation but results in a consequent ambiguity in the conversion of decibels to numerical values: one must know whether a voltage or power is being described to know whether 10 dB is a factor of 10 or 3.

Some examples of the use of logarithmic definitions for harmonic signals are shown in Table 1.5. It is noteworthy that a milliwatt, which seems like a tiny amount of power, represents a sizable and easily measured 0.32-V peak into a 50 Ω load.

Table 1.5: Examples of Power Levels and Corresponding Voltages

Power (W)	Power (dBm)	Peak Voltage (V)	RMS Voltage (V)
1	30	10	7.1
0.1	20	3.2	2.2
0.001	0	0.32	0.22
10^{-6}	−30	0.01	7.1×10^{-3}
10^{-12}	−90	10^{-5}	7.1×10^{-6}

We already alluded several times to the importance of *noise* but have so far said little about its origin or magnitude. There are many possible sources of electrical noise, but at the frequencies of interest for most digital communications, the first contributor to noise we must consider is the universally present *thermal noise* generated by any resistive load at a finite temperature. The physical origin of this noise is the same as that of the blackbody radiation emitted by any object not at absolute zero: by the equipartition theorem of statistical mechanics, all degrees of freedom of a classical system at finite temperature will contain an equal amount of energy, of order kT where k is Boltzmann's constant (1.38×10^{-23} joules/K, but not to fear: we will use this only once and forget it), and T is the absolute temperature in Kelvins, obtained by adding 273 to the temperature in centigrade. (Note that this is only true when the separation between energy levels is small compared with kT; otherwise Fermi or Bose-Einstein statistics must be used and life gets more complex. Thus, the statements we're going to make about electrical noise at microwave frequencies cannot be generalized to the noise in optical fiber systems at hundreds of teraHz.)

Because the resistor voltage is a degree of freedom of a resistor, we must expect that if the resistor is at a finite temperature, a tiny random thermal noise voltage will be present. At microwave frequencies, the amount of power is essentially independent of frequency (for the aficionado, this is the low-frequency limit of the Planck distribution), and so it is convenient to describe the noise in terms of the noise power per Hz of bandwidth. The mean-square voltage is

$$\left\langle v_n^2 \right\rangle \approx 4\mathrm{k}T[BW]R \tag{1.19}$$

The power delivered to a matched load (load resistance = source resistance R, so half the voltage appears on the load) is thus kT[BW] independent of the value of the resistor. This power level at room temperature (300 K—a very warm cozy room with a fire and candles) is equal to 4×10^{-21} W/Hz or −174 dBm/Hz. This is a very important number and well worth memorizing: it represents the lowest noise entering any receiver at room temperature from any source with a finite impedance of 50 V, such as an antenna receiving a signal. The noise floor of any room-temperature receiver cannot be better than −174 dBm/Hz.

How much power does this represent for bandwidths of practical interest? In 1 kHz (the ballpark bandwidth used in some types of cellular phones) the thermal noise is −144 dBm. In 1 MHz (closer to the kind of channels used in WLANs) the corresponding power is −114 dBm. Generally speaking, a receiver must receive more power than this to have any chance of extracting useful information, though specialized techniques can be used to extract a signal at lower power than the noise at the cost of reduced data rates.

Real radios always add more than just the minimum amount of thermal noise. The excess noise is measured in terms of what it does to the *S*/*N*. The noise factor of a receiver is defined as the ratio of output to input (*S*/*N*) and represents the effect of excess noise. The logarithm of the noise factor is known as the noise figure; the great utility of the noise figure is that it can simply be added to the thermal noise floor to arrive at an effective noise floor for a receiver. Thus, if we wish to examine a 1-MHz-wide radio channel using a receiver with a 3-dB noise figure, the effective noise floor will be $(-114 + 3) = -111$ dBm.

We've noted previously that the transmitted signal will ideally decrease as the inverse square of the distance between transmitter and receiver. Let us spend a moment to quantify this *path loss* in terms of the distance and the receive antenna. Assume for the present that the transmitting antenna is perfectly isotropic, sending its energy uniformly in all directions. (Such an antenna cannot actually be built, but any antenna that is much smaller than a wavelength provides a fair approximation.) Further, let us assume the receiving antenna captures all the energy impinging on some effective *A* (Figure 1.41). The collected fraction of the radiated energy—the path loss—is then just the ratio of the area *A* to the total surface area of the sphere at distance *r*. If the transmitting antenna is in fact directional (and pointed at the receiver), more power will be received. The directivity of the antenna is measured in dB relative to an isotropic antenna or dBi; the received power is the sum of the isotropic received power in dBm and the antenna directivity in dBi, for ideally efficient antennas.

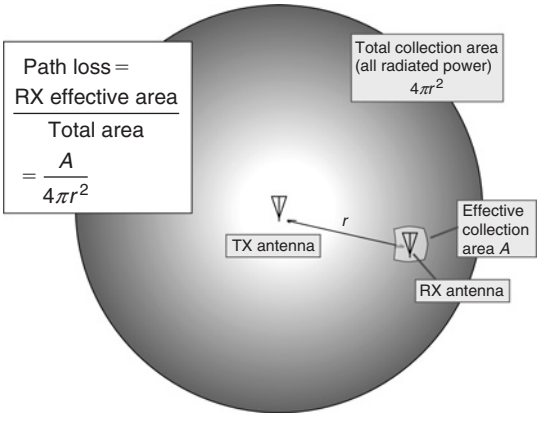


Figure 1.41: Path Loss for an Isotropic Transmitter in Free Space

We're now equipped to take a first crack at figuring out the *link budget* for a plausible radio link. The calculation is depicted graphically in Figure 1.42. We'll set the frequency to 2.4 GHz, though at this point the choice has no obvious effect. The transmit power will be 100 mW, or 20 dBm, which turns out to be a reasonable power for a high-quality WLAN client card or access point. For the present, we'll assume the transmitting and receiving antennas are 100% efficient and deliver all the power they take in either to or from the electromagnetic medium; real antennas may not reach this nice quality level. We'll allow the transmitting antenna to be slightly directional—3 dBi—so that the received signal should be increased by 3 dB over the nominal isotropic value. The receiving antenna, located 60 m from the transmitter, is assigned an effective area of 25 cm^2. Putting it all together, we find that the path loss in free space is about 73 dB. The received signal, compensating for the directionality of the transmitter, is (20 dBm − 73 + 3) = −50 dBm, or 10 nW.

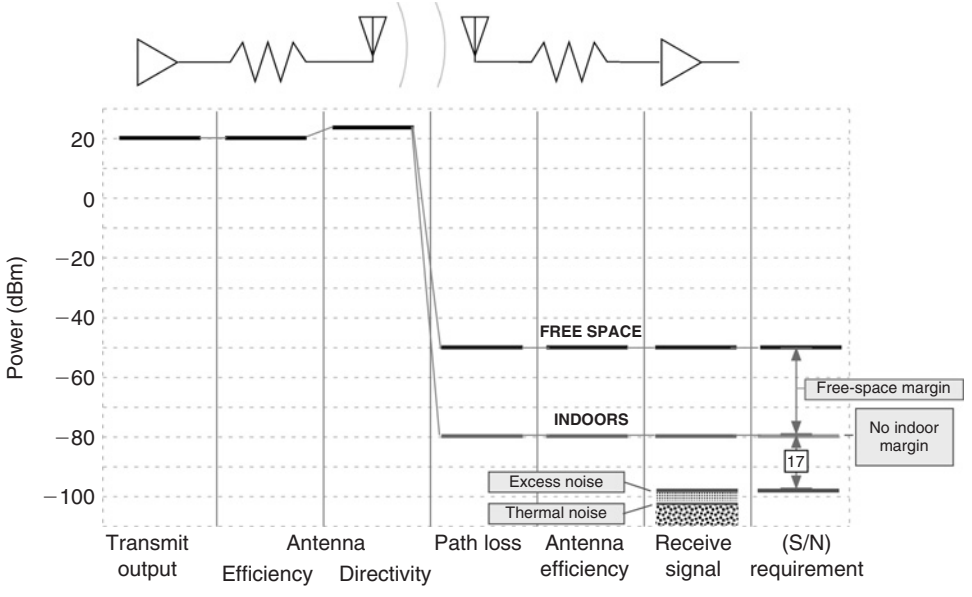


Figure 1.42: Link Budget Example

That doesn't seem like a lot, but recall that what matters is how it compares with the noise. The thermal noise in a 16-MHz Wi-Fi channel is about −102 dBm. Thus, if the receiver is ideal, the (S/N) of the link is (−50 − (−102)) = 52 dB. A realistic receiver might have a noise figure of 5 dB, so the real noise floor is −97 dBm and the (*S*/*N*) is reduced to 47 dB.

How much do we need? Referring back to Table 1.2, we see that various modulations require (*S*/*N*) between around 10 and 25 dB to achieve nearly error-free reception. We'll use a reasonable value of 17 dB (chosen to get round numbers out, of course). In this case, the received (*S*/*N*) of 47 dB exceeds the requirement by 30 dB: we have a *link margin* of 30 dB. With such a

large margin, we could have good confidence that the link would work reliably, despite minor variations in power, noise figure, or distance.

Unfortunately, in a real indoor environment, the path loss can vary widely. It would not be unusual to have a path loss at 60 m indoors of as high as 103 dB (the exact value obviously chosen again to give a round number). In this case our received signal decreases to −80 dBm, the (*S*/*N*) is 17 dB, and we have no link margin at all. A link with no margin will fail to perform to specifications at least half the time. In practice, the radio would attempt to adapt to the circumstances by "falling back" to a modulation with less stringent (*S*/*N*) requirements and thus by implication to a lesser data rate.

The goal of the system designer is to ensure that the link budget for each link is adequate to reliably transport data at the desired rate, despite the constant efforts of nature and humanity to disrupt it.

1.5 Capsule Summary: Chapter 1

Let us briefly recap the main points of this chapter. The goal of a wireless communications system is to exploit propagating electromagnetic waves to convey information. The electromagnetic potential is fundamentally a shared communications medium, so measures must be taken to enable many users to coexist on the medium through multiplexing. The most basic and (almost) universal multiplexing approach is to confine radiation by frequency. However, to transmit information, waves must be modulated, resulting in a finite bandwidth for a given data rate and thus a finite amount of spectrum to share between differing uses and users. The modulation techniques that can be used differ in the amount of information they transmit per symbol, their resilience in the presence of noise, and their tolerance to multipath, but no matter what technique is used, the maximum amount of information that can be transmitted in a single noisy channel is less than Shannon's limit. We examined two specialized modulation techniques, OFDM and pulse radio, of increasing interest in the modern wireless world.

We gave a brief qualitative review of the considerations that go into creating a working wireless link. Once the modulation and data rate are known, the required (*S*/*N*) can be specified. A given radio link can then be examined to see whether this required signal quality can be delivered on the assigned frequency from the power available across the path of interest, despite inevitable uncontrolled variations in the equipment and the environment.

Further Reading

Modulations

Digital Modulation and Coding, Stephen Wilson, Prentice-Hall, 1996: *Includes a nice introduction to probability and information theory as applied to data communications.*

Information Transmission, Modulation, and Noise (4th edition), Steven and Mischa Schwartz, McGraw-Hill, 1990: *Thorough and readable, but dated.*

OFDM

"Implementing OFDM in Wireless Designs," Steve Halford, Intersil; tutorial T204, Communications Design Conference, San Jose, CA, 2002.

UWB/Pulse Radio

"Efficient Modulation Techniques for UWB Signals," Pierre Gandolfo, Wireless System Design Conference, San Jose, CA, 2003.

This page intentionally left blank

CHAPTER 2

Basics of Wireless Local Area Networks

Daniel M. Dobkin

2.1 Networks Large and Small

In this chapter, we provide a brief introduction into the topic of networking in general and review the characteristics of several wireless local network technologies. In this review, the reader will encounter real examples of how the concepts introduced in Chapter 1 are put to use in getting data from one place to another wirelessly.

Our treatment is necessarily cursory and emphasizes aspects relevant to our main topic of wireless transmission and reception, giving short shrift to frame formats, management entities, and many other topics directed toward network implementation.

Data networks have existed in various forms since the invention of telegraphy in the nineteenth century. However, the use of data networks has expanded tremendously with the proliferation of computers and other digital devices. Today, networks extend everywhere, and the Internet—the network of networks—allows computers almost anywhere in the world to communicate with each other using these networks.

Data networks can be organized in a hierarchy based on physical extent. The physical reach of the network is a useful figure of merit both for technical and practical reasons. Information cannot move faster than the speed of light, so as the distance covered by a network grows, its *latency*—the minimum time needed to send a bit between stations—grows too. Latency has important consequences on how networks are designed and operated: if the transmitter has to wait a long time for a reply to a message, it makes sense to send large chunks (*packets*) of information at once, because otherwise the medium will simply be idle while awaiting confirmation. Long links also need low error rates, because the time required to correct an error is long. Shorter links can take a different approach, sending a packet and then waiting for a response, because the time required is not large. Small and large networks are also practically and commercially distinct: small networks are usually owned and operated by their users, whereas large networks are owned and operated by service providing companies who are not the primary users of their capacity.

With that brief preface, let us categorize data networks by size:

- *Personal area networks* (PANs): PANs extend a few meters and connect adjacent devices together. To speak formally, a data connection between a single transmitting station and a single receiving station is a link, not a network, so the connection between, for example, a desktop computer and an external modem (if there are any readers who remember such an impractical arrangement) is a data link rather than a data network. However, more sophisticated cabling systems such as the small computer system interface, which allow sharing of a single continuous cable bus between multiple stations, may be regarded as networks. More recently, wireless PANs (WPANs) have become available, with the ambition to replace the tangle of cabling that moves data between devices today. Wired PANs were purely dedicated to moving data, though some WPAN technologies (such as Bluetooth) also support voice traffic.

- *Local area networks* (LANs): LANs were invented to connect computers within a single facility, originally defined as a room or small building and later extended to larger facilities and multisite campuses. LANs extend a few hundred meters to a few kilometers. They are generally owned and operated by the same folks who own the site: private companies, individuals in their homes, government agencies, and so on. They are generally used indoors and historically have been solely for the movement of data, though the recent implementation of voice-over-Internet-protocol technologies has allowed LANs to provide voice service as well. Ethernet (see section 2.2) is by far the most prevalent LAN technology.

- *Metropolitan area networks* (MANs): MANs connect different buildings and facilities within a city or populated region together. There is a significant technological and historical discontinuity between LANs and MANs: LANs were invented by the computer-using community for data transfer, whereas MANs descended primarily from the telephone network, traditionally organized to move time-synchronous voice traffic. MANs are generally owned by local telephone exchanges (incumbent local exchange companies) or their competitors (competitive local exchange companies). They are organized around a large number of feeders to a small number of telephone central offices, where traffic is aggregated. Most MANs deployed today are based on a hierarchical system of increasingly faster synchronous data technologies. T-1 lines (in the United States) provide 1.5 megabits per second (Mbps) of data over twisted pairs of copper wires and in sheer numbers still dominate over all other connections to the MAN: there are thousands or tens of thousands of these in a large central office. T-1 and T-3 (45 Mbps) are further aggregated into faster connections, usually using fiber optic transmission over synchronous optical network links: OC-3 at 155 Mbps, OC-12 at 622 Mbps, and so on. Traditional MANs support both voice and data transfer.

- *Wide area networks* (WANs): WANs connect cities and countries together. They are descended from the long-distance telephone services developed in the mid-twentieth century and are generally owned and operated by the descendants of long-distance telephone providers or their competitors, where present. Almost all long-distance telecommunications today is carried over fiber-optic cables, commonly at OC-12 (622 Mbps), OC-48 (2.5 Gbps), or OC-192 (10 Gbps) rates; OC-768 (40 Gbps) is in the early stages of deployment. WAN connections cross the oceans and continents and carry the voice and data commerce of the world. The Internet initially evolved separately from WANs, but its explosive growth in the late 1990s resulted in a complex commingling of the traditional WAN and LAN communities, businesses, and standards bodies.

In the simplest view, there are small networks (PANs and LANs) and big networks (MANs and WANs). Small networks deliver best-effort services over short distances and are cost sensitive. Big networks deliver guaranteed-reliable services over long distances and are quality sensitive and cost competitive. This book focuses on how small networks are converted from wires to wireless links.

All networks, big or small, have certain elements in common. A message (typically a packet in a data network) must be addressed to the destination station. The message must have a format the destination station can understand. The message has to get access to some physical medium to be sent. Errors in transmission must be corrected. These activities can be grouped into a hierarchical arrangement that helps provide structure for designing and operating the networks. *Applications* live on the top of the hierarchy and use the networking services. *Networks* support applications and deal with routing messages to the destination station. *Links* transfer data between one station and another in the network and are responsible for access to the physical medium (a wire or a radio link), packaging the packet in an appropriate format readable by the receiving station and then reading received packets and producing the actual voltages or signals. (Note that an important function of networks, *correcting errors* in messages, can go anywhere in the hierarchy and may be going on simultaneously at some or all of the levels, leading on occasion to regrettable results.)

A very widely used example of this sort of hierarchical arrangement is the *Open Systems Interconnect* (OSI) protocol stack. A simplified view of the OSI stack, arranged assuming a wireless LAN (WLAN) link to one of the stations, is depicted in Figure 2.1. The arrangement is in general correspondence with the requirements noted above, though additional layers have been interposed to provide all the manifold functions needed in complex networks.

The standards work of the Institute of Electrical and Electronics Engineers (IEEE), with which we shall be greatly concerned in this chapter, generally divides the data link layer into an upper logical link control layer and a lower medium access control (MAC) layer. In this book, we are almost exclusively interested in what goes on in the link layer (in this chapter) and the physical (PHY) layer (here and elsewhere in the book). The logical link control layer is focused

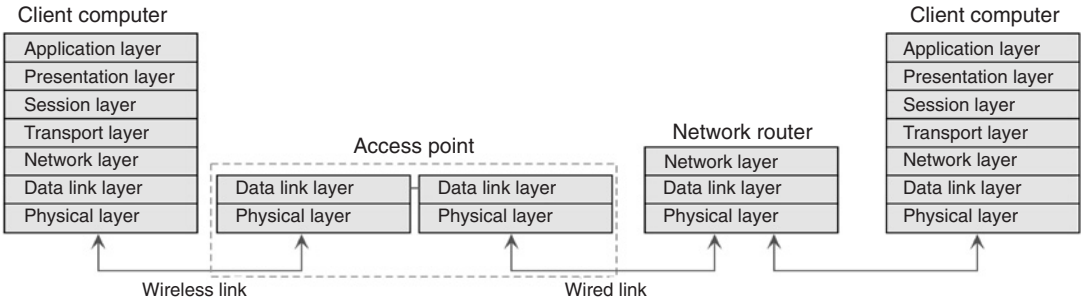


Figure 2.1: The OSI Protocol Stack; the Configuration Shown Is Typical of an 802.11 Infrastructure Network

mainly on the requirements of the higher layers, and in practice the logical link control layer used for many WLAN technologies is the same as the one used for wired technologies. We do not discuss it further. However, the MAC layer is intimately involved in the efficient use of the wireless medium, and it is of great interest for any user of a wireless link.

2.2 WLANs from LANs

LANs connect nearby computers together so that they can exchange data. Ethernet, by far the most popular LAN technology today, was invented by Bob Metcalfe of Xerox Palo Alto Research Center in 1973, in connection with the prescient work on personal computers then being done at Palo Alto Research Center. Ethernet is an *asynchronous* technology: there are no fixed time slots assigned to users. There is also no central control of access to the transmitting medium. Instead, Ethernet exploits three basic ideas to minimize the impact of two or more stations trying to use the medium at the same time; two of them depend on the ability of a station to listen to the voltages on the cable while it sends its message. *Carrier sensing* is used by each station to make sure no one is using the medium before beginning a transmission. (Note that in fact the most popular forms of Ethernet use baseband signaling, that is, the voltages on the cable correspond directly to binary bits and are not modulating a high-frequency carrier, so there's no carrier to sense. Why did you expect it to make sense?) *Collision detection*—noticing that the signal on the cable isn't the one the station is sending—allows the station to detect that another station has tried to send a message at the same time it is transmitting. (The combination of these two techniques is often abbreviated *CSMA/CD*, for carrier-sense multiple access with collision detection.) Finally, *random backoff* is used after a collision to ensure that stations won't simply continue to collide: after all stations become silent, each station that has a pending packet to send randomly chooses a time delay before attempting to send it. If another collision results, the stations do the same exercise, but over a larger time window. This approach has the effect of forcing the *offered traffic* to the network to go down when collisions are frequent, so that data transfer continues at a lower rate but the

network doesn't get tangled up: ethernet networks fail gracefully when they get busy. Thus, Ethernet was an early example of a MAC layer designed to allow peer stations to share a medium without central coordination.

Wireless stations are mobile, and access to the wireless medium cannot be as readily constrained as a wired connection. Wireless links are much noisier and less reliable in general than wired links. The MAC and PHY layers of a wireless network must therefore deal with a number of issues that are rare or absent from the provision of a wired link:

- *Getting connected*: Wireless stations by default are likely to be mobile. How does a wireless station let other mobile or fixed stations know it is present and wants to join the network? This is the problem of *associating* with the network. A closely related problem is the need of portable mobile stations to save power by shutting down when they have no traffic, without losing their associated status; the network coordinator must remember who is awake and who is asleep, allow new stations to join, and figure out that a station has left.

- *Authentication*: In wired networks, a station plugs into a cable placed by the network administrator. If you've entered the building, you're presumed to be an authorized user. Wireless propagation is not so well controlled, so it is important for the network to ensure that a station ought to be allowed access and equally important for the station to ensure that it is connecting to the intended network and not an impostor.

- *Medium access control*: The wireless medium is shared by all (local) users. Who gets to transmit what when? A good wireless network protocol must efficiently multiplex the shared medium.

- *Security*: Even if the network refuses to talk to stations it doesn't recognize, they might still be listening in. Wireless networks may provide additional security for the data they carry by *encryption* of the data stream. It is important to note that security may also be provided by other layers of the network protocol stack; the lack of local encryption in the wireless link does not necessarily imply insecurity, and encryption of the wireless data doesn't ensure security from eavesdropping at some other point in the network.

- *Error correction*: The wireless medium is complex and time varying. There is no way to transmit all packets without errors. The link layer may insist on receiving a positive *acknowledgment* (ACK) before a station may consider a transmission successful. Incoming data packets may be *fragmented* into smaller chunks if packet errors are high; for example, in the presence of microwave oven interference (see Chapter 7), which peaks in synchrony with power lines at 60 Hz, fragmentation may be the only way to successfully transmit very long packets. Fragmented packets must be identified as such and reassembled at the receiving station.

- *Coding and interleaving*: Errors are inevitable in the wireless medium. Protection against bit errors can be provided by encoding the data, so that errors are detected by the receiving station. Codes often allow errors of up to a certain size to be corrected by the receiver without further discussion with the transmitting station: this is known as *forward error correction*. Interleaving is the process of redistributing bits into an effectively random order to guard against bursts of errors or interference destroying all the neighboring bits and thus defeating the ability of the codes to correct local errors.

- *Packet construction*: Data packets get *preambles* prepended onto them. The preambles typically contain synchronization sequences that allow the receiving station to capture the timing of the transmitter and, in the case of more sophisticated modulations, may also allow the receiver to determine the carrier phase and frequency. The preambles also contain digital information specific to the wireless medium.

- *Modulation and demodulation*: The resulting packets must then be modulated onto the carrier, transmitted on the wireless medium, and received, amplified, and converted back into bits.

Let us examine how these problems are addressed in some current and upcoming WLAN and WPAN technologies.

2.3 802.11 WLANs

In 1985 the U.S. Federal Communications Commission (FCC) issued new regulations that allowed unlicensed communications use of several bands, including the 2.4-GHz band, that had previously been reserved for unintended emissions from industrial equipment. Interest in possible uses of this band grew, and in the late 1980s researchers at NCR in Holland, who had experience in analog telephone modems, initiated work to develop a wireless data link. The initial experimental units used an existing Ethernet MAC chip. As discussed in more detail in section 2.3.2, there is no easy way to detect a collision in a wireless network; the experimenters worked around this limitation by informing the MAC chip that a collision occurred any time a positive ACK was not received, thus initiating the Ethernet backoff algorithm. In this fashion they were able to make working wireless links with minimal modification of the existing Ethernet MAC. The early workers understood that to make a true volume market for these products, standardization was necessary. When standardization efforts were introduced at the IEEE shortly thereafter, the MAC was elaborated to deal with the many wireless-specific issues that don't exist in Ethernet but remained based on Ethernet, which had been introduced because of its availability and was successful because of its simplicity and robustness. Thus, the standard today continues to reflect both the strengths and limitations of the original Ethernet MAC.

The IEEE 802 working group deals with standards for wired LANs and MANs. IEEE 802.3 is the formal standardization of the Ethernet wired LAN. The IEEE decided to incorporate WLANs as part of the 802 working group and created the 802.11 activity, culminating in the first release in 1997 of the 802.11 standard. In recent years, in addition to various elaborations of the 802.11 standards, which are discussed below, other working groups have formed within 802 to consider related applications of wireless data links. (In practice, WPANs are also included as part of the 802.15 activity.)

The 802.11 standard actually allowed three physical layers: an infrared link, a *frequency-hopping* (FH) radio link, and a *direct-sequence spread-spectrum* (DSSS) radio link. These links supported data rates of 1 to 2 Mbps. The infrared link physical layer has had little commercial significance (in fact, the author has been unable to find any evidence of any commercial product using this protocol ever having been manufactured). Commercial products were deployed with the FH radio, and the FH approach does offer certain advantages in terms of the number of independent collocated networks that can be supported with minimal interference. However, in terms of current commercial importance, the DSSS physical layer is completely dominant, due to its later extension to higher data rates in the 802.11b standard in 1999. Therefore, here we shall concentrate only on this physical layer. The IEEE standards bodies do many wonderful things, but creation of convenient nomenclature is not one of them; therefore, we introduce the terminology 802.11 classic to refer to the original 1997 802.11 specification and its later releases and in our case predominantly to the DSSS variant. We can then unambiguously use the unmodified moniker "802.11" to refer to all the 802.11 working group standards, including the alphabet soup of elaborations released in later years.

2.3.1 802.11 Architecture

The 802.11 standard allows both infrastructure networks, which are connected to a wired network (typically Ethernet) using an *access point*, and independent networks connecting peer computers wirelessly with no wired network present. Most installations are of the infrastructure variety, and we focus on them. From the point of view of the Ethernet network, which is usually connected to both ends of an 802.11 link, the wireless link is just another way of moving an Ethernet packet, formally known as a *frame* (Figure 2.2), from one station to another (Figure 2.3).

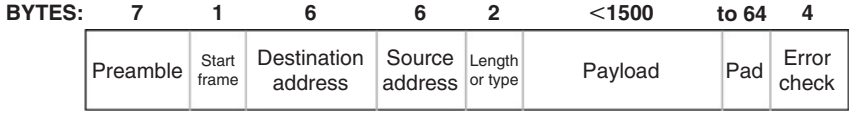


Figure 2.2: An Ethernet (802.3) Frame (Packet)

However, the possibility (well, likelihood) that some or most wireless stations are mobile stations means that the architecture of a wired Ethernet network with wireless stations is likely

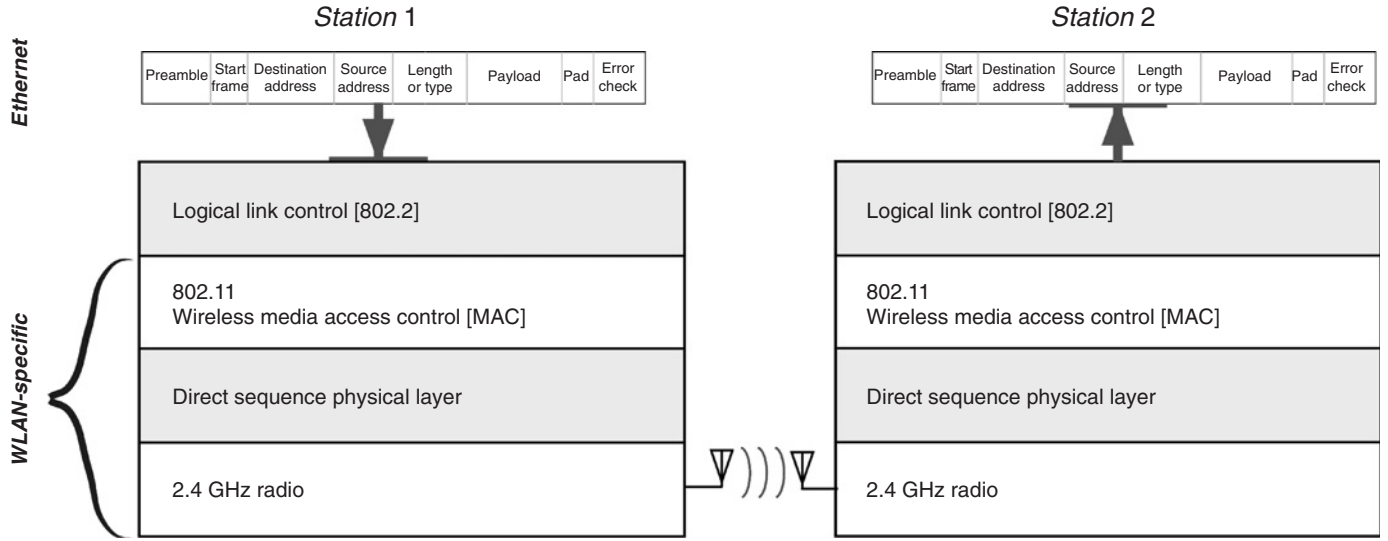


Figure 2.3: 802.11 as a Means of Moving Ethernet Frames

to be fundamentally different from a conventional wired network. The stations associated to a particular access point constitute a basic service set (BSS).

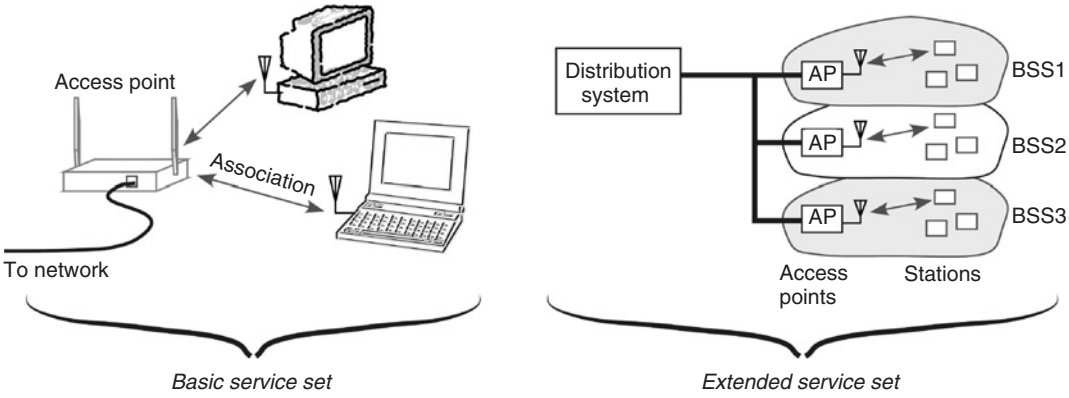


Figure 2.4: Definition of the Basic Service Set (BSS) and Extended Service Set (ESS)

The collection of BSSs connected to a single wired network forms an *extended service set* (ESS) (Figure 2.4). A BSS has a unique identifier, the *BSSID*. An ESS also has an identifier, the *ESSID*, that is unique to that ESS but shared by all the component BSSs (is that enough acronyms in one sentence for you?). Unfortunately, interfaces for many commercial 802.11 implementations use the nomenclature SSID, which in practice usually refers to the ESSID but is obviously somewhat ambiguous and potentially confusing. The ESS architecture provides a framework in which to deal with the problem of mobile stations *roaming* from one BSS to another in the same ESS. It would obviously be nice if a mobile station could be carried from the coverage area of one access point (i.e., one BSS) to that of another without having to go through a laborious process of reauthenticating, or worse, obtaining a new Internet protocol (IP) address; ideally, data transfers could continue seamlessly as the user moved. A distribution system is assigned responsibility for keeping track of which stations are in which BSSs and routing their packets appropriately. However, the original and enhanced 802.11 standards did not specify how such roaming ought to be conducted; only with the release of the 802.11 f *Inter-Access Point Protocol* (IAPP) standard in 2003 did the operation of the distribution system receive a nonproprietary specification. The IAPP protocol uses the nearly universal internet protocol (TCP/IP) combined with RADIUS servers to provide secure communication between access points, so that moving clients can reassociate and expect to receive forwarded packets from the distribution system. In the future, local roaming on 802.11 networks should be (hopefully) transparent to the user, if compliant access points and network configurations become widely established.

2.3.2 MAC and CSMA/CA

Despite its name, Ethernet was conceived as a cable-based technology. Putting all the stations on a cable and limiting the cable length allowed meant that all stations could always hear one another, and any collision would be detected by all the stations on the cable. However, in a wireless network it is unlikely that all stations will be able to receive all transmissions all the time: this is known as the *hidden station* problem. An example is depicted in Figure 2.5: the access point can communicate with stations A and B, but A and B cannot receive each other's transmissions directly. Because of this fact, collision detection would fail if A and B both attempted to transmit at the same time. Furthermore,as we've already seen, received signals are tiny. Provisions to separate a tiny received signal from a large transmitted signal are possible if the two are at differing frequencies, but this requires bandwidth that was not available in the limited 2.4-GHz ISM band. It is difficult and relatively expensive to reliably receive a tiny transmitted signal while simultaneously transmitting at the same frequency, so collision detection is not practical for a WLAN station. Thus, in a wireless implementation, CSMA/CD cannot be used, because there is no way to confidently associate the absence of a carrier with a free medium or reliably detect all collisions.

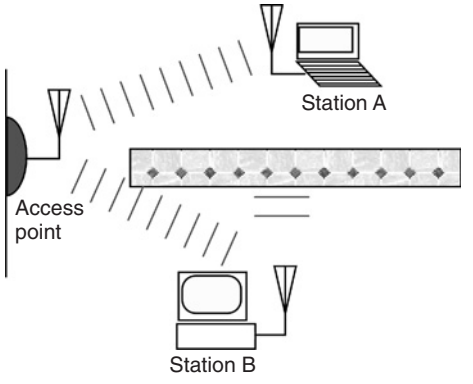


Figure 2.5: Stations A and B Are Hidden From One Another, Though Both Are Visible to the Access Point

To get around this difficulty, 802.11 stations use *carrier-sense multiple access with collision avoidance* (CSMA/CA, Figure 2.6). All stations listen to the current radio channel before transmitting; if a signal is detected, the medium is considered to be busy and the station defers its transmission. A virtual carrier sense mechanism—the *network allocation vector* (NAV)—is provided to further reduce the likelihood of collisions. Each packet header involved in an exchange that lasts longer than a single frame, such as a data frame that expects to be acknowledged by the receiver, will provide a NAV value in the header of the frame. All stations that receive the frame note the value of the NAV and defer for the additional required time even if they can't detect any signal during that time.

Requirements on the timing between frames—interframe spaces—are used to enable flexible access control of the medium using the carrier sense mechanisms. Once a transmission has begun, the transmitting station has captured the medium from all other stations that can hear its transmissions using the physical carrier sense mechanism. A station wishing to send a new packet must first ensure that both the physical carrier sensing mechanism and the NAV indicate the medium to be free for a time equal to the *distributed interframe space* (DIFS). The distributed interframe space is relatively long, so that stations that are completing exchanges already in progress, which are permitted to transmit after a *short interframe space* (SIFS), can capture the medium to transmit their response packets before any new station is allowed to contend for the right to transmit. The NAV value of each packet in a sequence is also set to capture the medium for the entire remaining sequence of packets. Once all the necessary parts of an exchange are complete, the medium may then become free for the distributed interframe space, at which point stations waiting to transmit frames randomly select a time slot and start transmission if no carrier has been detected.

Because collisions cannot be detected, the opposite approach is taken: 802.11 depends on positive acknowledgement of a successful transmission through the receipt of an ACK packet by the transmitting station. Failure to receive an ACK, whether due to poor signal strength, collisions with other stations, or interference, is considered to be indicative of a collision; the transmitting station will wait until the medium is again free, choose a random time slot within a larger possible window (the Ethernet backoff mechanism), and attempt to transmit the packet again.

Two additional optional provisions can be taken to improve performance when, for whatever reason, packets are lost frequently. First, the sending station (typically a client rather than an access point) can precede its data transmission with a *request to send* (RTS) packet. This sort of packet informs the access point and any other stations in range that a station would like to send a packet and provides the length of the packet to be sent. The access point responds after the short interframe space with a *clear to send* (CTS) packet, whose NAV reserves the medium for the time required for the remainder of the data transmission and ACK. The advantage of the scheme is that one can hope that all associated stations can hear the access point CTS packet even if they are not able to detect the RTS packet. An exchange like this is depicted schematically in Figure 2.6. The exchange begins with the client's RTS packet, which reserves the NAV for the whole of the envisioned exchange (the length of the data packet, three short interframe spaces, a CTS frame, and an ACK). The access point responds with clearance, reserving the medium for the remainder of the exchange with its NAV setting. All stations then defer until the end of the exchange. After a distributed interframe space has passed without any more transmissions, a station waiting to send data can use one of the contention slots and transmit.

The second backup mechanism for improving the chances of getting a packet through is *fragmentation*. A large packet from the wired medium can be split into smaller fragments,

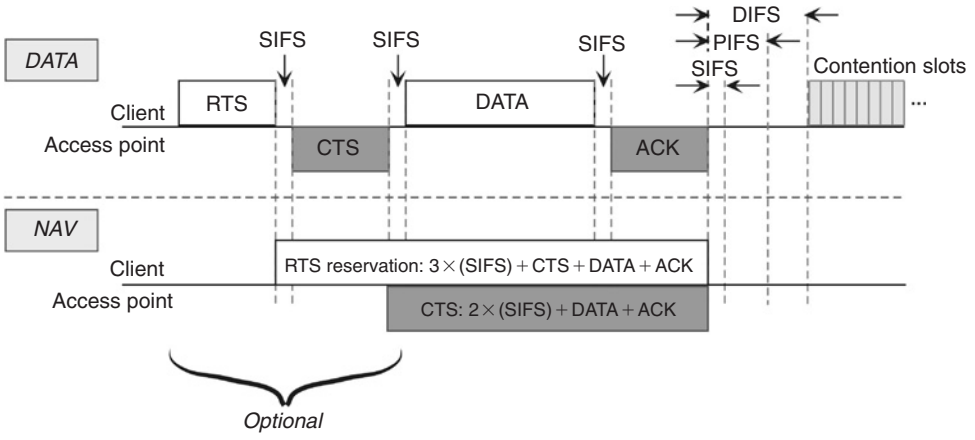


Figure 2.6: Example of Packet Exchange Under CSMA/CA

each of which is more likely to be transmitted without a problem. If an error is encountered in one of the fragments, only that fragment needs to be resent. Like the RTS/CTS exchange, a station transmitting a packet sets the NAV value to reserve the medium for the whole exchange consisting of all the fragments of the packet and an ACK for each fragment.

An important consequence of the MAC is that an 802.11 radio is never transmitting and receiving simultaneously: it is a *half-duplex* system in radio terms. This choice simplifies the design of the radio front end by eliminating the need to distinguish between the powerful transmitted signal and a tiny received signal.

The MAC layer defined by the original 802.11 standard is used almost unchanged for the enhanced versions (802.11b, a, and g) of the standard and thus is worth a moment's reflection. The MAC, although not as simple as the original Ethernet MAC, is nevertheless a fairly basic object. We have not discussed the point-coordination function because it is rarely used. Absent this, the MAC provides no central coordination of contending stations and no guarantees of performance except for the continuity of an ongoing packet exchange.

2.3.3 802.11 Classic Direct-Sequence PHY

To maximize the probability that multiple users could share the unlicensed bands without unduly interfering with each other, the FCC placed restrictions on communications systems that could operate there. A fundamental part of the requirement was that the systems not transmit all their energy in one narrow segment of the band but should spread their radiation over a significant part of the band, presumably a much larger segment than actually required by the data bandwidth, that is, users were required to apply *spread spectrum* techniques (a proviso since relaxed). In this fashion, it was hoped that interference between collocated systems would be minimized.

One approach to spreading the spectrum is to operate on many channels in some pseudo-random sequence; this is known as *frequency hopping* and is used in the classic FH PHY and the Bluetooth (802.15) PHY.[1] It is obvious that if two neighboring radios are operating on separate channels, interference ought to be minimized except when by chance they happen to select the same frequency. The disadvantage of this approach is that the channels must be quite narrow if there are to be a large number of them and therefore little chance of overlap; the 1-MHz maximum width specified in the regulations limits the data rate that can be transmitted at the modest signal-to-noise ratios (S/Ns) expected on low-cost, unlicensed, wireless links (see Chapter 1 for a more detailed discussion of the relationship between bandwidth and data rate).

A quite distinct and rather more subtle approach is called *direct-sequence spread* DSSS. Direct-sequence methods were developed for military applications and are also used in those cellular telephones that are based on code-division multiple access (CDMA) standards. In DSSS, the relatively slow data bits (or more generally symbols) are multiplied by a much faster pseudo-random sequence of *chips*, and the product of the two is used as the transmitted signal. Recall from Chapter 1 that the bandwidth of a signal is determined by the number of symbols per second transmitted, whether or not those symbols contain useful information. Thus, the bandwidth of a DSSS signal is determined by the chip rate, which in general is much larger than the data rate; a DSSS signal can satisfy the FCC's requirement that the signal be spread. Further, the received signal can be multiplied again by the same sequence to recover the original lower data rate. In spectral terms, multiplying the received wide-bandwidth signal by the chip sequence collapses all its energy into a narrower bandwidth occupied by the actual data while simultaneously randomizing any narrowband interfering signal and spreading its energy out. Thus, the use of direct sequence codes provides *spreading gain*, an improvement in the link budget due to intelligent use of extra bandwidth beyond what is needed by the data.

In 802.11 classic DSSS, each data bit is multiplied by an 11-chip Barker sequence, shown in Figure 2.7. (Here data are shown as $+1$ or -1 both to assist the reader in thinking in terms of multiplication and due to the direct translation of the result into binary phase-shift keying [BPSK] transmitted chips; a formally equivalent treatment in terms of modulo-2 addition of binary bits is also possible.) Thus, a 1 bit becomes the sequence ($+1$ -1 $+1$ $+1$ -1 $+1$ $+1$ $+1$ -1 -1 -1) and a 0 bit becomes (-1 $+1$ -1 -1 $+1$ -1 -1 -1 $+1$ $+1$ $+1$).

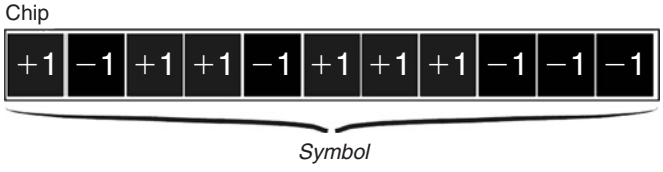


Figure 2.7: Barker Sequence

[1] In a remarkable example of unexpected ingenuity, frequency hopping was invented by the Austrian-born actress Hedy Lamarr, aided by composer George Antheil (U.S. patent 2,292,387), though she received little credit during her lifetime for this achievement.

The Barker sequence is chosen for its *autocorrelation properties*: if two Barker sequences offset in time are multiplied together and the result added (a *correlation* of the two sequences), the sum is small except when the offset is 0. Thus, by trying various offsets and checking the resulting sum, one can locate the beginning of an instance of the sequence readily; that is, synchronization to the transmitter is easy. It is also possible to have several DSSS signals share the same frequency band with little interference by using different orthogonal spreading codes for each signal: codes whose correlation with each other is small or zero. This technique is key to the operation of CDMA cellular phone standards but is not used in the 802.11 standards.

The basic symbol rate is 1 Mbps; each symbol consists of 11 chips, so the chip rate is 11 Mbps. The *spreading gain*, the ratio of the actual bandwidth to that required by the underlying data, is thus 11:1 or about 10.4 dB, meeting the FCC's original minimum requirement of 10 dB. Because the chip rate is 11 Mbps, one would expect the bandwidth required to be modestly in excess of 11 MHz. The actual bandwidth of the transmitted signal depends on the details of precisely how the signal is filtered and transmitted; rather than specifying such implementation-specific aspects, the standard simply defines a spectral mask, which provides a limit on how much power a compliant transmitter is allowed to radiate at a given distance from the nominal center frequency. The spectral mask for 802.11 classic is shown in Figure 2.8. The frequency reference is the nominal frequency for the given channel (about which we'll have more to say in a moment); the amplitude reference is the power density in the region around the nominal frequency. The boxes represent the maximum allowed power density; the smooth line is a cartoon of a typical power spectrum. Observe that, although the standard allows the spectrum to be as much as 22 MHz wide, a typical real signal would have a bandwidth measured, for example, 10 dB from the peak of around 16 MHz.

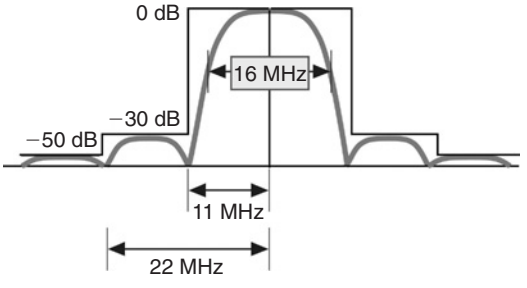


Figure 2.8: 802.11 Spectral Mask

The nominal center frequency for a transmission is chosen from a set of channel frequencies spaced by 5 MHz, shown in Table 2.1. (In the United States, channels 1–11 are available; channels 1–13 are allowed in most European jurisdictions, and channel 14 is for Japan.) Note that if the bandwidth is 16 MHz, transmissions on adjacent channels will interfere with each

Table 2.1: 802.11 Channels

Channel	*f* (GHz)
1	2.412
2	2.417
3	2.422
4	2.427
5	2.432
6	2.437
7	2.442
8	2.447
9	2.452
10	2.457
11	2.462
12	2.467
13	2.472
14	2.484

other quite noticeably. A separation of five channels (25 MHz) is needed to remove most overlap; thus, there are actually only three nonoverlapping channels available in the United States, channels 1, 6, and 11 (Figure 2.9).

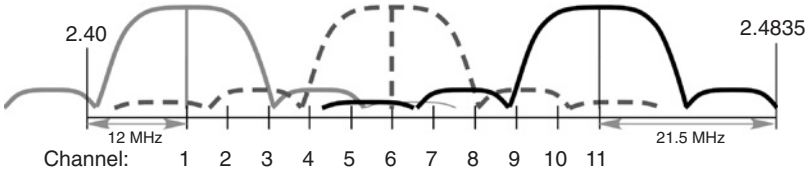


Figure 2.9: Nonoverlapping 802.11 Channels, U.S. ISM Band

Two data rates are supported in 802.11 classic: 1 and 2 Mbps. Both rates use the same chip rate of 11 Mbps and have essentially the same bandwidth. The difference in data rate is achieved by using differing modulations. Packets at the basic rate of 1 Mbps, and the preambles of packets that otherwise use the extended rate of 2 Mbps, use *differential BPSK* (DBPSK), as shown in Figure 2.10. The data portion of extended-rate packets uses differential quaternary phase-shift keying (DQPSK). Recall from Chapter 1 that the bandwidth of a signal is mainly determined by the symbol rate, not by the nature of the individual symbols; because QPSK carries 2 bits per symbol, it is possible to roughly double the data rate without expanding the bandwidth, at a modest cost in required (S/N). In each case, the data bits are scrambled in order before transmission; this procedure avoids the transmission of long sequences of 1s or 0s that might be present in the source data, which would give rise to spectral artifacts.

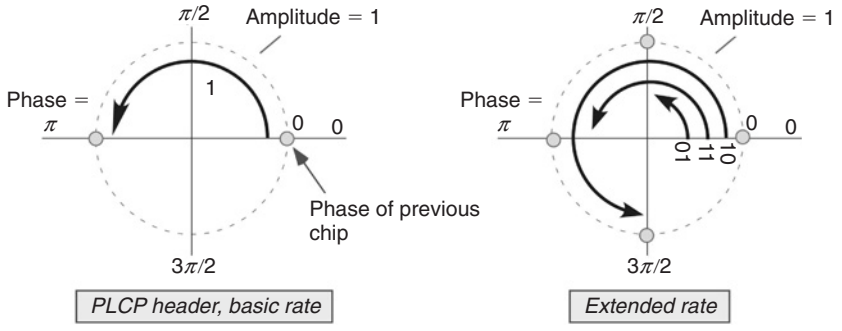


Figure 2.10: 802.11 Classic Modulations: DBPSK (Left) and DQPSK (Right)

The 802.11 modulations are differential variants of the BPSK and QPSK schemes we discussed in Chapter 1, that is, the phase of each symbol is defined only with respect to the symbol that preceded it. Thus, if the phase of a BPSK signal is the same for two consecutive chips, a 0 has been transmitted. (Note that the arrows in Figure 2.10 showing the differences in phase are indicative only of the phase change between the moments at which the signal is sampled, not of the path the signal takes between sample times. This apparently subtle point has significant consequences in terms of the requirements on the radio, as we discuss in Chapter 3.) A moment's reflection will clarify why such a change might be desirable. Imagine that we wish to accurately state the absolute phase of a received QPSK signal by, for example, comparing it with a local oscillator that at the beginning of the packet is locked to the phase of the carrier. A 1500-byte (12,000 bit) packet at 2 Mbps lasts 6000 μsec. During this time, the total phase change of the carrier is 2π (2.4 GHz) (6000×10^{-6} sec) $= 9 \times 10^7$ radians. If we wish the phase of the local oscillator to remain accurate to, for example, $\pi/12$ radians, so that this drift is small compared with the $\pi/2$ radian changes we're trying to resolve between chips, we need the oscillator to maintain phase to about 3 parts per billion. This is implausibly difficult and expensive. (Of course, more clever methods are available to perform such a task at more modest requirements on the hardware; the calculation is done to provide a frame of reference.) On the other hand, to maintain phase lock between successive chips at 11 Mbps, we only need hold this accuracy over 1/11 μsec: 220 cycles, or 1400 radians. The required accuracy in this case is 200 parts per million, which is easily achieved with inexpensive crystal-referenced synthesized sources. The price of this alternative is merely that we lose any information present in the first symbol of a sequence: because we can exploit our synchronization sequence (which doesn't carry any data anyway) for this purpose, it is obvious that differential modulation carries considerable benefits and little penalty.

The maximum transmit power of an 802.11 radio is limited in the standard to comply with the requirements of regulatory bodies. In the United States, power must be less than 1 W, but in practice a typical access point uses a transmit power of about 30 to 100 mW. The standard

then requires that a receiver achieve a frame error rate of less than 8% for 1024-byte frames of 2 Mbps QPSK at a signal power of −80 decibels from a milliwatt (dBm). That sounds like a lot of errors, but because each frame has $1024 \times 8 = 8192$ bytes, the bit error rate is a respectable $0.08/8192 = 10^{-5}$. To understand what the transmit and receive power levels imply, let's do a link budget calculation like the one we performed in Chapter 1. Let the transmitted power be 100 mW (20 dBm). Assume the transmitting antenna concentrates its energy in, for example, the horizontal plane, achieving a 6-dB increase in signal power over that of an ideal isotropic antenna. Assume the receiving antenna has the same effective area as an ideal isotropic receiver. Our allowed path loss—the link budget—is then $(20 + 6 - (-80)) = 106$ dB. Recalling that an isotropic antenna has an equivalent area of around 12 cm^2, this is equivalent to a free space path of 2000 m or 2 km! However, recall that propagation indoors is unlikely to achieve the same performance as unimpeded ideal propagation in free space. For the present let us simply add 30 dB of path loss for a long indoor path, leaving us with an allowed range of $(2000/10^{1.5}) = 63$ m. The transmit power levels and receiver sensitivities envisioned in the 802.11 standard make a lot of sense for indoor communications over ranges of 10 s to perhaps 100 m, just as one might expect for a technology designed for the LAN environment (see section 2.1).

The receive sensitivity requirement also tells us something about the radio we need to build. Recall that a QPSK signal needs to have an (S/N) of about 12.5 dB for reasonably error-free reception. At first blush, one would believe this implies that the noise level of the receiver must be less than $(-80 - 12.5) = -92.5$ dBm. However, recall that 11 QPSK chips at 11 Mbps are "averaged" together (correlated with the Barker sequence) to arrive at one 2-bit data symbol. This averaging process provides us with some extra gain—the spreading gain described above—so that we can tolerate an (S/N) of $12.5 - 10{:}5 = 2$ dB and still get the data right. The noise level we can tolerate in the receiver is thus $-80 - 2 = -82$ dBm. Recalling that the signal is about 16 MHz wide, the unavoidable thermal noise in the receiver is $(-174\,\text{dBm} + 10(\log 16) + 60) = -102$ dBm. The specification has left room for an additional 20 dB of excess noise in the receiver. Receivers with noise figures of 20 dB at 2.4 GHz are very easy to build, and in fact 802.11 commercial products do much better than this. The standard has specified a level of performance that can be reached inexpensively, appropriate to equipment meant for an LAN. The link budget calculation is displayed graphically in Figure 2.11.

2.3.4 802.11 Alphabet Soup

Products implementing the 802.11 classic standard were manufactured and sold in the late 1990s by vendors such as Proxim, Breezecom (now part of Alvarion), Lucent (under the WaveLAN brand), Raytheon, Symbol Technologies, and Aironet (now part of Cisco). However, the maximum data rate of 2 Mbps represented a significant obstacle to the intended target usage model of wirelessly extending Ethernet: even the slowest variant of Ethernet has a native bit rate of 10 Mbps and a true throughput of around 8–9 Mbps. Work had started as

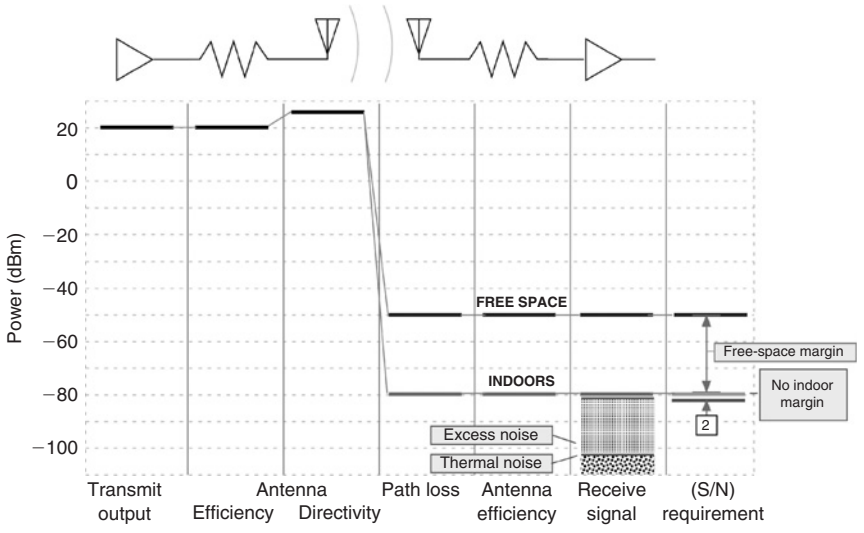


Figure 2.11: Link Budget Estimates for an 802.11–Compliant Radio Link

early as 1993 on improved physical layers, and by 1999 two new higher performance PHY layers were released as enhancements to the original 802.11 classic interfaces. *802.11a* was a fairly radical modification of the physical layer, having essentially no commonalty with the DSSS or FH PHY layers of the classic standard and targeted for operation in a different slice of spectrum around 5 GHz. *802.11b* was a much more conservative attempt to improve performance of the 802.11 PHY layer without changing it very much. The 802.11b standard received the most early commercial attention, including support in the consumer market from Apple Computer's Airport product line and Cisco's Aironet products in the "enterprise" market (LANs for large industrial companies). The elaboration of compatibility efforts begun in association with the release of 802.11 classic led to the formation of the Wireless Ethernet Compatibility Alliance, or WECA, which (recognizing the awkwardness of the IEEE's nomenclature) provided *Wi-Fi* compliance certification to ensure interoperable 802.11b products from multiple vendors. By 2000 it was clear that 802.11b had found a "sweet spot," combining adequate data rates and performance with simplicity and low cost, and was poised for major commercial successes and widespread deployment. Sales volumes grew rapidly, and prices for both client cards and access points fell even more rapidly as numerous vendors entered the field.

The more audacious 802.11a standard, which would provide enhanced data rates of up to 54 Mbps, was seen by most industrial participants as a tool to hold in reserve for future use, but the start-up company Atheros proceeded to release chips implementing the 802.11a standard, with products available from Intel and Proxim in 2002. These products, although modestly successful at introduction, clarified the desirability of combining the high data rates of the 802.11a standard with backward compatibility to the installed base of 802.11b products. The

"G" task group, which was charged to take on this awkward task, after some struggles and dissension approved the 802.11g standard in 2003, noticeably after the release of several prestandard commercial products. At the time of this writing, 802.11g products have nearly replaced 802.11b products on the shelves of consumer electronics stores: the promise of backward compatibility in conjunction with a much higher data rate at nearly the same price is an effective sales incentive, even though many consumer applications do not currently require the high data rates provided by the standard. Dual-band access points and clients, supporting both 802.11a and 802.11b/g, are also widely available.

A number of task groups are also actively attempting to fill in holes in the original standards work and add new capabilities. Task group "I" is charged with improving the limited authentication and encryption capabilities provided in the original standard, about which we have more to say in section 2.3.8. Task group "F" is responsible for providing the missing description of the distribution system functions, so that roaming between access points can be supported across any compliant vendor's products; an IAPP specification was approved in June 2003.

Recall that the original Ethernet standard, and the 802.11 MAC derived from it, deliver *best-effort* data services, with no guarantees of how long the delivery will take. Services such as voice or real-time video delivery require not only that data be delivered but that it arrive at the destination within a specified time window. This requirement is commonly placed under the not-very-informative moniker of *quality of service*. Task group "E" is defining quality of service standards for time-sensitive traffic over 802.11 networks, although to some extent that effort has refocused on the 802.15 work we describe in the next section. Task group "H" is defining two important additions to the 802.11a standard: *dynamic frequency selection and transmit power control*. These features are required for operation in most European jurisdictions, and the recent FCC decisions on spectrum in the United States will also require such capability for operation in most of the 5-GHz bands.

In the next three subsections we examine the important 802.11b and 802.11a PHY layers and briefly touch on their admixture in 802.11g. We also provide a cursory discussion of WLAN (in)security and some cures for the deficiencies revealed therein. We must regrettably refer the reader to the IEEE web sites described in section 2.7 for more details on the other task groups within 802.11.

2.3.5 The Wi-Fi PHY (802.11b)

The 802.11b physical layer uses the same 2.4-GHz band and channelization as the classic PHY. Furthermore, the basic signaling structure of 11 megasamples/second (Msps) of either BPSK or QPSK symbols is unchanged, and so it may be expected that the frequency spectrum of the transmitted signals will be similar to those of the classic PHY. However, the use of these symbols is significantly different.

To maintain compatibility with classic systems, packets with preambles transmitted at the lowest rate of 1 Mbps, using DBPSK modulation, must be supported. However, the new PHY adds the option to use short preambles with 2 Mbps DQPSK modulation to reduce the overhead imposed by the very slow long preamble on short high-rate packets. More importantly, the 802.11b PHY introduces two completely new approaches to encoding the incoming data onto the QPSK symbols: *complementary code keying* (CCK) and *packet binary convolutional coding* (PBCC). Each method may be used to support two new data rates, 5.5 and 11 Mbps. Both methods completely abandon the Barker sequence and conventional direct-sequence spreading.

Let us first examine CCK. CCK is a *block code*: chunks of symbols of fixed size are used as code words, and the subset of *allowed* code words is much smaller than the total possible set of code words. Errors are detected and corrected by comparing the received code word with the possible code words: if the received code word is not an allowed word but is close to an allowed word, one may with good confidence assume that the nearby allowed code word is in fact what was transmitted. Block codes are relatively easy to implement.

We look at the high-rate 11 Mbps coding in detail. In the particular code used in the 802.11b CCK-11 Mbps option, transmitted QPSK symbols are grouped into blocks of eight to form code words. Because each symbol carries 2 bits, there are $4^8 = 65{,}536$ possible code words. Of this large domain, only 256 code words, corresponding to the 8 input bits that define the chosen CCK block are allowed. They are found in the following fashion, shown schematically in Figure 2.12. (Figure 2.12 is headed "evennumbered symbol" because alternate code words have slightly different phase definitions, important for controlling the spectrum of the output but confusing when one is trying to figure out how the scheme works.) The input bits are grouped in pairs (dibits). Each dibit defines one of four possible values of an intermediate phase φ. The intermediate phases are then added up (modulo-2π) as shown in the chart to

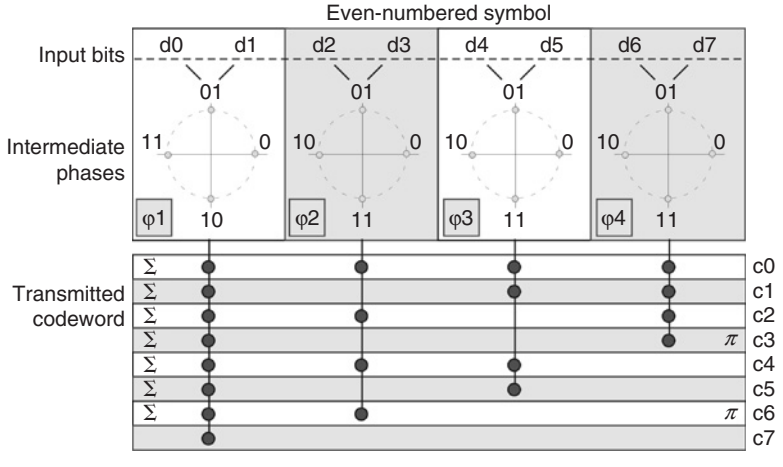


Figure 2.12: CCK-11 Code Word Encoding Scheme (Even-Numbered Symbols)

determine each QPSK symbol c of the transmitted code word. Thus, the last symbol c7 shares the phase of $\phi 1$ (and is used as the phase reference for the remainder of the code word).

The phase of c4 is ($\phi 1 + \phi 2 + \phi 3$). Symbols c3 and c6 have π radians added to the argument (i.e., they are multiplied by $\phi 1$).

This procedure is a bit confusing to describe but easy to implement and ensures that the resulting allowed code words are uniformly distributed among all the possible code words. Recall that we defined the distance between two QPSK symbols by reference to the phase-amplitude plane; for a normalized amplitude of 1, two nearest neighbors are $2/\sqrt{2}$ apart (Figure 2.20, Table 2.1); if we square the distance we find nearest neighbors differ by 2 in units of the *symbol energy*, E_s. We can similarly define the difference between code words by adding up the differences between the individual QPSK symbols, after squaring to make the terms positive definite (>0). Having done so, we find that CCK-11 code words have their nearest neighbors at a squared distance of 8. This means that no possible single-chip QPSK error could turn one allowed code word into another, so single-chip errors can always be detected and corrected. In RF terms, the bit error rate is reduced for a fixed (S/N), or equivalently the (S/N) can be increased for the same error rate. The change in (S/N) is known as *coding gain*. The coding gain of CCK is about 2 dB. A slight variant of the above scheme is used to deliver 5.5 Mbps.

The standard also provides for a separate and distinct method of achieving the same 5.5- and 11-Mbps rates: PBCC. PBCC is based on a *convolutional code* (Figure 2.13). The coding is performed using a series of shift registers, shown here as z^1 through z^6. At each step, a new input bit is presented to the coder and all bits are shifted to the right one step. The "+" symbols represent modulo-2 additions that produce the output bits y_0 and y_1 from the input and the stored bits. The coder can be regarded as a machine with $2^6 = 64$ possible states. In any given state of the coder, two transitions to the next state are possible (which occurs being determined by the new input bit) and two of the four possible outputs are allowed (again determined by the input). Thus, the sequence of allowed outputs is highly constrained relative to the sequence of possible outputs. This code is known as a *rate 1/2* code, because each input bit produces two output bits.

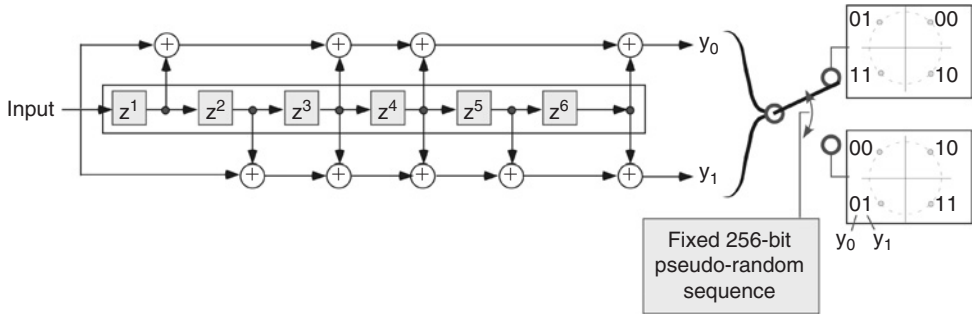


Figure 2.13: PBCC-11 Convolutional Code and Cover Sequence

The output bits are mapped onto the next QPSK symbol in two possible ways (rotated 90 degrees with respect to each other) according to a 256-bit fixed sequence. Because there are 11 megachips per second and each chip carries one input data bit coded into two y's, the net data rate is 11 Mbps. The 256-bit sequence helps to remove any periodicity in the output signal and thus helps smooth the spectrum of the output. The relatively long period of the sequence means that any other interfering transmission is unlikely to be synchronized with it and will appear as noise rather than as valid code words, making it easier to reject the interference.

Convolutional codes are usually decoded with the aid of a *Viterbi trellis* decoder. (Trellis algorithms can also be used to decode block codes.) The trellis tracks all the possible state transitions of the code and chooses the trajectory with the lowest total error measure. It would seem at first glance that such a procedure would lead to an exponentially growing mess as the number of possible trajectories doubles with each additional bit received, but by pruning out the worst choices at each stage of the trellis, the complexity of the algorithm is reduced to something manageable. However, implementation is still rather more complex than in the case of a block code like CCK.

The PBCC-11 code has slightly better performance than the corresponding CCK code: about 3.5 additional dB of coding gain (Figure 2.14). However, the computational complexity of decoding the convolutional code is about 3.5 times larger than the CCK code. To the author's knowledge, the PBCC variant has been implemented only in a few products from D-Link and US Robotics and has enjoyed very little commercial success. Commercial politics may have played a significant role in this history, as Texas Instruments enjoyed certain intellectual-property rights in PBCC that may have made other vendors reluctant to adopt it.

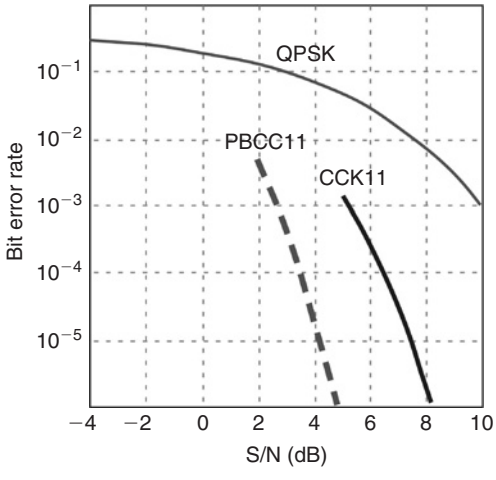


Figure 2.14: Error Rate of Uncoded QPSK, PBCC11, and CCK 11 vs. (S/N) (After Heegard et al., in Bing 2002)

It is also important to note that in most applications of a WLAN, performance is secondary to price and convenience; the modest advantage in link budget of PBCC was not enough to produce a commercial advantage in the marketplace.

2.3.6 802.11a PHY

The 802.11a PHY is a radical departure from the approaches above. The first major change is the use of the Unlicensed National Information Infrastructure (UNII) band at 5.15–5.825 GHz instead of the 2.4-GHz ISM band. At the time the PHY was promulgated, 5-GHz radios implemented in standard silicon processing were not widely available, so this choice implied relatively high costs for hardware. The motivation for the change was the realization that a successful commercial implementation of 802.11 classic and/or 802.11b devices, combined with the other proposed and existing occupants of the 2.4-GHz band (Bluetooth devices, cordless telephones, and microwave ovens, among others), would eventually result in serious interference problems in this band. The UNII band was (and is) relatively unoccupied and provides a considerable expansion in available bandwidth: in the United States, 300 MHz at the time versus 80 MHz available at ISM.

Rules for operating in the UNII band have changed significantly in the United States since the promulgation of the 802.11a standard: in November 2003, the FCC added 255 MHz to the available spectrum and imposed additional restrictions on operation in the existing 5.250- to 5.350-GHz band. The original and modified U.S. assignments are depicted in Figure 2.15. The allowed bands at the time of the standard consisted of a lower, middle, and upper band, each with 100 MHz of bandwidth. The lower band was dedicated to indoor use only and limited to 40-mW output power. The middle band allowed dual use, and the upper band was targeted to outdoor uses with a much higher allowed output of 800 mW.

The FCC's changes in late 2003 added an additional band, almost as big as what had been available, for dual use at 200 mW. Use of this band, and retroactively of the old UNII mid-band, is only allowed if devices implement transmit power control and dynamic frequency selection. The former reduces the power of each transmitting device to the minimum needed to achieve a reliable link, reducing overall interference from a community of users.The latter causes devices to change their operating frequency to an unoccupied channel when possible. These changes in band definitions and usage requirements improve consistency between U.S. and European (ETSI) requirements in the 5-GHz band and will presumably increase the market for compliant devices. The 802.11h standard, approved in 2003 to allow compliant devices that meet European standards, adds several capabilities relevant to power control and channel management. Clients inform access points about their power and channel capabilities; quiet periods are added when no station transmits to enable monitoring of the channel for interference, and access points can use the point coordination function interframe space to capture the medium to coordinate switching channels when interference is detected.

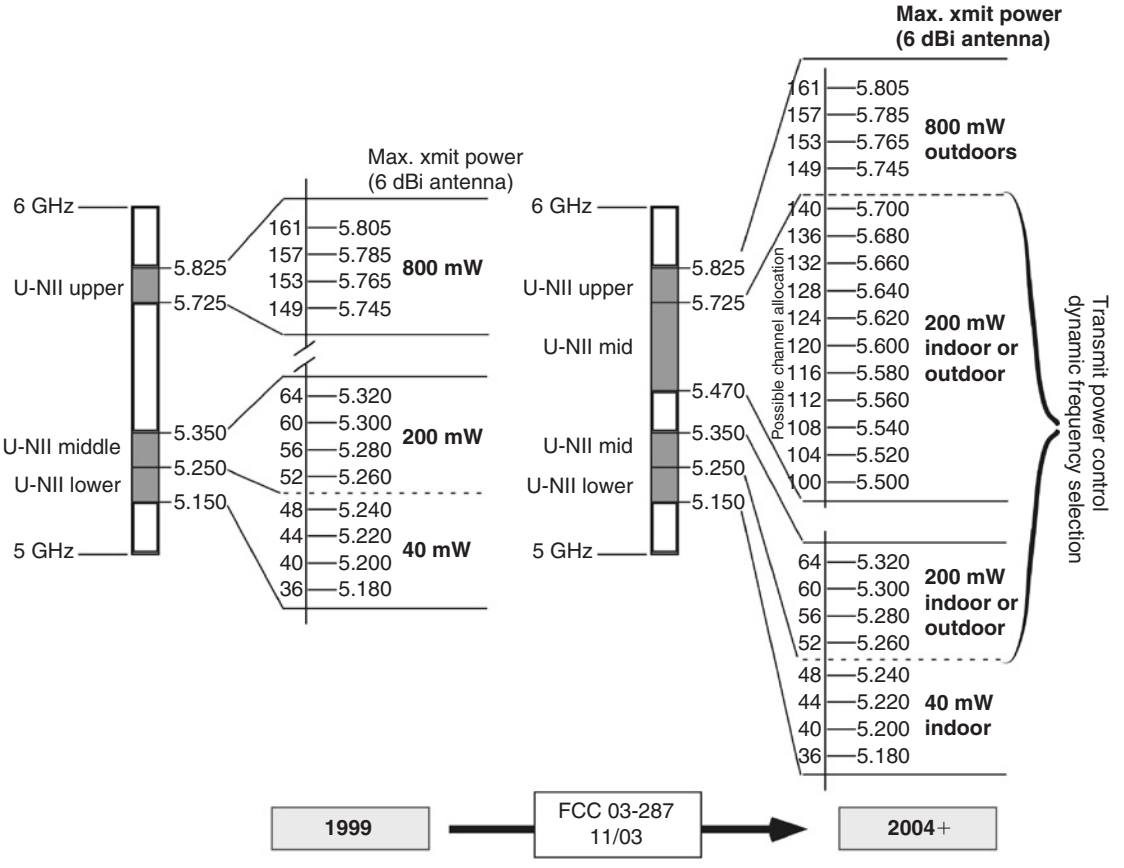


Figure 2.15: UNII Band in the United States, Then and Now

Devices that are compliant to 802.11h should also meet the new FCC standards for UNII operation in the United States.

The spectral mask limiting the bandwidth of the transmitted signal is shown in Figure 2.16. Note that although the mask is somewhat more complex and apparently narrower than the old 802.11 mask (Figure 2.8), in practice both signals end up being around 16-MHz wide at 10 dB down from the maximum intensity at the center of the transmitted spectrum. This design choice was made to allow the same analog-to-digital conversion hardware to be used for the baseband in either standard, which is beneficial in those cases where a single baseband/MAC design is to serve both 802.11a and 802.11b radio chips. This important fact allowed the 802.11a PHY to be kidnapped and transported whole into the ISM band in the 802.11g standard, as we discuss in section 2.3.7.

Nonoverlapping channels in the old band definition, and a possible channel arrangement in the new bands, are shown in Figure 2.17. The assumed channel shape is based on the

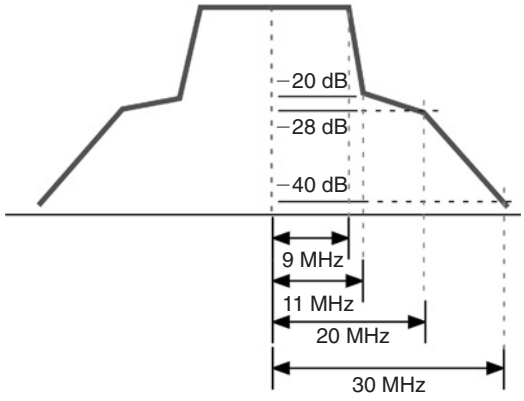


Figure 2.16: 802.11a Transmit Spectral Mask

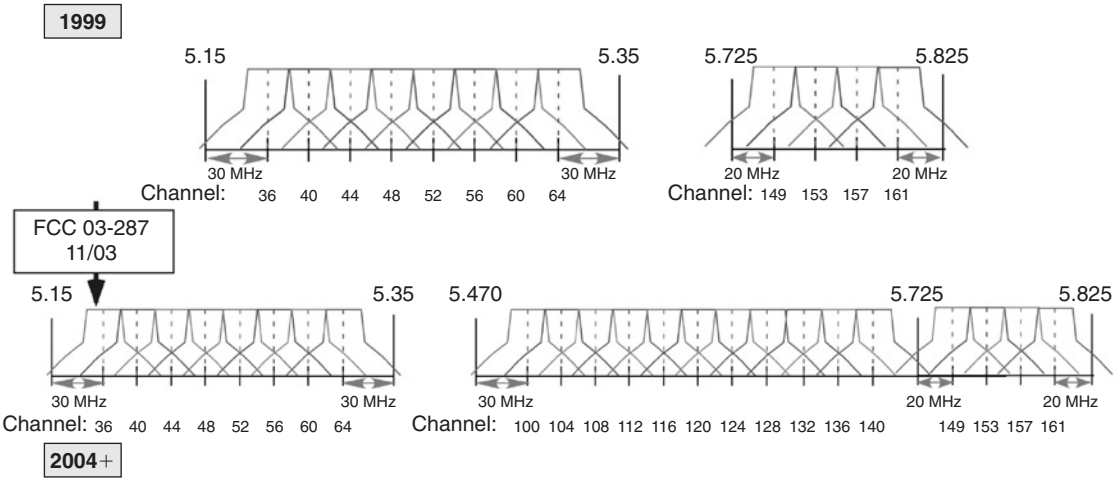


Figure 2.17: Possible Nonoverlapping Channel Assignments in Old and New UNII Bands

spectral mask in Figure 2.16. We can see that the old definition had room for as many as 12 nonoverlapping channels, though operation at the high band at full power in the lowest and highest channels would result in out-of-band emissions exceeding FCC allowances, so that in practice fewer channels might be used. In the new band assignment, there is room for up to 19 nonoverlapping channels in the lower and middle bands alone. Recall that the ISM band allows only three nonoverlapping 802.11 channels (Figure 2.9).

The availability of more than three channels is a significant benefit for implementing networks that are intended to provide complete coverage to a large contiguous area. With four nonoverlapping channels, one can construct a network of individual access points according to a *frequency plan* that ensures that access points on the same channel are

separated by at least four cell radii; with seven channels to work with, a minimum spacing of 5.5 radii is achieved (Figure 2.18). Thus, interference between adjacent cells is minimized, and each cell is able to provide full capacity to its occupants. With the large number of independent channels now available, individual access points can be allocated more than one channel each to increase capacity in heavily used cells while still maintaining minimal interference.

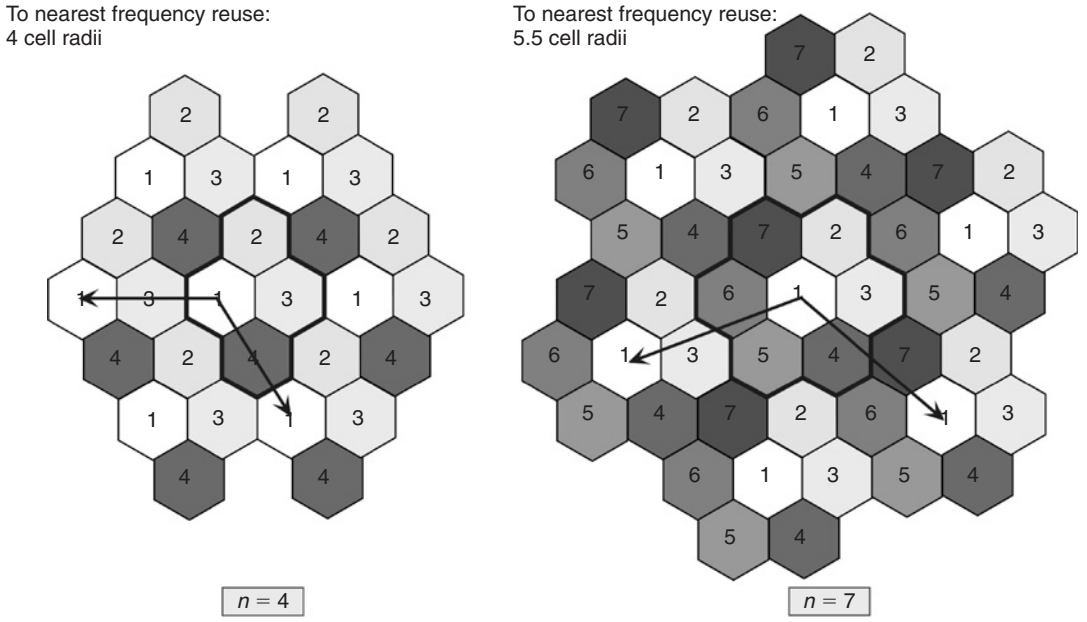


Figure 2.18: Examples of Cellular Coverage Plans for 4 and 7 Nonoverlapping Channels

Coding in 802.11a is somewhat similar to the PBCC option of the 802.11b PHY: after scrambling, incoming bits are processed by rate 1/2 convolutional encoder as shown in Figure 2.19. Like the PBCC encoder, the 802.11a code has a six-stage shift register, though the sequence of taps that define the output bits (formally, the generating polynomials of the code) have been changed. Codes with rates of 3/4 and 2/3 (fewer output bits per input bit and thus higher throughput of useful data) are created by puncturing the output of the rate 1/2 code, that is, some of the output bits are simply ignored. Such a procedure is roughly equivalent to intentionally introducing some bit errors in the data stream, though it is easier to correct

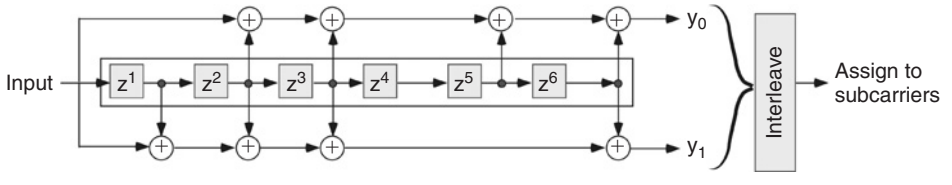


Figure 2.19: 802.11a Convolutional Encoder

for because the location of the missing bits is known; if done cleverly, it allows for higher throughput with little performance loss.

The use of the resulting output bits is significantly different from PBCC, because instead of single-carrier modulation, the 802.11a standard uses orthogonal frequency-division multiplexing (OFDM), a technique introduced in Chapter 1, section 1.3.2. A data symbol is composed of 48 subcarriers, which carry data, and 4 additional pilot subcarriers, which transmit a known pseudo-random sequence to assist the receiver in maintaining synchronization with the transmitter. The subcarriers are assigned to 64 frequency slots, separated by 312.5 kHz; the slots at the band edges and the slot in the center of the band (i.e., at the carrier frequency) are not used. If all the subcarriers were occupied, the signal would be approximately $64 \times 312.5\,\text{kHz} = 20\,\text{MHz}$ wide (63 spacings between subcarriers + half of the width of a subcarrier on each side); the actual (ideal, undistorted) signal is about 16.6 MHz wide, thus fitting nicely within the spectral mask in Figure 2.16.

Recall from Chapter 1 that to use OFDM, the time over which a symbol is integrated must be an integer number of cycles for all the subcarriers in order that orthogonality is maintained. The native symbol time of the 802.11a OFDM symbol is thus one cycle of the lowest frequency subcarrier, or 3:2 μsec. A cyclic prefix is appended to the symbol, as we described in Chapter 1, to allow for a guard interval that eliminates the effects of multipath. The cyclic prefix is 0.8 μsec long (about 20% of the total symbol), so that the total symbol length is 4 μsec. Therefore, the OFDM symbol rate is 250 Ksps. Each symbol contains 48 active subcarriers, so this is equivalent to sending 12 million single-carrier symbols per second.

To allow for varying conditions, many different combinations of code rate and subcarrier modulation are allowed. These are shown in Table 2.2. Recall from Chapter 1 that BPSK transports one bit per (subcarrier) symbol, QPSK two bits, and so on up to 64 quadrature-amplitude-modulation (QAM), which carriers 6 bits per symbol. The product of the bits per symbol of the modulation and the rate of the convolutional code gives the number of bits each

Table 2.2: 802.11a Modulations and Code Rates

Modulation	Code Rate	Data Rate (Mbps)
BPSK	1/2	6
BPSK	3/4	9
QPSK	1/2	12
QPSK	3/4	18
16QAM	1/2	24
16QAM	3/4	36
64QAM	2/3	48
64QAM	3/4	54

subcarrier contributes to each OFDM symbol; multiplication by 48 provides the number of bits transported per OFDM symbol.

Note that to avoid bursts of bit error from disruption of a few neighboring subcarriers, the output bits from the code are interleaved—that is, distributed in a pseudo-random fashion over the various subcarriers.

Several views of a simulated OFDM frame are shown in Figure 2.20. The frame consists of a preamble and simulated (random) data. The spectrum of the output signal is about 16-MHz wide, in good agreement with the simple estimate given above. Recall that a signal in time can be described by its in-phase and quadrature components, I and Q. The I and Q amplitudes for the frame are shown at the lower left. The frame begins with a set of simplified synchronization characters, which use only 12 subcarriers, and are readily visible at the left of the image. The remainder of the preamble and data frame appear to vary wildly in amplitude. This apparently random variation in the signal is also shown in the close-up of I and Q amplitudes over the time corresponding to a single OFDM symbol shown at the top right of the figure.

At the bottom right we show the *combined cumulative distribution function* for the signal power, averaged over the packet length, for a packet at 54 and 6 Mbps. The combined cumulative distribution function is a graphic display of the frequency with which the instantaneous power in the signal exceeds the average by a given value. The 54-Mbps packet instantaneous power is greater than 9 dB above the average around 0.1% of the time. The 6-Mbps packet displays an 8-dB enhancement in power with this frequency of occurrence. Although that may not seem like much, recall that the spectral mask (Figure 2.16) requires that the transmitted signal 20 MHz away from the carrier must be reduced by more than 28 dB. As we learn in Chapter 3, a high-power signal may become distorted, resulting in spectral components far from the carrier. If such components were comparable in size with the main signal and occurred only 1% of the time, their intensity would be 20 dB below the main signal. Infrequent power peaks contribute significantly to the output spectrum when distortion is present. This problem of a high ratio of peak power to average power represents one of the important disadvantages of using OFDM modulations, because it forces transmitters to reduce their average output power and requires higher precision and linearity from receivers than is the case for simpler modulations.

Recall that the OFDM symbol is constructed using an inverse fast Fourier transform (FFT), with the reverse operation being performed on the received signal to recover the subcarriers. In the case of 802.11a, a 64-entry FFT involving some hundreds of arithmetic operations must be performed at fairly high precision every 4 μsec, followed by a trellis decode of the resulting 48 data symbols. The rapid progress of digital integrated circuit scaling has permitted such a relatively complex computational system to be implemented inexpensively in standard silicon complementary metal-oxide semiconductor (CMOS) circuitry, a remarkable achievement.

2.3.7 802.11g PHY

The great disadvantage of the 802.11a PHY is its incompatibility with the older classic and 802.11b installed base. To overcome this obstacle while preserving the attractive high peak data rates of the 802.11a standard, task group "G" provided two enhancements to the Wi-Fi PHY. The first variant is essentially the transfer of the 802.11a OFDM physical layer *in toto* to the ISM band. This trick is possible because the bandwidth of the 802.11a symbols (Figure 2.20) is about 16 MHz, just as the classic symbols were.

The problem presented by this radical reinvention is that older systems have no ability to receive and interpret the complex OFDM symbols and simply see them as noise. Because the CSMA/CA MAC layer is heavily dependent on all stations being able to hear (at least) the access point, invisible preambles are a serious obstacle: receiving stations cannot set their NAV to reserve the medium (Figure 2.6) if they can't read the packet preambles. In order for "g" stations to coexist with older "b" stations, several options are possible. First, in the presence of mixed traffic, "g" stations can use the RTS/CTS approach to reserve the medium using 802.11b packets and then

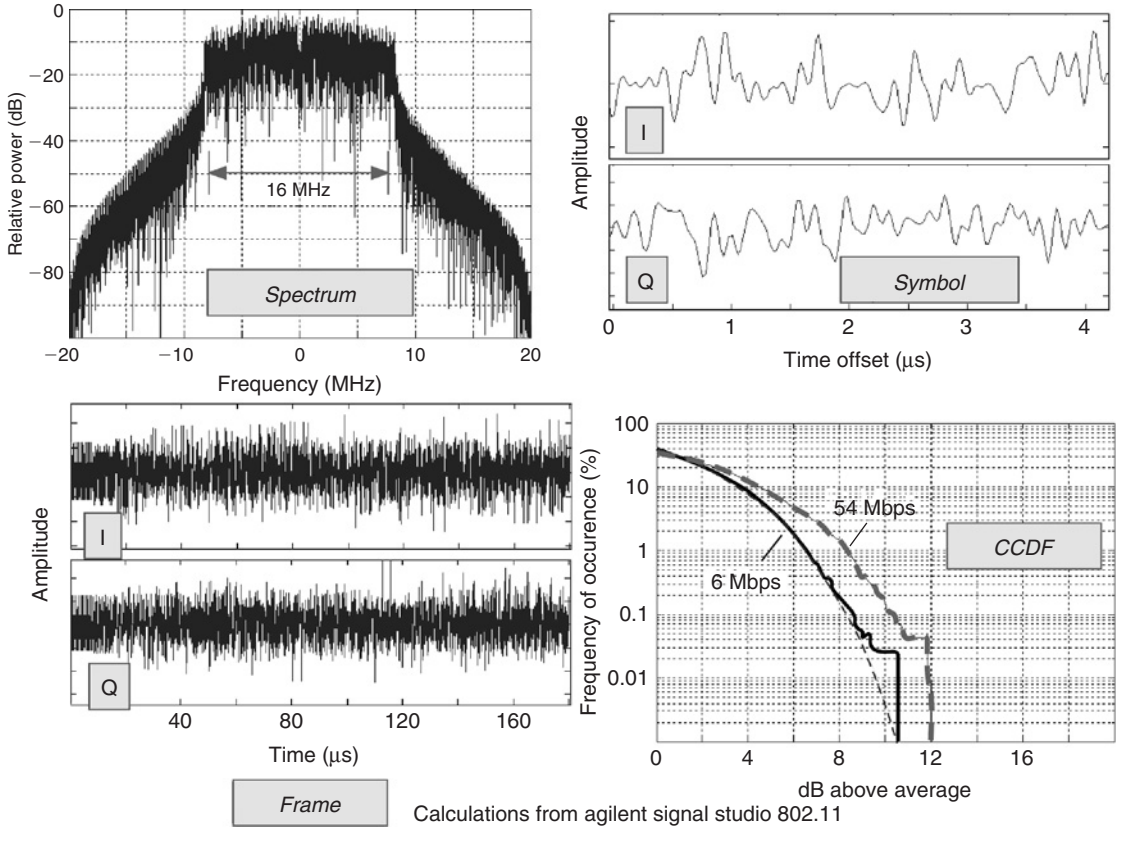


Figure 2.20: Several Views of an 802.11 OFDM Frame

use the reserved time to send faster 802.11g packets. This approach is simple but adds overhead to the "g" packet exchanges. Alternately, a station can send a CTS packet with itself as the destination: this approach, known as a CTS-to-self, is obviously more appropriate for an access point, which expects that every associated station can hear it, than a client.

A "g" station can also use a mixed frame format, in which the preamble is sent in a conventional 802.11b form and then the appended data are transferred as OFDM. In this case there is additional overhead because of the low data rate of the older preamble but no additional packet exchange. The final chip of the preamble is used as the phase reference for the first OFDM symbol.

Finally, "g" also defines a PHY using an extension of the PBCC coding system; in the highest-rate variant, 8PSK is substituted for QPSK modulation. It is perhaps too early in the history of 802.11g to draw confident conclusions about its evolution, but in view of the limited success of PBCC in 802.11b deployment, and successful if limited deployment of OFDM-based 802.11a chipsets by major manufacturers, it seems plausible that the OFDM variant will dominate 802.11g devices.

Trade-offs are necessarily encountered in the enhancement of the PHY. Higher data rates use higher modulation states of the subcarriers and thus require better (S/N) than the old QPSK chips. This implies that higher rate communications will have a shorter range than the lower rates. The use of OFDM signals gives rise to higher peak-to average power ratios, allowing less transmission power from the same amplifiers to avoid distortion. The benefit of coexistence is achieved at the cost of reduced actual data rates in the presence of legacy stations due to the overhead of any of the coexistence protocols discussed above. These problems are fairly minor bumps in the road, however: it seems likely that 802.11g products will dominate the WLAN marketplace until crowding in the ISM band forces wider adoption of 802.11a/dualband devices.

2.3.8 802.11 (In)Security

Most participants in wireless networking, or indeed networking of any kind, are at least vaguely aware that 802.11 presents security problems. One can hardly have a discussion of 802.11 without dealing with encryption and security; however, as whole books have already been written on this subject, the treatment here will be quite cursory.

We must first put the problem in its wider context. Most enterprise LANs have minimal internal authentication and no security, because it is presumed that physical security of the Ethernet ports substitutes for systemic provisions of this nature (though those hardy souls charged with administering networks at universities may find such presumptions laughable at best). Access to the network from the outside, however, is typically limited by a firewall that controls the type and destination of packets allowed to enter and exit the local network. Remote users who have a need to use resources from an enterprise network may use a *virtual*

private network (VPN), which is simply a logical link formed between two clients, one at the remote site and the other at the enterprise network, with all traffic between them being encrypted so as to defeat all but the most determined eavesdropper. Traffic on the Internet is similarly completely open to interception at every intermediate router; however, reasonably secure communications over the Internet can be managed with the aid of the Secure Sockets Layer (SSL), another encrypted link between the user of a web site and the site server.

It was apparent to the developers of wireless data devices that the wireless medium is more subject to interception and eavesdropping than the wired medium, though they were also aware of the many vulnerabilities in existing wired networks. They therefore attempted to produce a security mechanism that would emulate the modest privacy level obtained with physically controlled Ethernet wired ports in a fashion that was simple to implement and would not incur so much overhead as to reduce the WLAN data rate by an objectionable amount.

The resulting *Wired Equivalent Privacy* (WEP) security system was made optional in the classic standard but was widely implemented and extended. The system is a *symmetric key* system, in which the same key is used to encrypt and decrypt the data, in contrast to asymmetric systems in which a public key is used to encrypt data and a separate private key is used to decrypt it. Public key systems are much more difficult to penetrate but very computation-intensive and are used for the secure exchange of symmetric keys but not for packet-by-packet encryption. The basic idea of the encryption approach is to use the key to create a pseudo-random binary string (the *cipher stream*) of the same length as the message to be encrypted. The data bits (the *plaintext*) can then be added modulo-2 (which is the same as a bit-by-bit exclusive-or [XOR]) to the cipher stream to create the encrypted data or *cipher text*, which when added again to the same cipher stream will recover the original data (plaintext). The general scheme is shown in Figure 2.21.

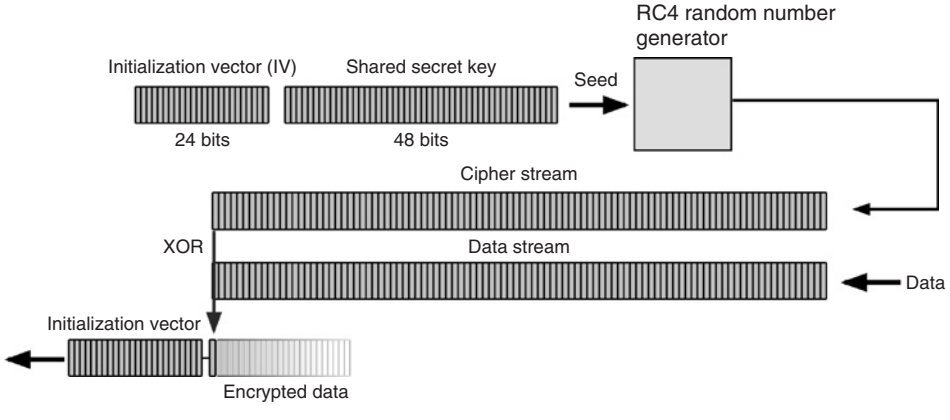


Figure 2.21: Simplified Schematic of WEP Encryption

A simple way to achieve such a result is to use the key as the seed value for a random number generator algorithm. The problem with this naive implementation is that every instance of the (pseudo) random sequence would then be identical; it wouldn't be very hard for an eavesdropper to reconstruct the cipher stream once a large number of packets, many containing known plaintext such as the contents of standard Internet packets, had been intercepted. Therefore, to avoid using the same cipher stream for every packet while also avoiding the complexity of exchanging a new key between the sender and receiver each time a packet needs to be transmitted, the WEP algorithm uses an *initialization vector* (IV): 24 additional bits that are prepended to the 48-bit secret key to form the seed for the random number generator. In this fashion, each new packet gets a new seed for the random number generator and thus a unique cipher stream.

In addition to security from interception, a secure communications system should provide some means of *authentication*: each side should have some confidence that they are in fact communicating with the machine (and ultimately the person) that is intended. The 802.11 standard appropriates WEP to support authentication by allowing an access point to challenge a potential client with a plaintext packet, which the client then returns encrypted with the shared WEP key. If the decrypted packet matches the original plaintext, the client is presumed to have knowledge of the shared key and thus be privileged to make use of the resources of the access point.

As has been so vividly shown with the public release of the remarkable means by which British and American researchers cracked the German Enigma code during World War II, attacks on cryptographic systems are rarely frontal but instead exploit weaknesses in the system and implementation. WEP has a number of weaknesses and limitations, some of which were quickly apparent and others more subtle and only revealed after significant work by outside researchers.

The first weakness is in the definition of the IV. The IV is necessarily sent in the clear because it is needed to decrypt the message content. As we noted above, to get decent security out of a random-number-generator scheme, it is necessary to avoid reusing the seed value, because this will generate a reused cipher stream. However, the IV space is only 24 bits. Furthermore, there is no requirement in the specification about how the IV is to be treated between successive packets; any IV value may be used and must be accepted by the receiving station. Many implementations simply incremented the IV value by 1 for each packet sent. In such a network, the IV is bound to be reused after $2^{24} = 16.8$ million packets. A heavily loaded network would deliver that much traffic in a couple of hours. Over the course of a week or two of eavesdropping, a huge number of packets with identical IVs and thus identical cipher streams could be collected and analyzed.

Once any cipher stream is obtained, it can be reused even in the absence of knowledge of the shared key, because any IV must be accepted. Any packets sent with the same IV can be decrypted (at least to the length of the known cipher stream); further, a packet of the same

length with the intruder's data can be encrypted, prepended with the same IV, and injected into the system.

The authentication system is another weakness. If an eavesdropper can hear the challenge and the response, the XOR of the challenge (plaintext) and corresponding response (cipher text) reveals the cipher stream. In this fashion one can accumulate cipher streams for packet injection or other attacks.

Obviously, it is even better to obtain the shared key and thus encrypt and decrypt freely than to extract a few cipher streams. Attacks to find the shared key can exploit the fact that the key is defined as a bit stream, but human users can't remember bit streams very well and tend to prefer words or streams of recognizable characters. Many 802.11 management systems allow the entry of a text password, which is processed (hashed) into a 48-bit shared key. Early hashing algorithms did not use the whole key space—that is, not all 48-bit numbers were possible results—so that the effective size of the keys to be searched was only 21 bits, which is about 2 million possibilities. (Some early cards even converted the ASCII values of letters directly into a key, resulting in a greatly reduced key space that is also easily guessed.) A very simple brute force recitation of all possible passwords is then possible, given that one can automate the recognition of a successful decryption. Because certain types of packets (standard handshakes for IPs) are present in almost any data stream, the latter is straightforward.

Because IVs can be reused, if there isn't enough traffic on a network to crack it, an attacker can always introduce more packets once he or she has acquired a few cipher streams, thus generating additional traffic in response.

Certain attacks are fixed or rendered much less effective by simply increasing the length of the shared key and the IV. Many vendors have implemented WEP variants with longer keys, though these are not specified in the standard and thus interoperability is questionable. For example, the larger space for keys, combined with improved hashing algorithms, renders the brute force password guessing attack ineffective on 128-bit WEP key systems.

The widely publicized "cracking" of the WEP algorithm by Fluhrer, Mantin, and Shamir (FMS) was based on a demonstration that the particular random number generator, RC4, used in WEP is not completely random. RC4 is a proprietary algorithm that is not publicly released or described in detail in the standard. It has been reverse engineered and shown to be based on fairly simple swaps of small segments of memory. Fluhrer and colleagues showed that certain IV values "leak" some information about the key in their cipher streams. To benefit from this knowledge, an attacker needs to know a few bytes of the plaintext, but because of the fixed nature of formats for many common packets, this is not difficult. The number of weak IVs needed to extract the shared key is modest, and the frequency of weak IVs increases for longer key streams, so that for a network selecting its IVs at random, the time required to crack, for example, a 104-bit key is only roughly twice that needed for a 40-bit key. Long keys provide

essentially no added security against the FMS attack. Shortly after Fluhrer et al. became available, Stubblefield, Ioannidis, and Rubin implemented the algorithm and showed that keys could be extracted after interception of on the order of 1,000,000 packets.

Finally, we should note that there is no provision whatsoever in the standard to allow a client to authenticate the access point to which it is associating, so that an attacker with a shared key can set up a spoof access point and intercept traffic intended for a legitimate network.

All the above attacks depend on another weakness in the standard: no provision was made for how the shared secret keys ought to be exchanged. In practice, this meant that in most systems the secret keys are manually entered by the user. Needless to say, users are not eager to change their keys every week, to say nothing of every day, and manual reentry of shared keys more than once an hour is impractical. Further, manual key exchange over a large community of users is an administrative nightmare. As a consequence, it is all too likely that WEP shared secret keys will remain fixed for weeks or months at a time, making the aforesaid attacks relatively easy to carry out.

Some of the problems noted above can be solved completely within the existing standards. Weak IVs can be avoided proactively, and some vendors have already implemented filters that ensure that weak IVs are never used by their stations. If all stations on a network avoid weak IVs, the FMS attack cannot be carried out. Weak hashing algorithms for short keys can be avoided by entering keys directly as hexadecimal numbers. In home networks with few users and relatively light traffic, manual key exchange on a weekly or even monthly basis will most likely raise the effort involved in a successful attack above any gains that can be realized.

The simple measures cited above can hardly be regarded as adequate for sensitive industrial or governmental data. In 2003, products conforming to a preliminary standard promulgated by the WECA industrial consortium, known as Wi-Fi Protected Access (WPA), became available. WPA is a partially backward-compatible enhancement of WEP. WPA uses the IEEE 802.1x standard as a framework for authentication of users and access points; within 802.1x various authentication algorithms with varying complexity and security can be used.

WPA uses a variant of the *Temporal Key Integrity Protocol* created by Cisco to improve security of the packet encryption process. An initial 128-bit encryption key, provided in a presumably secure fashion using the 802.1x authentication process, is XOR'd with the sending station's MAC address to provide a unique known key for each client station. This unique intermediate key is mixed with a 48-bit sequence number to create a per-packet key, which is then handed over to the WEP encryption engine as if it were composed of a 24-bit IV and 104-bit WEP key. The sequence number is required to increment on each packet, and any out-of-sequence packets are dropped, preventing IV-reuse attacks. The 48-bit sequence space means that a sequence number will not be reused on the order of 1000 years at today's data rates, so there are no repeated cipher streams to intercept. A sophisticated integrity checking

mechanism is also included to guard against an attacker injecting slight variations of valid transmitted packets.

WPA addresses all the currently known attacks on WEP, though total security also depends on proper selection and implementation of algorithms within the 802.1x authentication process. It is certainly secure enough for home networks and for most enterprise/industrial implementations. At the time of this writing (mid-2004), the 802.11i task group has approved an enhanced standard based on the Advanced Encryption Standard rather than the WEP RC4 algorithm, which will provide an adequate level of security for most uses.

Enterprises that have serious concerns about sensitive data can also implement end to end security through the use of VPNs and SSL web security. The advantage of this security approach is that protection against eavesdropping at any stage of the communications process, not just the wireless link, is obtained. However, VPNs are complex to set up and maintain and may not support roaming of the wireless device from one access point to another, particularly if the IP address of the mobile device undergoes a change due to the roaming process.

2.4 HiperLAN and HiperLAN 2

Some of the researchers who participated in defining the 802.11a standard were also active in similar European efforts: the resulting standards are known as HiperLAN. The HiperLAN 2 physical layer is very similar to that used in 802.11a: it is an OFDM system operating in the 5-GHz band, using an almost identical set of modulations and code rates to support nearly the same set of data rates. The physical layer is also compliant with ETSI requirements for dynamic frequency selection and power control.

However, the MAC layer is quite different. Instead of being based on the Ethernet standard, it is based on a totally different traffic approach, *asynchronous transfer mode* (ATM). ATM networking was developed by telephone service providers in the 1980s and 1990s in an attempt to provide a network that would support efficient transport of video, data, and voice, with control over the quality of service and traffic capacity assigned to various users. Native ATM is based on fixed 53-byte packets and virtual connections rather than the variable packets and global addressing used in IP networking over Ethernet. The HiperLAN MAC was constructed to provide a smooth interface to an ATM data network, by using fixed time slots assigned to stations by an access point acting as the central controller of the network.

ATM was originally envisioned as extending all the way from the WAN to the desktop but in practice has seen very little commercial implementation beyond the data networks of large telecommunications service providers. Similarly, no commercial products based on HiperLAN have achieved significant distribution, and the author is not aware of any products in current distribution based on HiperLAN 2. It seems likely, particularly in view of recent FCC decisions bringing U.S. regulations into compliance with European regulations, that

802.11a products will achieve sufficient economies of scale to prevent the wide distribution of HiperLAN-based products in most applications. Although HiperLAN provides superior quality-of-service controls to 802.11, many major vendors for the key quality of service-sensitive service, video delivery, have transferred their attention to the ultrawideband (UWB) standardization activities in the 802.15.3a, discussed in the next section.

2.5 From LANs to PANs

WLANs are an attempt to provide a wireless extension of a computer network service. WPANs are intended for a somewhat different purpose and differ from their WLAN cousins as a consequence. The basic purpose of a WPAN is to replace a cable, not necessarily to integrate into a network attached to that cable. WPANs seek to replace serial and parallel printer cables, universal serial bus connections, and simple analog cables for speakers, microphones, and headphones with a single wireless digital data connection. Personal areas vary from person to person and culture to culture but are usually much smaller than a building, so PANs do not need to have the same range expected of WLANs. The cables being replaced are generally very inexpensive, so even more than WLANs, WPAN products must be inexpensive to build and use. Many WPAN products are likely to be included in portable battery-powered devices and should use power sparingly.

In this section we discuss three differing WPAN standards activities, all (at least today) taking place under the aegis of the 802.15 working group of IEEE. As is often the case, with the passing of time the definition of a category is stretched: what we've defined as a PAN for replacing low-rate cables is certainly evolving to replace higher rate cables and may be becoming a path to the universal home data/entertainment network that many companies have already unsuccessfully sought to popularize.

2.5.1 Bluetooth: Skip the Danegeld, Keep the Dane

The first WPAN effort to achieve major visibility was the Bluetooth Special Interest Group, initiated in 1998 by Ericsson with support from Intel, IBM, Nokia, and Toshiba. The Bluetooth trademark is owned by Ericsson and licensed to users. Although the group clearly displayed more marketing savvy than all the IEEE working groups put together in its choice of a memorable moniker—the name refers to Harald "Bluetooth" Blätand, king of Denmark from about AD 940 to 985—the separation of this activity from the IEEE helped produce a standard that was not compatible in any way with 802.11 networks. Given the likelihood of collocated devices using the two standards, this is unfortunate. The Bluetooth standard is now also endorsed by the IEEE as 802.15.1.

To make life simple for users, the Bluetooth PAN was intended to operate with essentially no user configuration. Thus, Bluetooth networks are based on *ad hoc* discovery of neighboring devices and formation of networks. Simple modulation, low transmit power, low data rate,

and modest sensitivity requirements all contribute to a standard that can be inexpensively implemented with minimal power requirements. Frequency hopping was used to allow multiple independent collocated networks. The network is based on synchronized time slots allocated by the master device, so that resources could be provided on a regular deterministic basis for voice and other time-sensitive traffic.

The architecture of Bluetooth networks is based on the *piconet* (Figure 2.22). In terminology also perhaps harkening back to the time of King Blätand, a piconet consists of one device that acts as the *master* and a number of other devices that are *slaves*. Each piconet is identified by the FH pattern it uses. The hopping clock is set by the master device, and slaves must remain synchronized to the master to communicate. Provisions are made to permit devices both to discover nearby piconets and spontaneously form their own.

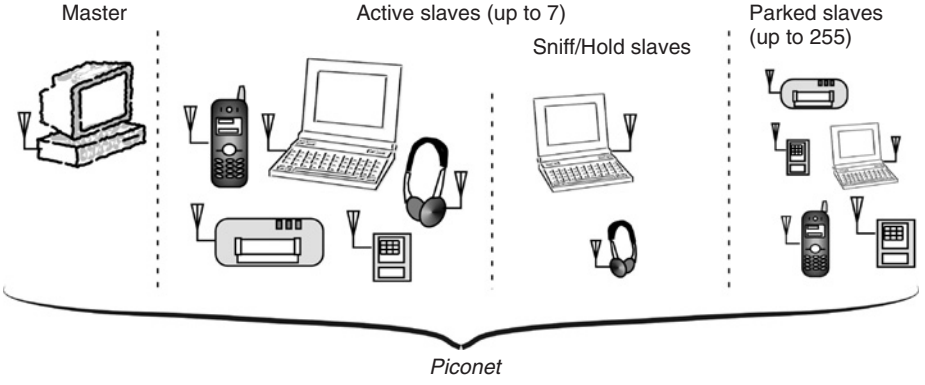


Figure 2.22: Bluetooth Piconet

Only seven slave devices can be active at any given time. Some of the active devices can be placed in a power-saving *sniff* state, in which they only listen for their packets occasionally, or *hold* states, in which they do not actively participate in the piconet for some period of time before checking back in. Additional members of the piconet may be *parked*: in the parked state, slaves maintain synchronization with the master clock by listening to beacons periodically but do not otherwise participate until instructed to become active.

Piconets can physically overlap with one another with minimal interference, as they use distinct hopping schemes. A slave in one piconet could be a master in another. By having some devices participate in multiple piconets, they can be linked together to form *scatternets*. Devices within a piconet can exchange information on their capabilities, so that the user may use printers, cameras, and other peripherals without the need for manual configuration. Devices use *inquiry* and *paging* modes to discover their neighbors and form new piconets.

The Bluetooth/802.15.1 PHY layer operates in the 2.4-GHz ISM band, which is available at least partially throughout most of the world. The band in the United States is divided into 79

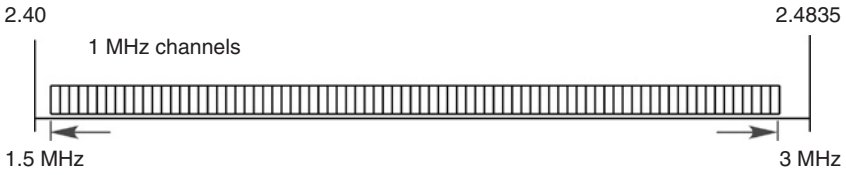


Figure 2.23: Bluetooth Channelization

1-MHz channels (Figure 2.23), with modest guard bands at each end to minimize out-of-band emissions.

During normal operation, all members of the piconet hop from one frequency to another 1600 times per second. The dwell time of 625 μsec provides for 625 bit times between hops. When a device is in inquiry or paging mode, it hops twice as fast to reduce the time needed to find other devices.

The signal is modulated using *Gaussian minimum-shift keying*, GMSK (also known as Gaussian frequency-shift keying). GMSK can be regarded as either a frequency modulation or phase modulation technique. In each bit time, frequency is either kept constant or changed so as to shift the phase of the signal by π by the end of the period *T* (Figure 2.24). Because the amplitude remains constant during this process, the peak-to-average ratio of a GMSK signal is essentially 1, so such signals place modest requirements on transmitter and receiver linearity. GMSK is similar to BPSK, but BPSK systems have no such constraint on the trajectory taken between phase points and thus may have large amplitude variations during the transition from one symbol to another. Each GMSK symbol carries only a single bit, and thus 1 Msps = 1 Mbps of data transfer capacity.

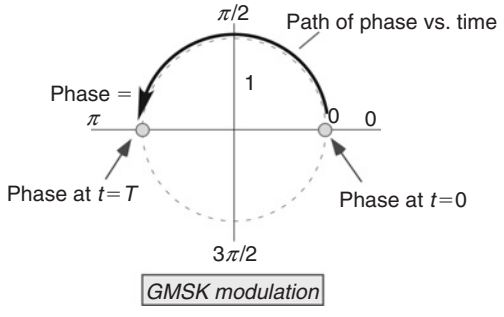


Figure 2.24: Gaussian Minimum-Shift Keying

Bluetooth radios come in three classes: class 1 can transmit up to 20 dBm (100 mW), class 2 is limited to 4 dBm (2.5 mW), and class 3 devices, the most common, operate at 1 mW (0 dBm). Bluetooth receivers are required to achieve a bit error rate of 0.1% for a received signal at −70 dBm: that's a raw bit error rate of about 60% for 600-bit packets. These specifications

are rather less demanding than the corresponding 802.11 requirements: recall that an 802.11 receiver is required to achieve a sensitivity of −80 dBm at 2 Mbps and typical transmit power is 30–100 mW. A link between two compliant class 3 devices has 30 dB less path loss available than a link between two 802.11 devices. The receiver noise and transmit power requirements are correspondingly undemanding: Bluetooth devices must be inexpensive to build.

An example of a link budget calculation is shown in Figure 2.25. We've assumed a transmit power of 1 dBm and modest antenna directivity (1 dB relative to an isotropic antenna [dBi]) and efficiency (80%) as the antennas are constrained to be quite small to fit into portable/handheld devices. At 3 m the free-space path loss is about 51 dB; we assumed that if a person is in the way of the direct path, an additional 10 dB of obstructed loss is encountered. If we assume a modest 4 dB (S/N) is required by the GMSK modulation, we find that we have room for 39 dB of excess noise from the receiver while still providing 10 dB of link margin in the presence of an obstruction. This is a huge noise figure; practical Bluetooth devices can do much better. The Bluetooth specifications are appropriate for inexpensive devices meant to communicate at moderate rates over ranges of a few meters.

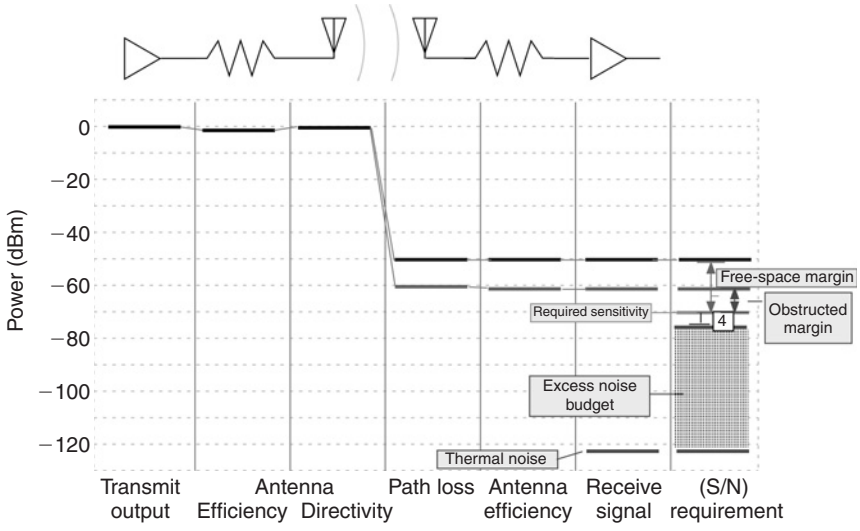


Figure 2.25: Bluetooth Example Link Budget: 10 m, 0 dBm TX Power

A few words about the security of a Bluetooth link seem appropriate. Bluetooth links have two fundamental security advantages over 802.11 irrespective of any encryption technology. The first is a physical limit: a Bluetooth class 3 transmitter at 0 dBm is harder to intercept from afar than an 802.11 access point at 20-dBm transmit power. The second is related to the usage of the devices. Most 802.11 devices are connected to a local network of some kind, which provides many resources of interest to an attacker: servers containing possibly valuable data and a likely

connection to the global Internet. Many Bluetooth devices may only provide access to a single data stream or a few neighbor devices. The attacker who intercepts data sent to a printer from a few meters away might just as well walk over to the printer and look at what's coming out. Finally, any given Bluetooth link is likely to be used less intensively and provides a lower peak data rate than an 802.11 link, so an attacker has a harder time gathering data to analyze.

That being said, Bluetooth provides security services that are in some ways superior to those in 802.11. Authentication and encryption are both provided and use distinct keys. The cipher stream generation uses the E0 sequence generator, which is at the present regarded as harder to attack than RC4. However, just like in 802.11, weaknesses are present due to the usage models. Keys for devices with a user interface are based on manually entered identifier (PIN) codes. Users won't remember long complex codes and don't do a very good job of choosing the codes they use at random, so links are likely to be vulnerable to brute force key-guessing attacks. Devices without a user interface have fixed PIN codes, obviously more vulnerable to attack. Devices that wish to communicate securely must exchange keys in a pairing operation; if an eavesdropper overhears this operation, they are in an improved position to attack the data.

Although there hasn't been as much publicity about cracking Bluetooth, as was the case with 802.11, it is prudent to use higher layer security (VPN or SSL, etc.) if large amounts of sensitive data are to be transferred over a Bluetooth link.

2.5.2 Enhanced PANs: 802.15.3

The 802.15.3 task group was created to enhance the capabilities of 802.15.1 (Bluetooth) while preserving the architecture, voice support, low cost of implementation, and range of the original. Data rates as high as 55 Mbps were targeted to allow applications such as streaming audio and video, printing high-resolution images, and transmitting presentations to a digital projector. Because the work was now under the aegis of the IEEE, the group sought to create a PHY with some compatibility with existing 802.11 devices; thus a symbol rate of 11 Msps was chosen, and the channels were to be compatible with the 5-MHz 802.11 channels. However, the group sought to use a slightly reduced bandwidth of 15 MHz, so that more nonoverlapping channels could be provided in the same area than 802.11 classic or b/g allow.

In July 2003, the 802.15.3 task group approved a new PHY layer standard that provides higher data rates while otherwise essentially using the Bluetooth MAC and protocol stack. The general approach is to use more sophisticated coding than was available a decade ago to allow more data to be transferred in less bandwidth. As with 802.11a and g, higher modulations are used to send more data per symbol; 32QAM uses the 16QAM constellation with the addition of partial rows and columns.

A novel coding approach, *trellis-coded modulation* (TCM), is used to achieve good performance at modest (S/N). In the conventional coding schemes we examined so far, each set of bits determines a modulation symbol (for example, a dibit determines a QPSK point) and the

received voltage is first converted back to a digital value (an estimate of what the dibit was) and then used to figure out what the transmitted code word probably was. Coding is a separate and distinct operation from modulation. TCM mixes coding and modulation operations by dividing the constellation into subsets, within each of which the distance between signal points is large. The choice of subset is then made using a convolutional code on some of the input bits, whereas the choice of points within the subset is made using the remainder of the input bits uncoded. The subset points are widely separated and noise resistant, so uncoded selection works fine. The subsets are more difficult to distinguish from one another, but this choice is protected by the code, making it more robust to errors. The net result is typically around 3 dB of improved noise margin in the same bandwidth versus a conventional modulation.

Table 2.3: 802.15.3 Data Rates

Modulation	Coding	Code Rate	Data Rate (Mbps)	Sensitivity (dBm)
QPSK	TCM	1/2	11	−82
DQPSK	None	–	22	−75
16QAM	TCM	3/4	33	−74
32QAM	TCM	4/5	44	−71
64QAM	TCM	2/3	55	−68

The benefits of this approach are revealed in the spectral emission mask (Figure 2.26). The 802.15.3 signal is able to deliver the same data rate as 802.11g but fits completely within a 15-MHz window. The practical consequence is that one can fit five nonoverlapping channels into the U.S. ISM band, providing more flexibility in channel assignment and reducing interference. Some exposure to increased intersymbol interference and sensitivity to multipath may result from the narrower bandwidth, but because 802.15.3 is a PAN technology not intended to be used at long ranges, little performance limitation should result.

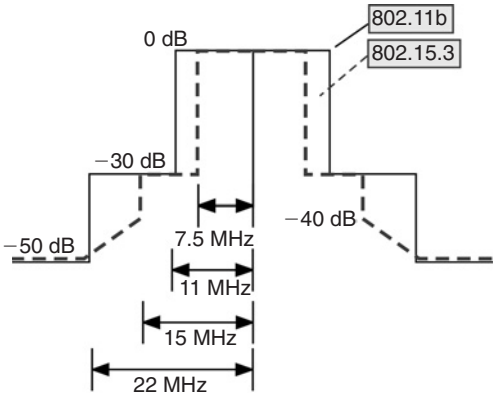


Figure 2.26: Comparison of 802.11b and 802.15.3 Emission Spectral Mask

2.5.3. UWB PANs: A Progress Report

Although 55 Mbps may seem like a lot of bits, it is a marginal capacity for wireless transport of high-definition television signals (a technology that may finally reach mainstream deployment in this decade in conjunction with large-screen displays). Task group 802.15.3a was charged with achieving even higher data rates, at minimum 100 Mbps with a target of 400 Mbps, to support versatile short-range multimedia file transfer for homes and business applications.

To achieve such ambitious goals, the task group turned to a new approach, based again on recent FCC actions. In 2002 the FCC allowed a novel use of spectrum, ultrawideband transmission. UWB radios are allowed to transmit right on top of spectrum licensed for other uses (Figure 2.27). However, UWB radios operate under several restrictions designed to minimize interference with legacy users of the spectrum. First, the absolute power emitted at any band is restricted to less than −40 dBm, which is comparable with the emissions allowed by nonintentional emitters like computers or personal digital assistants. The restrictions are more stringent in the 1- to 3-GHz region to protect cellular telephony and global position satellite navigation. Finally, like the original requirements on the ISM band, UWB users are required to spread their signals over at least 500 MHz within the allowed band.

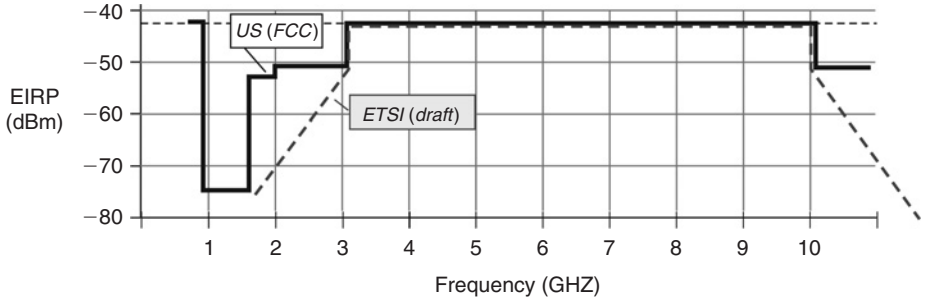


Figure 2.27: U.S. FCC and Proposed ETSI Limits for Ultrawideband Emitters (EIRP Is Equivalent Isotropic Radiated Power; see Chapter 5 for a Discussion of this Concept)

The wide bandwidth available for a UWB signal means that the peak data rate can be very high indeed. However, the limited total power implies that the range of transmission is not very large. As with any transmission method, UWB range can be extended at the cost of reduced data rate by using more complex coding methods to allow a tiny signal to be extracted from the noise.

The resulting trade-off between range and rate is shown in Figure 2.28. Here a full band transmission is one that exploits the whole 7-GHz bandwidth available under the FCC specs, whereas a subband transmission uses only the minimum 500-MHz bandwidth. It is clear that a

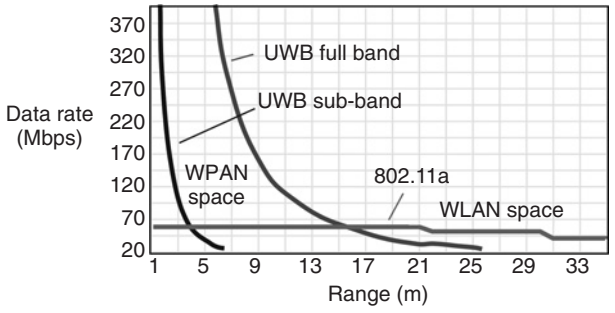


Figure 2.28: UWB Range vs. Data Rate, Shown for Both Full Band (7 GHz) and Subband (500 MHz) Approaches (After Gandolfo, Wireless Systems Design Conference, 2003)

subband approach can deliver 100 Mbps at a few meters, appropriate to a PAN application. As the demanded range increases, even the full UWB band cannot provide enough coding gain to substitute for transmit power, and the performance of the UWB transmitter falls below that of a conventional 802.11a system. UWB systems do not replace conventional WLAN systems but have ideal characteristics for WPAN applications.

At the time of this writing, two distinct proposals are under consideration by the 802.15.3a task group as a WPAN standard. We first examine the direct-sequence CDMA proposal, based on 802.15–03 334/r3 due to Welborn et al., with some updates from /0137 . . . r0. This proposal divides the available bandwidth into two chunks, a low band and a high band, avoiding the 5-GHz UNII region to minimize interference with existing and projected unlicensed products (Figure 2.29). The radio can operate using either the low band only, the high band only, or both bands, achieving data rates of 450, 900, or 1350 Mbps, respectively. Alternatively, the dual-band operating mode can be used to support *full-duplex* communications, in which the radio can transmit in one band and simultaneously receive in the other.

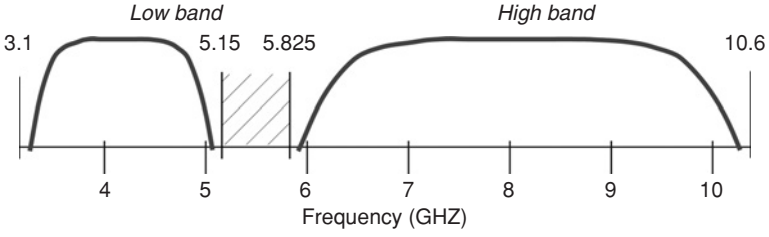


Figure 2.29: Pulse UWB Band Structure

The transmissions use Hermitian pulses, the nomenclature referring to the pulse shape and consequent spectrum (Figure 2.30). The high band pulse is shown; a longer pulse is used for the low band.

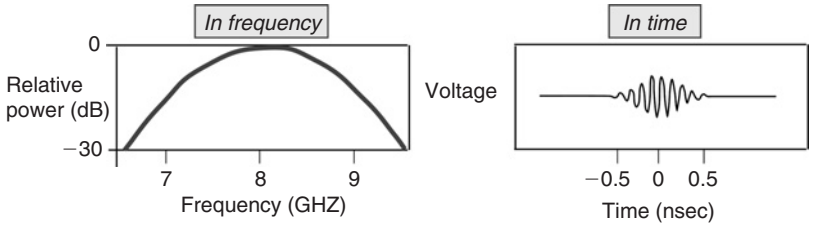


Figure 2.30: High-Band Modified Hermitian Pulse

The modulation techniques used are either BPSK or a variant of the biphase keying described in Chapter 1: *M-ary biorthogonal keying* (MBOK). In this scheme, the allowed symbols are specific sequences of positive and negative pulses; the sequences can be arranged in a hierarchy of mutually orthogonal symbols, in the sense that if they are multiplied together timewise and the sum taken (i.e., if they are correlated), the sum is 0 for distinct sequences. An example for $M = 4$ is shown in Figure 2.31. Each possible symbol is composed of four pulses, shown in cartoon form on the left. In the shorthand summary on the right, an inverted pulse is denoted "–" and a noninverted pulse is "+". Clearly, the first two symbols (00 and 01) are inversions of each other, as are the last two (10 and 11). Either symbol in the first group is orthogonal to either symbol in the second, for example, $00.10 = -1 + 1 - 1 + 1 = 0$, where we have assumed the product of a positive pulse and an inverted pulse to be 1.

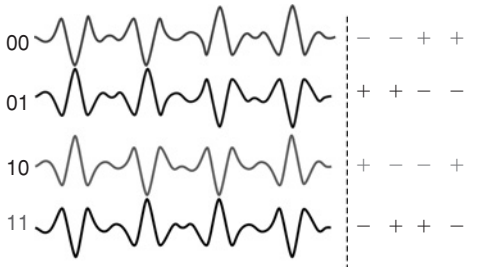


Figure 2.31: 4-Ary Biorthogonal Symbols

The proposed radio architecture uses the MBOK symbols as chips in a CDMA system. Each data bit is multiplied by a 24- or 32-chip code. This scheme is similar to the Barker code used to multiply data bits in the 802.11 classic PHY, but unlike 802.11, in this case multiple codes are provided to enable simultaneous noninterfering use of the bands. Four independent 24-chip codes are provided, allowing four independent users of each band or eight total collocated piconets. A *rake receiver* can be used to separately add the contributions of sequences of chips arising from paths with different delays; a *decision feedback equalizer* is assigned to deal with interference in the data symbols resulting from multipath. Both convolution and block code options are provided; the codes may be concatenated for improved performance at the cost of computational complexity.

Link performance at a transmit-receive distance of 4 m and data rate of 200 Mbps is summarized in Table 2.4. Note that the transmit power, −10 dBm, is 10 dB less than the lowest Bluetooth output and about 25 dB less than a typical 802.11 transmitter. The required bit energy over noise, E_b/N_o, is comparable with what we have encountered in conventional BPSK or QPSK modulation. Allowing for a couple of decibels of "implementation loss" (antenna limitations, cable losses, etc.) and a noise figure of 6.6 dB, the link should still achieve a respectable −75 dBm sensitivity, allowing some margin at a few meters range.

Table 2.4: Link Performance, DS-CDMA UWB at 200 Mbps

Parameter	Value
Data rate	200 Mbps
TX power	−10 dBm
Path loss	56 dB (4 m)
RX power	−66 dBm
Noise/bit	−91 dBm
RX noise figure	6.6 dB
Total noise	−84.4 dBm
Required E_b/N_o	6.8 dB
Implementation loss	2.5 dB
Link margin	8.7 dB
RX sensitivity	−75 dBm

The second current proposal is based on multiband OFDM, as described in IEEE 802.15–03/267r2, r6 Batra et al., with some updates from 802.15–04/0122r4 of March 2004. In this case, the available UWB bandwidth is partitioned into 528-MHz bands (recall that 500 MHz is the smallest bandwidth allowed by the FCC for an UWB application), as shown in Figure 2.32. The bands are assembled into five groups. Group 1 devices are expected to be introduced first, with more advanced group 2–5 devices providing improved performance. All devices are required to support group 1 bands; the higher bands are optional.

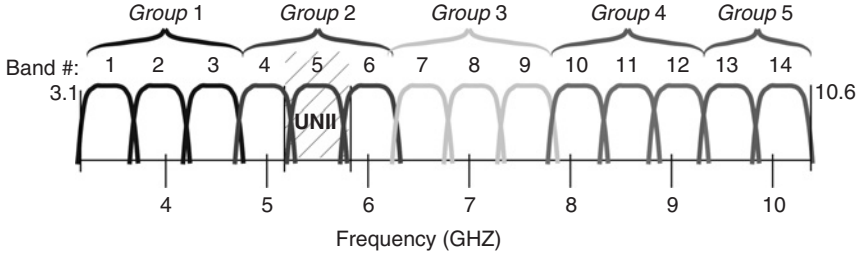


Figure 2.32: Band Definitions for OFDM UWB Proposal

The implementation requires a 128-point inverse FFT to create subcarriers spanning slightly more than 500 MHz of bandwidth. To reduce the computational demands, QPSK modulation on each subcarrier is used, instead of the higher QAM modulations used in 802.11 a/g. Furthermore, for data rates less than 80 Mbps, the subcarrier amplitudes are chosen to be conjugate symmetric around the carrier. That is, the nth subcarrier at positive frequency (relative to the carrier) is the complex conjugate of the nth subcarrier at negative frequency. A signal composed of complex conjugate frequencies produces a pure real voltage (this is just a fancy way of reminding us that $e^{i\omega t} + e^{-i\omega t} = 2\cos(\omega t)$). This means that the transmitter and receiver don't need separate branches to keep track of the in-phase and quadrature (I and Q) components of the signal and can be simplified.

Successive OFDM symbols are sent on different bands within a band group. Within each band group 1–4 there are four hopping patterns, each six symbols long, allowing for four collocated piconets. (Group 5 supports only two hopping patterns.) Thus, a total of 18 collocated piconets is possible.

The implementation parameters for the 200 Mbps are summarized in Table 2.5. A convolutional code with rate 5/8 (8 output bits for 5 input bits) is used. Of the 128 subcarriers, 22 are used as pilot tones and guard tones and 6 are not used at all. The resulting symbol is about 1/4 of a microsecond long. The cyclic prefix is no longer included, because this prefix adds a periodicity to the signal, giving rise to spectral lines which reduce the amount of power that can be transmitted. Instead, an equivalent amount of zero padding is added in frequency space (*zero-padded OFDM*). Each data symbol is spread by being sent twice, in separate frequency subbands. A brief guard interval is provided between symbols to allow the radio to change subbands. The use of separate subbands allows multiple collocated piconets.

Table 2.5: Wideband OFDM Parameters, 200 Mbps Rate

Parameter	Value
Data rate	200 Mbps
Modulation constellation	OFDM/QPSK
FFT size	128 tones
Coding rate	$R = 5/8$
Spreading rate	2
Pilot/guard tones	22
Data tones	100
Information length	242.4 nsec
Padded prefix	60.6 nsec
Guard interval	9.5 nsec
Symbol length	312.5 ns
Channel bit rate	640 Mbps

In Table 2.6 we compare link performance for the two proposals for a data rate of 200 Mbps at 4 m. OFDM has a slight advantage in noise tolerance, giving it modestly improved sensitivity. However, as we noted in connection with the 802.11b PHY, small distinctions in radio performance are likely to be less important than cost and ease of use in determining commercial viability.

Table 2.6: Link Performance, 200 Mbps

Parameter	Wideband OFDM	DS-CDMA
Information data rate	200 Mbps	200 Mbps
Average TX power	−10 dBm	−10 dBm
Total path loss	56 dB (4 m)	56 dB (4 m)
Average RX power	−66 dBm	−66 dBm
Noise power per bit	−91.0 dBm	−91 dBm
Complementary metal-oxide semiconductor RX noise figure	6.6 dB	66 dB
Total noise power	−84.4 dBm	−84.4 dBm
Required E_b/N_o	4.7 dB	6.8 dB
Implementation loss	2.5 dB	2.5 dB
Link margin	10.7 dB	8.7 dB
RX sensitivity level	−77.2 dBm	−75 dBm

Recent modifications (March 2004) to the direct-sequence proposal have changed the MBOK signaling and proposed a common signaling mechanism between the multiband OFDM and direct sequence approaches. The selection process appears to be favoring the multiband OFDM proposal, but it has not yet been approved. At the time of this writing, it is not known whether either of these proposals will be approved by the IEEE as standards or whether the resulting standard will enjoy any success in the marketplace. However, even if never deployed, these proposals provide an indication of the likely approaches for future high-rate short-range radio technologies.

2.6 Capsule Summary: Chapter 2

WLANs are constructed to extend LANs and share some of the properties of their wired forebears. However, the wireless transition places requirements on any WLAN protocol for connection management, medium allocation, error detection and correction, and additional link security. IEEE 802.11b WLANs provided a good combination of range and data rate at low cost and easy integration with popular Ethernet-based wired LANs; the successful WECA consortium also provided assurance of interoperability between vendors. These factors made 802.11b Wi-Fi networks popular despite limited support for time-sensitive traffic and porous link security. Enhancements of the 802.11b standard to higher data rates (802.11g) and

different bands (802.11a) provide an evolution path to support more users and applications; WPA and eventually 802.11i enable a more robust encryption and authentication environment for protecting sensitive information.

Contemporaneous with the evolution of WLANs, WPANs have been developed to replace local cables. For historical and technical reasons, a distinct PAN architecture has evolved, using master/slave self-organizing piconets. WPAN radios are targeted for short-range communications and require modest transmit power and limited receiver sensitivity. Support for time-sensitive traffic is provided in the design. WPANs are evolving toward higher data rates and in a possible UWB implementation may provide the basis for the long-envisioned convergence of the various home entertainment/information media into a single short-range wireless network.

2.7 Further Reading

802.11 Networks

802.11 Networks: The Definitive Guide, Matthew Gast: *The best overall introduction. Goes far beyond what we have provided here in examining framing and network management.*

Wireless Local Area Networks, Benny Bing (ed.), Wiley-Interscience, 2002: *Because this is a collection of chapters by different authors from different organizations, the quality of the chapters varies considerably, but the book covers a lot of ground. Chapter 2 (Heegard et al.) provides a very nice discussion of the performance of 802.11b codes.*

How Secure is Your Wireless Network?, Lee Barken, Prentice–Hall, 2004: *Provides a somewhat more detailed discussion of the flaws in WEP and a practical introduction to the use of WPA, EAP, and VPN setup.*

WEP Attacks

"Weaknesses in the Key Scheduling Algorithm of RC4," S. Fluhrer, I. Mantin, and A. Shamir, Selected Areas in Cryptography conference, Toronto, Canada, 16–17 August 2001

"Using the Fluhrer, Mantin and Shamir Attack to Break a WEP," A. Stubblefield, J. Ioannidis, and A. Rubin, Network and Distributed Security Symposium Conference, 2002

Bluetooth

Bluetooth Revealed, Brent Miller and Chatschik Bisdikian, Prentice–Hall, 2001: *Nice description of how the Bluetooth protocols work to provide discovery, association, and network traffic management.*

Trellis-Coded Modulations

Digital Modulations and Coding, Stephen Wilson: *See section 6.6 for a more detailed examination of how TCM works.*

Standards

http://www.ieee802.org/ is the home page for all the 802 standards activities. At the time of this writing, 802 standards are available for free download as Adobe Acrobat PDF files 6 months after they are approved. Many working documents are also available on the respective task group web pages. Draft standards that are not yet available for free download can be purchased from the IEEE store (but beware of awkward digital-rights-management limitations on these files, at least as of 4Q 2003).

http://www.etsi.org/SERVICES_PRODUCTS/FREESTANDARD/HOME.HTM provides access to ETSI standards as PDF downloads. The HiperLAN standards are described in a number of ETSI documents; three useful ones to start with are as follows:

ETSI TS 101 475 v 1.3.1 (2001): "Broadband Radio Access Networks . . . HiperLAN 2 PHYSICAL LAYER"

TR 101 031 v 1.1.1 (1997) "Radio Equipment and Systems . . . (HiperLAN) . . . " and

ETSI EN 301 893 v 1.2.3 (2003) "Broadband Radio Access Networks . . . 5 GHz High Performance RLAN"

http://www.bluetooth.com/ is the home page for the Bluetooth Special Interest Group (SIG), but at the time of this writing it appears that standardization activity is not included at this web site.

This page intentionally left blank

CHAPTER 3

Radio Transmitters and Receivers

Daniel M. Dobkin

3.1 Overview of Radios

3.1.1 The Radio Problem

It's easy to build a radio. When I was a kid, one could purchase "crystal radio" kits, consisting of a wirewound inductor antenna, a diode, and an earpiece; by tuning a plug in the inductor, one could pick up nearby AM radio stations quite audibly (at least with young pre-rock-amplifier ears). To build a good radio is harder, and to build a good cheap radio that works at microwave frequencies is harder still. That such radios are now built in the tens of millions for wireless local area networks (WLANs) (and the hundreds of millions for cellular telephony) is a testament to the skills and persistence of radio engineers and their manufacturing colleagues, the advancement of semiconductor technology, and the large financial rewards that are at least potentially available from the sale and operation of these radios.

The key requirements imposed on a good WLAN radio receiver are as follows:

- *Sensitivity*: A good radio must successfully receive and interpret very small signals. Recall that the thermal noise in a 1-MHz bandwidth is about -114 decibels from a milliwatt (dBm) or about 4 fW (4×10^{-15} W). Depending on the application, the requirements on the radio may be less stringent than that, but in general extremely tiny wanted signals must be captured with enough signal-to-noise ratio (S/N) to allow accurate demodulation.
- *Selectivity*: Not only must a radio receive a tiny signal, it must receive that signal in the presence of vastly more powerful *interferers*. An access point on channel 1 should be able to receive the signal from a client across the room and down the hall at -90 dBm, even though a client of a different access point on channel 6 is right below the access point, hitting it with a signal at -40 dBm (that's a hundred thousand times more power).
- *Dynamic range*: The same access point should be able to receive its client radio's transmissions as the user moves their laptop from their office to the conference room directly under the antenna—that is, the receiver must adapt to incoming wanted signal power over a range of 60 or 70 dB. The same receiver that can accurately detect a few femtowatts needs to deal with tens of microwatts without undue distortion of the signal.

- *Switch states*: All current WLAN and wireless personal area network technologies are *half-duplex*—the same channel is used for transmitting and receiving, and the transmitter and receiver alternate their use of the antennas. A receiver needs to be turned off when the transmitter is on, both to minimize any chance of a damaging overload from leakage of the transmitted signal and to minimize power consumption; it then needs to return to full sensitivity when the transmission is done, quickly enough so that no data are missed. Most WLAN radios also allow for the use of either of two *diversity* antennas: the receiver must decide which antenna to use quickly enough to be ready for the transmitted data.

A WLAN radio transmitter has a different set of requirements:

- *Accuracy*: The transmitter must accurately modulate the carrier frequency with the desired baseband signal and maintain the carrier at the desired frequency.
- *Efficiency*: The transmitter must deliver this undistorted signal at the desired absolute output power without wasting too much DC power. The final amplifier of the transmitter is often the single largest consumer of DC power in a radio.
- *Spurious radiation*: Distortion of the transmitted signal can lead to radiation at frequencies outside the authorized bands, which potentially can interfere with licensed users and is frowned upon by most regulatory authorities. (We discuss in more detail how this *spurious* output arises in section 3.2.2.) Production of clean spur-free signals is often a trade-off between the amount of radio frequency (RF) power to be transmitted and the amount of DC power available for the purpose.
- *Switch states*: Like the receiver, the transmitter should turn off when not in use to save power and avoid creating a large interfering signal and turn back on again quickly, so as not to waste a chance to send data when there are data to send.

These requirements are actually relatively undemanding compared with other applications of microwave radios. For example, the use of a half-duplex architecture results from the limited bandwidth historically available in the unlicensed bands. Licensed cellular telephony generally uses *paired bands* for transmission and reception; for example, in the United States, a phone operating in the "cellular" bands will transmit at some channel in the range 825–849 MHz and receive at a corresponding channel 44 MHz higher in the 869–894 band. Older digital phones use separate time slots for transmission and reception and are therefore also half-duplex, but old analog phones as well as modern General Packet Radio Service and code-division multiple access phones are *full-duplex*: they transmit and receive at the same time. As WLAN evolves toward simultaneous operation in both the 2.4-GHz ISM and 5-GHz Unlicensed National Information Infrastructure (UNII) bands, simultaneous transmission and reception may become more common in WLAN. However, constructing a filter to separate 2.48 and 5.2 GHz (a 96% change) is much easier than playing the same trick between, for example, 830 and

874 MHz (a 5% change). In addition, for 802.11 radios, the medium access control (MAC) interprets all failed transmissions as collisions and tries again: the system fails gracefully in the presence of radio deficiencies and so can tolerate performance that might be unacceptable for other applications.

3.1.2 Radio Architectures

All modern digital radios have one or more antennas at one end and a set of *analog-to-digital converters* (ADCs) and *digital-to-analog converters* (DACs) at the other. The signal at the antenna is (for most WLAN and wireless personal are network applications) at a frequency in the GHz range, whereas the ADCs and DACs operate with signals up to a few tens of megahertz. Among their other services, radios must provide *frequency conversion* between the antenna and the digital circuitry.

Ever since its invention by Edwin Armstrong in 1917, the *superheterodyne* architecture (superhet for short) has dominated the design of practical radios. The superhet receiver uses two frequency conversion steps: first, the received frequency is converted to an *intermediate frequency* (IF), and second, after some amplification and filtering, the IF is again converted to baseband. For a WLAN application, the RF frequency of 2.4 GHz might be converted to a common IF of 374 MHz and then down to DC (Figure 3.1). Note that the final down-conversion is performed twice, to produce an I and Q output (recall I and Q are the in-phase and quadrature components of the signal). This is necessary because the down-conversion operation can't tell the difference between frequencies above and below the carrier, so they end up on top of each other when the carrier is converted to zero frequency. By doing two conversions using different phases of the carrier, we can preserve information about the sidebands. We provide a detailed example of the use of this sort of trick in our discussion of the image-reject mixer in section 3.2.3.

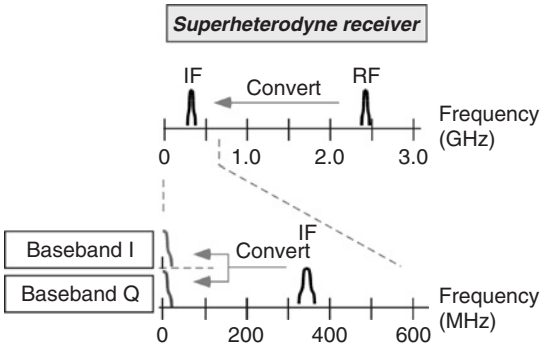


Figure 3.1: WLAN Superheterodyne Frequency Plan

A WLAN superheterodyne receiver might look like the block diagram in Figure 3.2. Here for simplicity we show only one final down-converter. Mixers are used to convert the RF signal

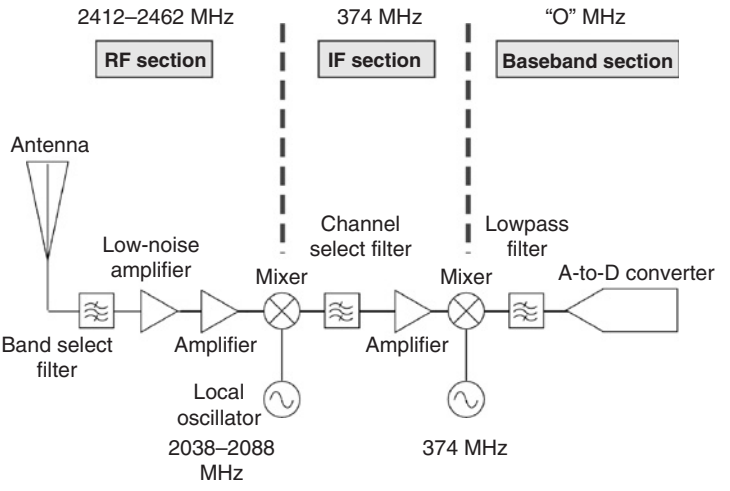


Figure 3.2: Block Diagram of a Superheterodyne Receiver

to IF and the IF signal to a baseband ("0" frequency) output. The band select filter typically accepts the whole RF band of interest (e.g., the whole 2.4- to-2.483-GHz ISM band), and the following amplifiers should work over this band. The first local oscillator (LO) must be tunable. The IF is the difference between the wanted RF frequency and the LO frequency; to receive, for example, channel 1 at 2412 MHz, we set the LO at $(2412 - 374) = 2038$ MHz. However, once this conversion is complete, the IF is always the same, and DC is DC: all other radio parts can be fixed in frequency. (In modern radios, the baseband filters may have adaptive bandwidths to improve performance in varying conditions.)

The use of an IF has a number of benefits. In most cases, we'd like to listen to signals on a particular channel (e.g., channel 1 at 2.412 GHz) and not neighboring channels. An electrical filter circuit is used to select the wanted channel from all channels. The task of filtering is much easier if it is performed at the IF rather than the original RF frequency. For example, a 20-MHz channel represents 0.83% of the RF frequency of 2.4 GHz but 5% of the 374-MHz IF; it is both plausible and true that distinguishing two frequencies 5% apart is much easier than the same exercise for a 1% difference. Furthermore, a radio needs to allow tuning to different channels: a user would hardly be thrilled about having to buy three radios to make use of the three channels in the ISM band. It is difficult to accurately tune a filter, but in a superhet architecture we don't need to: we can tune the frequency of the LO instead. The fact that the IF is fixed also makes it easier to design amplifiers and other circuitry for the IF part of the radio. Finally, all other things being equal, it is easier to provide a given amount of gain at a low frequency than at a high frequency, so it costs less to do most of one's amplification at the IF than at the original RF frequency. The choice of an IF is a trade-off between selectivity,

filtering, requirements on the LO, and component cost; we discuss this issue, frequency planning, in section 3.3.

The performance of a superhet receiver is dominated by certain key elements. The *low-noise amplifier* (LNA) is the main source of excess noise and typically determines the sensitivity of the radio. The cumulative distortion of all the components (amplifiers and mixers) before the channel filter plays a large role in determining how selective the receiver is: a distorted interfering signal may acquire some power at the wanted frequency and sneak through the channel filter. Finally, enough gain adjustment must be provided to allow the output signal to lie within the operating limits of the ADC, which are typically much narrower than the range of input RF powers that will be encountered.

A superhet transmitter is the opposite of a receiver: the I and Q baseband signals are mixed to an IF, combined, and then after amplification and filtering converted to the desired RF (Figure 3.3; again we show only one branch of the baseband for clarity). In a WLAN radio, the IF is generally the same for both transmit and receive to allow the use of a single channel filter in both directions and to share LOs between transmit and receive functions.

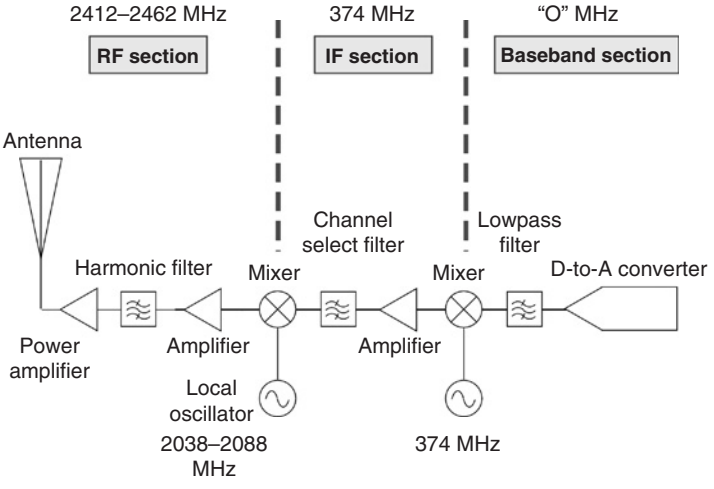


Figure 3.3: Block Diagram of a Superheterodyne Transmitter

As with the superhet receiver, in the transmitter all components before the final mixer operate at fixed frequencies (baseband or IF). The final mixer and amplifiers must operate over the whole RF bandwidth desired. The transmitter performance is often dominated by the power amplifier. Distortion in the power amplifier causes radiation of undesired spurious signals, both into neighboring channels and into other bands. To minimize distortion, the output power can be decreased at fixed DC power, but then the efficiency of the power amplifier is reduced.

Because the power amplifier dominates transmitter performance, the superhet architecture offers fewer benefits in transmission than in reception.

Even though IF filtering is easier than RF filtering, IF filters still constitute an expensive part of a superhet radio; IF amplifiers are also more expensive than amplifiers at baseband. Why not skip the IF stage? Two increasingly common architectural alternatives to superhet radios, *direct conversion* and *near-zero IF* (NZIF), attempt exactly this approach. A direct-conversion radio is what it purports to be: RF signals are converted to baseband in a single step (Figure 3.4). The channel filter becomes an inexpensive low-pass filter from zero frequency to a few megahertz; furthermore, at these low frequencies, active filters using amplifiers can be implemented, allowing versatile digital adjustment of bandwidth and gain.

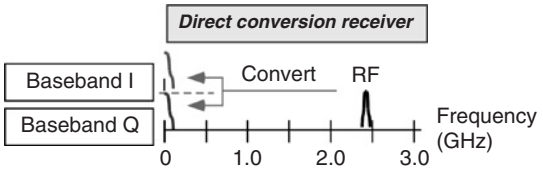


Figure 3.4: WLAN Direct Conversion Frequency Plan

Direct-conversion radios encounter serious challenges as well. Because there is no IF stage, the gain that the IF provided must be placed somewhere else, typically at baseband because it is cheaper to do so. DC offset voltages, which may appear for various reasons, including distortion of the LO signal, may be amplified so much at baseband that they drive the ADC voltage to its rail, swamping the wanted signal. Similarly, low-frequency noise from many causes, including electrical noise in metal-oxide-semiconductor field-effect transistors (MOSFETs) and *microphonic* noise (the result of mechanical vibrations in the radio), can be amplified to an objectionable level by the large baseband gain. Because the LO is at the same frequency as the RF signal in a direct conversion system, it cannot be filtered out and may radiate unintentionally during the receive operation, interfering with other users. The radiated signal may bounce back off external objects in a time-varying fashion and interfere with the wanted signal.

One solution to some of these problems is not to convert to zero but to a very low frequency, just big enough to allow the whole of the received signal to fit. For a WLAN signal, we would choose an IF of about 8 MHz: an *NZIF* receiver. The IF is low enough that no additional conversion is needed; the ADC can accept the IF signal directly. Because now there is no signal at DC, we can filter out any DC offsets without affecting the receiver. Low-frequency noise at less than a few megahertz (encompassing much electrical noise and all microphonic noise) can also be filtered out. However, because the IF in this case is small, the image of the wanted frequency is very close to the wanted frequency and can't be filtered. Some other means of image rejection must be provided. We discuss these alternative architectures in more detail in section 3.3.

3.1.3 A "Typical" WLAN Radio

A radio consists of a transmitter, receiver, and whatever ancillary functions are needed to support their operation. A complete WLAN radio might look something like the block diagram in Figure 3.5. The radio is composed of a superhet transmitter and receiver. In general, the IFs for transmit and receive would be the same so that the LOs and filters could be shared between the two functions, though here we depicted them as separate components for clarity. Both transmit and receive convert in-phase and quadrature (I and Q) signals separately, because as noted above this step is necessary to provide separate access to sidebands above and below the carrier frequency. The heart of the radio is typically implemented in one or few chips. Older systems used silicon integrated circuits with bipolar transistors or heterostructure transistors with a silicon-germanium (SiGe) base layer. In recent years, improvements in integrated circuit technology have allowed most radio chips to migrate to less-expensive MOSFET implementations.

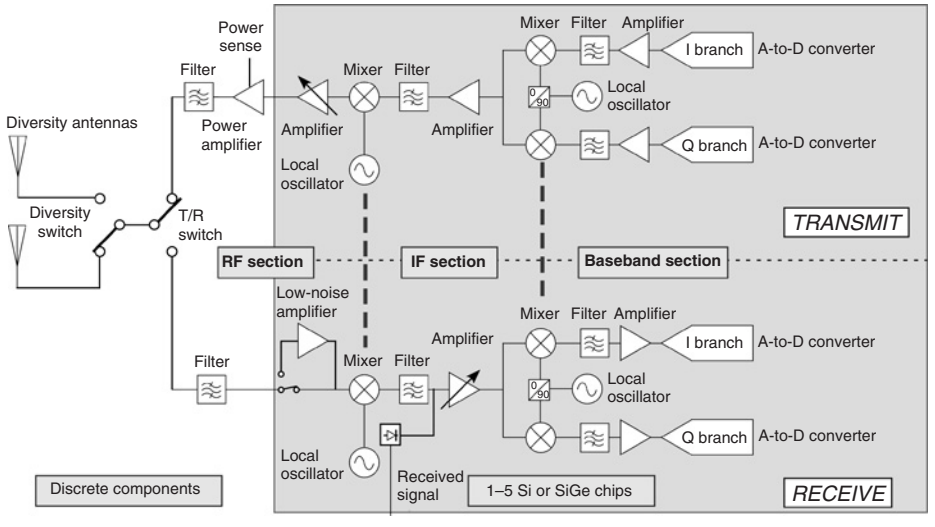


Figure 3.5: Simplified Block Diagram of an Archetypal Superhet WLAN Radio

The power amplifier is often a separate chip, using either a SiGe process or a compound semiconductor such as GaAs. Because WLANs are half-duplex, the antenna must be switched between the transmitter and receiver with a *transmit/receive* (T/R) switch. Most WLAN radios provide the option (at least for the receiver) of using one of two diversity antennas, requiring an additional switch. Switches are currently implemented using one or more separate chips, though we discuss an example of an integrated complementary metal-oxide semiconductor (CMOS) implementation in section 3.2.6. Finally, an important capability of a WLAN radio is the provision of a *received signal strength indication* to the MAC layer to help determine when the channel is clear for transmission.

Many variants of the simple block diagram of Figure 3.5 are possible. Today, multiband radios are available, which may provide simultaneous operation at 2.4 and 5.2 GHz with a single set of antennas and a single baseband/MAC chip. Some WLAN systems use antenna arrays much more complex than a simple diversity pair and may use modified radio architectures. As we noted in section 3.1.2 above, many radios use direct-conversion or NZIF architectures. However, all these radios depend on the same set of functional blocks:

- *ADCs and DACs* to convert between the analog and digital worlds
- *Amplifiers* to increase signal power
- *Mixers* to convert between frequencies
- *Oscillators* to provide the means for conversion and define the frequency of the transmitted or received signals
- *Filters* to select desired frequencies from a multitude of interferers and spurs
- *Switches* to select the required input at the right time

In the next section we examine the key performance parameters that determine the suitability of each type of component for its application and how each block affects the operation of the overall radio system.

3.2 Radio Components

3.2.1 ADCs and DACs

This book is primarily about analog components and issues, but because ADCs and DACs are only half-digital devices we ought to say a few words about them.

Any time-dependent signal, such as a varying voltage or current, that is band limited (i.e., composed of frequencies within a restricted bandwidth) can be completely reconstructed from its value at a tiny number of discrete points, that is, from *samples* of the waveform. According to another theorem of the inevitable Dr. Nyquist, the frequency with which the samples must be acquired is twice the bandwidth of the signal. An example of a sampled signal is depicted in Figure 3.6. The rather complex analog signal is the sum of a signal at normalized frequency 1, another component at $f = 2$, a third at $f = 5$, and a final component at $f = 7$. We show the sampling operation taking place on each component separately, which is correct because sampling is a linear operation (a sample of $(A + B)$ = the sum of a sample of A and a sample of B). Because the highest frequency is 7, 14 samples are needed to define the signal. It is obvious that this is more than we need to determine the amplitude of the $f = 1$ and $f = 2$ components but just enough to find the $f = 7$ component. Note that the theorem assumes

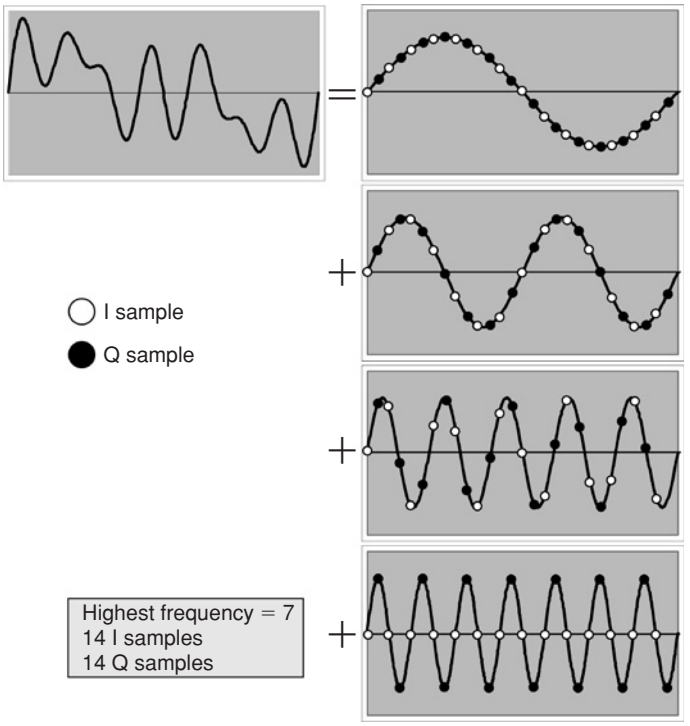


Figure 3.6: Nyquist Sampling of a Band-Limited Signal

that the sampled values are complex; in terms of real signals, this means that the signal must be sampled both in-phase and at quadrature or equivalently that the signal must be mixed to produce an I and Q output, each of which is sampled separately at 2(bandwidth) to provide the whole.

In principle, a modulated carrier (i.e., the received RF signal) could be directly sampled at a sampling rate equal to the bandwidth of the modulation, because the information is contained in the modulation rather than the carrier. Such a scheme, known as a subsampling mixer, simultaneously achieves frequency conversion and ADC. However, this approach is not commonly used; most often, the RF signal is down-converted to baseband using analog components, and the resulting I and Q signals are sampled at (at least) twice the information bandwidth. Note that the bandwidth required to send a data rate R is approximately $R/2$. (You can see why this should be so by imagining a string of alternating 1s and 0s, the highest frequency digital data stream: the fundamental is obviously a sinusoid of frequency equal to half the data rate.)

The same reasoning works in reverse: a band-limited analog signal can be reconstructed by providing samples at the Nyquist rate. Thus, similar requirements are imposed on

the DAC: a minimum of 11 megasamples/second (Msps) in the case of 802.11 classic or 802.11 b must be output to define the desired transmitted signal. A DAC attempts to produce a constant voltage starting at the moment of the sample and continuing until the next sample output; the resulting stairstep waveform is passed through a filter to smooth the transitions from one value to the next, reconstructing the desired band-limited analog signal.

The simplest way to construct an ADC is to put a tapped resistor between two voltages (e.g., the supply voltage and ground) and then compare the input voltage with the voltage on each tap using a comparator. The value of the voltage will be shown as the tap at which the comparator outputs change. This architecture is known as a *flash ADC* (Figure 3.7). Flash ADCs are extremely fast, because all the operations take place in parallel, and have been constructed to operate at multi-gigahertz sampling rates. However, they don't scale very well: to obtain an accurate estimate of the signal voltage—equivalently, to have many bits of resolution—many taps are required, with an amplifier and comparator for each tap. The resistor voltage divider must be very accurate, and the number of components (and power consumption) doubles with each bit of resolution.

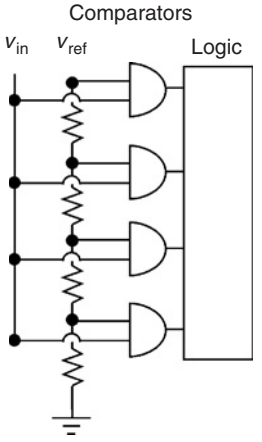


Figure 3.7: Flash ADC

Two alternative architectures allow the use of low-resolution ADC modules to achieve high-resolution overall conversion and provide better scaling and more sensible implementation for WLAN applications. The first is the *pipelined ADC* (Figure 3.8). A low-resolution ADC (here a 4-bit example is shown) digitizes the input signal to provide a coarse estimate—that is, the most significant bits—of the input sampled voltage. The result, as well as being output, is sent to a DAC, which provides modest resolution but high precision for the few states it does produce. The DAC output is subtracted from the input voltage and after amplification (not

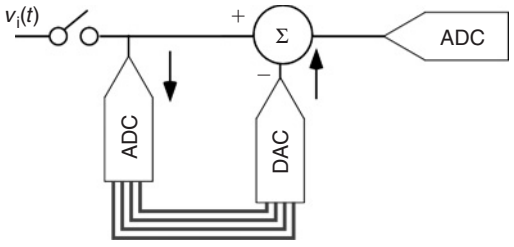


Figure 3.8: Pipelined ADC

shown) is applied to a second-stage ADC, again low resolution, to get a fine estimate of the residual voltage—the least significant bits.

Thus, one can combine two 4-bit ADCs to get a total of 8 bits of resolution. In practice, additional stages are used, and the actual output resolution is somewhat less than the sum of the resolution of all the ADCs to allow correction for overlap errors. Because several steps are involved in converting each sample, pipelined ADCs have some latency between the receipt of a voltage and the transmission of a digital value. By loading several samples into the pipeline at once and converting in parallel, the overall delay associated with the ADC operation can be minimized. Pipelined ADCs can permit sample rates of tens of megasamples per second at reasonable size and power consumption.

Another common architecture, the *sigma-delta ADC*, uses a feedback loop to provide high resolution from a modest-resolution conversion stage (Figure 3.9). At each sampling time, a voltage is provided to the input of the ADC. The voltage is subtracted from the current estimate of the sampled value, as provided by a DAC. The resulting difference is then integrated and provided to an ADC, which may be low resolution because its job is only to correct the estimate. After a time, hopefully short compared with the sample time, the estimate converges to the sampled input, and the correction voltage goes to 0.

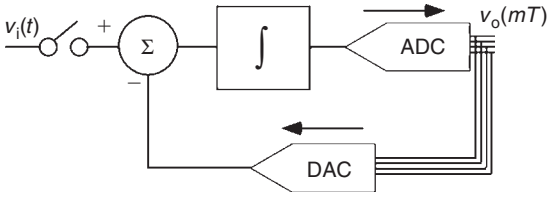


Figure 3.9: Sigma-Delta Analog-to-Digital Converter

The sigma-delta design trades speed for resolution. One-bit converters are often used, and the sampling rate is much higher than the desired output rate. Digital filters on the output remove the resulting noise to produce a reasonably accurate estimate of the sampled voltage.

Each ADC architecture has a corresponding DAC architecture. For example, the analog of a flash ADC is a *current-steering DAC*, in which binary-weighted current sources are directed by switches to the output based on the binary value of the current output string. Demands on DACs are generally somewhat less stringent than ADCs, because there are no interferers, wild variations in input signal, or other unpredictable conditions to deal with. It should be noted, however, that in higher power access points, the broadband noise of the DACs can become a significant contributor to out-of-band power, particularly in direct conversion architectures where the noise cannot be readily filtered. More discussion of noise and architecture can be found in section 3.3 of this chapter.

What speed and resolution are required for a WLAN application? In the case of an 802.11 classic signal, at least 11 Msps would be required to capture the 11 Mchips/second signal. Many systems take more than the minimum number of samples (*oversampling*) to obtain an improved estimate of the received signal. An 802.11a or g (or HiperLAN 2) signal contains an orthogonal frequency-division multiplexing (OFDM) symbol every 4 μsec, and each symbol must be sampled (at least) 64 times to produce the 64 possible subcarriers: $(64/4 \times 10^6) = 16$ Msps.

Recall from Chapter 1 that a binary phase-shift keying (BPSK) signal (802.11 classic) requires about 9.5 dB of signal to noise for accurate demodulation: this is equivalent to only 2–3 bits of ADC resolution (recalling that the ADC samples a voltage not a power, so 9.5 dB = a factor of 3 in voltage). Higher rate quaternary phase-shift keying (QPSK) signals need about 12 dB, or around 4 bits. It is more subtle to decide how much resolution is required for an OFDM signal. Recall that a 64 quadrature-amplitude-modulation (QAM) signal (a subcarrier in the high-rate modes) needs about 26 dB (S/N) for accurate demodulation: that's about 5 bits of resolution. However, (S/N) is measured relative to the average signal; the peak (corner) points of a 64 QAM constellation are about 4 dB above the average power, so we're up to 6 bits. Then we add a bunch of these signals together to form the OFDM symbol. One would naively expect that on the order of 6 extra bits would be needed to describe the 52 active carriers ($2^6 = 64$, after all), but simulations show that about 7–9 total bits are sufficient. (Details are provided in Côme et al. in section 3.6.) If we design the ADCs to provide exactly the necessary resolution to demodulate the RF signal, we'd better have exactly the right signal amplitude coming in, requiring perfect gain adjustment. In practice it seems prudent to provide a few bits of resolution to allow for imperfect adjustment of the analog signal gain. Thus, we'd expect to need around 6 bits for 802.11 classic/b and perhaps 8–10 bits for 802.11a or g.

A summary of ADC and DAC parameters for various reported 802.11 chip sets is provided in Table 3.1. A comparison with recent commercial stand-alone ADCs may be made by reference to Table 3.2. We see that our expectations for resolution are roughly met: 802.11b requires 6 bits or so of ADC resolution, whereas 802.11a or g chips provide 8 bits of resolution, both with oversampling used when the sampling rate is reported.

Table 3.1: Survey of 802.11 ADC/DAC Performance

Vendor	Intersil/Virata	Broadcom	Broadcom	Marvell	Thomson
Protocols supported	802.11b	802.11b, g	802.11a	802.11b	802.11a
DAC resolution/speed	6 b 22 Msps	8 b	8 b	9 b 88 Msps	8 b 160 Msps
ADC resolution/speed	6 b 22 Msps	8 b	8 b	6 b 44 Msps	8 b 80 Msps
DC power: transmit	720 mW	144 mW	380 mW	1250 mW	920 mW
DC power: receive	260 mW	200 mW	150 mW	350 mW	200 mW
Reference	Data sheets	Trachewsky et al. HotChips 2003	Trachewsky et al. op. cit.	Chien et al. ISSCC 2003 paper 20.5	Schwanenberger et al. ISSCC 2003 paper 20.1

Table 3.2: Survey of Commercial Stand-Alone ADCs

Part	Resolution (bits)	Rate (Msps)	(S/N) (dB)	Quantization (S/N) (dB)	Power (mW)
AD6645	14	105	73.5	90	1500
SPT7722	8	250	46	54	425
MAX1430	15	100	77	96	2000
ADC10080	10	80	60	66	75
ADC10S040	10	40	60	66	205
After "High Speed ADC's . . . , " David Morrison, *Electronic Design* 6/23/03, p. 41.					

The performance of integrated ADCs is modest compared with the state of the art for stand-alone devices (Table 3.2), and by comparison of power consumption one can see why: for example, the high-performance AD6645 provides 14 bits of resolution at 105 Msps, adequate to enable a very simplified radio architecture with reduced gain control, but its power consumption is greater than that of the whole Broadcom radio chip. Functional integration demands trade-offs between the performance, power consumption, and cost of each of the components that make up the radio. Nevertheless, performance achievable in integrated implementations is adequate for WLAN requirements.

3.2.2 Amplifiers

The key properties of an amplifier are gain, bandwidth, noise, distortion, and power. Gain is the amplifier's *raison d'être*: it is the ratio of the size of the output to that of the input. Amplifiers must respond rapidly to be part of the RF chain; bandwidth is less significant in the rest of the radio. Noise is what an amplifier should not add to the signal but does. The output signal ought to be just a multiplied version of the input signal, but it isn't; distortion in the time domain shows up as spurious outputs in the frequency domain. Distortion usually represents the main limitation on amplifier output power. Let's consider each of these aspects of amplifier performance in turn.

Gain is the ratio of the output signal to the input signal. In RF applications, power gain is usually reported, because power is easier to measure than voltage: G = power out)/(power in). In radio boards and systems, one can usually assume that the input and output of the amplifier are matched to a 50 V environment; however, amplifiers integrated into a chip may use differing impedance levels. At low IF frequencies or at baseband, high impedances may be used, and power gain becomes less relevant than voltage gain. Gain is harder to get at RF than at IF or baseband, so most designs use only as much high-frequency gain as necessary to achieve the desired noise performance.

Two general classes of devices are available for providing gain at gigahertz frequencies: field-effect transistors (FETs) and bipolar junction transistors (BJTs) (Figure 3.10). FETs are lateral devices: electrons move sideways in a *channel* under a *gate* from a *source* to a *drain*, with the number of electrons being determined by the potential on the gate. The output current is the product of the number of electrons in the channel and their velocity. The velocity the electrons can reach is mainly determined by the type of semiconductor used for the device. (The velocities of holes are so much lower than those of electrons that p-type devices are rarely if ever used for high-frequency amplifiers.) The number of electrons is roughly linear in the gate-source voltage.

Bipolar transistors are vertical devices: electrons are injected from an n-type *emitter* into a p-type *base*, in which they drift until they are collected by the n-type *collector* or (more rarely) recombine with a hole in the base region. The number of electrons in the base is set by the likelihood that an electron will be able to thermally hop over the emitter-base potential barrier and is thus exponentially dependent on the barrier height. Therefore, the current in a bipolar transistor is exponential in the input voltage; for small variations in input voltage, the corresponding change in output current grows linearly in the current:

$$\frac{d}{dv_{\text{in}}} I_{\text{out}} \approx \frac{d}{dv} e^{\frac{-qv_{\text{in}}}{\text{k}T}} = \frac{-q}{\text{k}T} e^{\frac{-qv_{\text{in}}}{\text{k}T}} = \frac{-q}{\text{k}T} I_{\text{out}} \tag{3.1}$$

where q is the electron charge, k Boltzmann's constant, and T the absolute temperature.

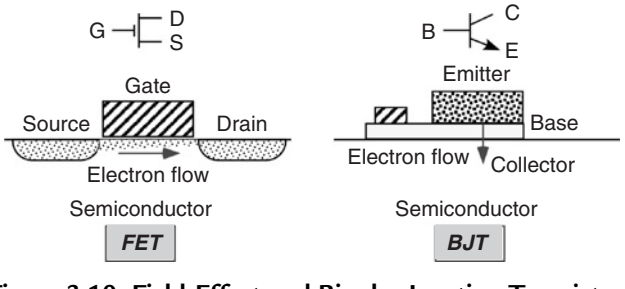


Figure 3.10: Field-Effect and Bipolar-Junction Transistors

To compare them, it is useful to think of the semiconductor device as a *transconductance* amplifier: it accepts a small RF voltage and produces a large output current. The ratio of the change in current to the change in voltage that produces it is the transconductance, $g_m = dI_{out}/dv_{in}$. To make a transconductor into a voltage amplifier, we put the current through a load resistor to convert it into a voltage. The power dissipated is the product of current and voltage. The gain of the amplifier at frequencies low enough that parasitic capacitances can be ignored is

$$G = \frac{P_{out}}{P_{in}} = \frac{(g_m v_{in})^2 R_L}{v_{in}^2 / R_{in}} = g_m^2 R_L R_{in} \tag{3.2}$$

To get high gain, one needs high transconductance. The transconductance of an FET device is approximately the product of the gate capacitance (the change in charge due to a change in voltage) and the velocity of the charge and is thus roughly determined by the width of the channel and the type of semiconductor. To achieve higher electron velocities, one can progress from silicon ($V_{el} \approx 8 \times 10^6$ cm/sec) to GaAs ($V_{el} \approx 1.5 \times 10^7$ cm/sec). To go farther, one can construct *high-electron-mobility transistors* (HEMTs) in which the electrons are located at a heterojunction between two semiconductors of differing bandgap, removing them from dopant atoms and consequent scattering. However, the amount of improvement is limited by the fact that semiconductors with the highest electron velocities, such as InGaP, have different crystal lattice sizes from common substrates such as GaAs. The lattice mismatch can be tolerated if the layers of InGaP are sufficiently thin: the resulting *pseudomorphic* HEMT, or pHEMT, can achieve effective electron velocities of around 2×10^7 cm/sec. Heterostructure devices also benefit from much improved electron mobility (the ratio of velocity to electric field at low velocities), which can be many times higher than mobility in silicon. In practice, FET devices can obtain transconductance of tens (silicon) to hundreds (pHEMT) of millisiemens for a channel width of a few hundred microns.

In a bipolar device, transconductance increases with collector current (equation [4.1]), and collector current is exponential in the DC input voltage. By turning up the current density, it is easy to obtain very large transconductances from a bipolar transistor: recalling that at room temperature kT/q is about 0.026 V, g_m for a bipolar is about $40(I_{out})$, irrespective of the technology used. A collector current of 100 mA (which is quite reasonable for a transistor with a periphery of a few hundred microns) will provide 4000 mS or 4 S of transconductance, much more than can be obtained from a FET of similar size. Low-frequency gain is cheap for a bipolar. Much of this gain is then often converted to linearity by exploiting negative feedback configurations. On the other hand, parasitic capacitances in BJTs are large relative to FET capacitances and vary strongly with applied voltage. This makes design of broadband amplifiers somewhat more difficult with BJTs than with FETs. Bandwidth is roughly determined by the ratio of some measure of the gain (for example, g_m) to some measure of the parasitics, typically a capacitance. Bipolar devices can have more transconductance but also more capacitance than FETs; the net result is that BJTs display similar bandwidth to FETs

of similar dimensions. One important advantage of BJTs is that, being vertical devices, the critical length in the direction in which the electrons flow is determined by the growth of a layer or diffusion of dopants into a layer, both of which are relatively easy to control compared with the width of a submicron feature such as a FET gate. Therefore, BJTs have historically been easier and cheaper to produce for high-frequency applications and are often the first technology applied to a microwave problem, with silicon-based FET technology catching up as lithographic dimensions decrease.

In the receiver, the job of the amplifier chain is to deliver an output voltage within the dynamic range of the ADC. Because the input signal power can vary over such a large range, gain adjustability must be provided. Some adjustment in either the magnitude of the digital output signal or the gain or both must also be provided for the transmitter, though in this case the output range is modest in comparison with that faced by the receiver: 10–20 dB variation is typical.

In Table 3.3 we show typical gain for some single-stage commercial amplifiers operating at 2 GHz, using various technologies and compared with a representative silicon CMOS result. Note that GaAs FET devices are fabricated with the metal gate directly on the surface of the semiconductor and are thus often referred to as MEtal-Semiconductor FETs or MESFETs. Power gain of around 15 dB for a single stage is a reasonable expectation at the sort of frequencies used in WLAN radios.

Table 3.3: Gain at 2 GHz, Commercial Devices, Various Technologies

Technology	Gain (2 GHz)	Reference
GaAs pHEMT	16 dB	Agilent 9/03
GaAs MESFET	14 dB	WJ Comm 9/03
SiGe HBT	17 dB	Sirenza 9/03
Si CMOS, 0:8 mm	15 dB	Kim MTT '98

To put this number in context, remember that a typical input signal of 80 dBm needs to be amplified to a convenient value for digital conversion: on the order of 1 V at 1 KV, or 1 mW (0 dBm). Thus, we need around 80 dB total net gain, partitioned between the RF sections, IF if present, and baseband. (Actually, more is needed to overcome losses in conversion and filtering.) We can see that several gain stages are likely to be needed, though at this point it isn't obvious how to allocate the gain between the parts of the radio.

To elucidate one of the key considerations in determining the amount of RF gain needed, let us proceed to consider noise and its minimization. Recall from Chapter 1, section 1.4.2, that a 50 Ω input load (a matched antenna) delivers −174 dBm/Hz of thermal noise at room temperature into the radio. The radio has its own resistors, as well as other noise sources, and

thus adds some excess noise over and above amplifying the noise at the input. The amount of noise added by an amplifier is generally quantified by defining a *noise factor*, the ratio of the output (S/N) divided by the input (S/N) (Figure 3.11). The noise factor is 1 if the amplifier contributes no excess noise and grows larger when excess noise is present. Note that the noise factor of a single amplifier is independent of the amplifier gain.

$$F = \frac{S/N_{in}}{S/N_{out}} = \frac{\left(\frac{P_{in}}{N_o}\right)}{\left(\frac{GP_{in}}{G[N_o + N_{ex}]}\right)} = \left(\frac{GN_o + GN_{ex}}{GN_o}\right) = 1 + \frac{N_{ex}}{N_o}$$

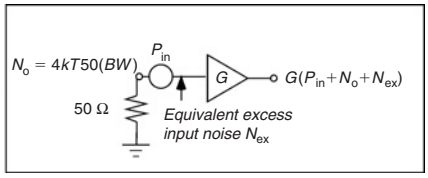


Figure 3.11: Noise Factor Definition

It is common to define the *noise figure* as the logarithm of the noise factor, reported in decibels. Because $\log(1) = 0$, the noise figure NF is 0 for an ideal noiseless amplifier and grows larger as more excess noise is added.

Amplifier gain G rears its head when we try calculating the noise factor of two amplifiers in sequence—*cascaded* amplifiers—as shown in Figure 3.12. The excess noise contributed by the first amplifier is amplified by both G_1 and G_2, whereas the excess noise of the second amplifier is magnified only by the second-stage gain G_2. In consequence, the noise factor of the cascaded combination is the sum of the noise factor of the first stage, F_1, and the excess noise of the second stage divided by the gain of the first stage. *If the first stage gain is large, overall system noise factor is dominated by the noise in the first stage.*

$$F = \frac{S/N_{in}}{S/N_{out}} = \frac{\left(\frac{P_{in}}{N_o}\right)}{\left(\frac{G_1G_2P_{in}}{G_1G_2[N_o + N_{ex1}] + G_2N_{ex2}}\right)}$$

$$= 1 + \frac{N_{ex1}}{N_o} + \frac{N_{ex2}}{G_1N_o} = F_1 + \frac{F_2 - 1}{G_1}$$

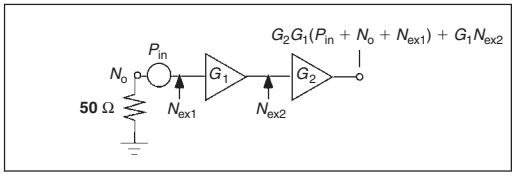


Figure 3.12: Noise Factor of Two Amplifiers in Cascade

For this reason, we always need to allocate enough gain in the first RF amplifier stage or two to minimize the noise contribution of the rest of the chain. A typical first-stage gain of 15 dB (a factor of 30) is usually enough to ensure that the first stage—the *low-noise amplifier* (*LNA*), so named because its (hopefully small) noise figure determines the noise in the whole receiver—does indeed fulfill its dominant role.

The noise figure of the chain, once obtained, provides a quick insight into the noise performance of the radio: the effective noise floor is just the thermal noise (in dBm) plus the noise figure (in dB). For example, a radio with a 1-MHz input bandwidth and a 3-dB noise figure has a noise floor of $(-174 + 60 + 3) = -111$ dBm.

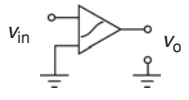
In Table 3.4 we include noise figure in our amplifier comparison. pHEMTs are the champions of the low-noise world and are widely used in applications like satellite receivers where noise figure is a key influence in overall performance. The other device technologies all readily achieve noise figures of a few decibels, adequate for typical WLAN/wireless personal network applications. (Note that these are not the lowest noise figures that can be achieved in each technology but are those obtained in practical devices optimized for overall performance.)

Table 3.4: Gain and Noise Figure at 2 GHz; Devices From Table 3.3

Technology	Gain (2 GHz)	NF (2 GHz)	Reference
GaAs pHEMT	16 dB	1.2 dB	Agilent 9/03
GaAs MESFET	14 dB	3 dB	WJ Comm 9/03
SiGe HBT	17 dB	3 dB	Sirenza 9/03
Si CMOS, 0:8 mm	15 dB	3 dB	Kim MTT '98

Noise sets the lower bound on the signals that can be acquired by a receiver. The upper bound, and more importantly the upper bound on interfering signals that can be present without *blocking* the reception of the tiny wanted signal, is set by distortion. Distortion redistributes power from the intended bandwidth to other *spurious* frequencies. In the receiver, distortion spreads the power from interferers on other channels into the wanted channel; in the transmitter, distortion splatters transmitted power onto other folks' channels and bands, potentially earning the wrath of neighbors and regulators. How does distortion do its dirty work, how is it measured, and how does the radio designer account for its effects?

Let us consider an amplifier with a simple *memoryless* distortion, depicted in Figure 3.13. (Memoryless refers to the fact that the output depends only on the amplitude of the input and not on time or frequency; real amplifiers remember what is done to them—though they often forgive—but incorporating phase into the treatment adds a lot of complexity for modest rewards and is beyond our simple needs.)



$$v_o = a_1 v_{in} - a_2 v_{in}^2 - a_3 v_{in}^3$$

Figure 3.13: Amplifier With Simple Memoryless Polynomial Transfer Characteristic

The output of the amplifier is taken to be accurately expressed as a polynomial function of the input voltage. The coefficients are taken here to be real numbers for simplicity. Further, we assume in most cases that the input voltage is small enough so that the nonlinear (quadratic and cubic) contributions are much smaller than the ideal linear amplification.

What happens when we plug a pure sine wave input into this distorting amplifier? First we'll do the math and then provide some pictures to show what it means. The output voltage is obtained by plugging the sinusoidal input into the polynomial:

$$v_0 = a_1 v_\mathrm{i} \sin(\omega_1 t) - a_2 v_\mathrm{i}^2 \sin^2(\omega_1 t) - a_3 v_\mathrm{i}^3 \sin^3(\omega_1 t) \tag{3.3}$$

To rephrase this equation in terms that have more direct spectral relevance, we exploit a couple of trigonometric identities:

$$\sin^2 x = \frac{1}{2} - \frac{1}{2}\cos(2x) \tag{3.4}$$

and

$$\sin^3 x = \frac{3}{4}\sin x - \frac{1}{4}\sin(3x) \tag{3.5}$$

After a bit of algebra we can partition the output in a fashion that clearly shows how each type of distortion creates specific contributions to the spectrum of the output (Figure 3.14). The quadratic term, or *second-order distortion*, creates a constant (DC) output voltage and a new component at twice the input frequency. The cubic term, a *third-order distortion*, produces no DC component but adds both an undesired nonlinear term at the input (*fundamental*) frequency and a new term at three times the fundamental. Naturally, the math proceeds in an essentially identical fashion for a cosine input instead of a sine.

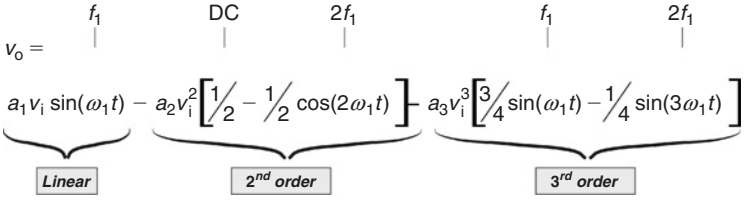


Figure 3.14: Output Voltage of a Distorting Amplifier Displayed as Linear, Second-, and Third-Order Contributions

Some insight into how these terms arise can be had by reference to Figures 3.15 and 3.16, showing (grossly exaggerated) images of the time-domain distortion. Figure 3.15 shows what would happen with a pure second-order amplifier. The amplifier only produces positive output voltages, so an input signal with an average voltage of 0 obviously must produce an output

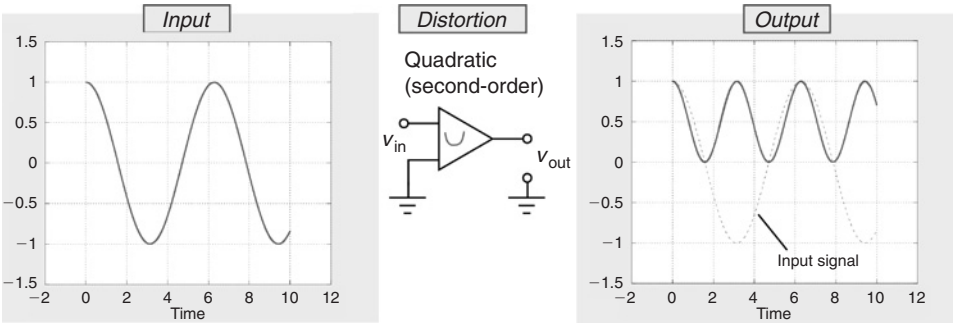


Figure 3.15: Pure Second-Order Distortion of a Sinusoidal Input Signal

with a positive average: a DC offset has been introduced by the distortion. The quadratic characteristic takes the negative-going peaks of a sine and inverts them; thus, there are two maxima per (input) cycle instead of one: the frequency of the input has been doubled.

A pure cubic distortion is shown in Figure 3.16. The transfer characteristic is symmetric about 0, so no DC offset results. Clearly, the output generally resembles the input signal: a component is present at the input frequency. However, the output is not the same shape as the input. In fact, we can see that the distortion alternates sign six times in one cycle of the fundamental (at about $t = 0{:}5$, 1.6, 2.6, 3.7, 4.7, and 5.8, where the distorted signal and the input signal cross), that is, there is a component at three times the frequency of the fundamental.

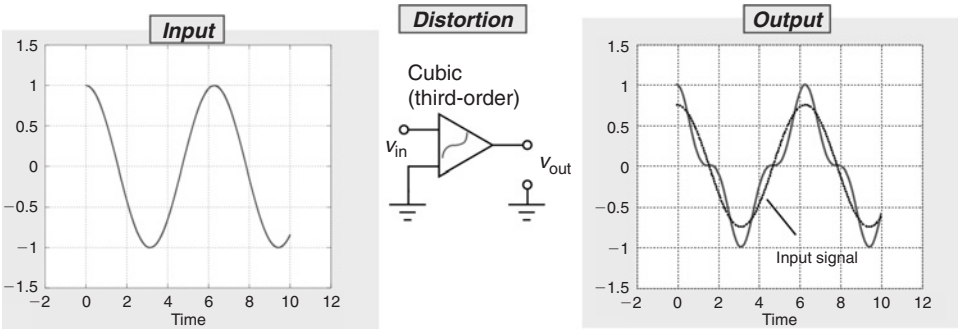


Figure 3.16: Pure Third-Order Distortion of a Sinusoidal Input Signal

The resulting spectra are depicted in Figure 3.17. The spectrum contains a DC offset due to the presence of second-order distortion, a change in the amplitude of the fundamental (relative to an ideal linearly amplified input) due to the third-order distortion, and second and third harmonics of the input frequency.

So far it's not entirely obvious why a WLAN radio designer is very concerned. For example, an 802.11b radio operates in the ISM band. Say the fundamental is at channel 1, 2.412 GHz. The

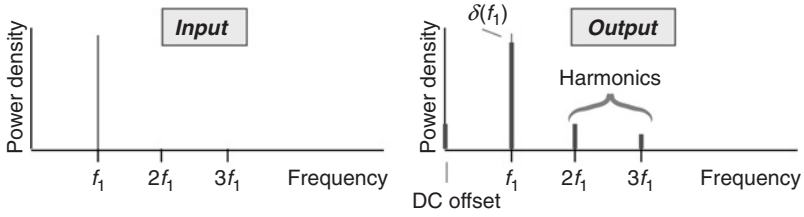


Figure 3.17: Spectrum of a Distorted Signal

second and third harmonics, at, respectively, 4.824 and 7.236 GHz, are so far from the band of interest that they are easily filtered out; the DC component can be removed by using a capacitor between stages. Only the tiny distortion of the main signal appears to be consequential. (Note that this statement is a bit too glib if a direct-conversion architecture is being used: the DC term in the output may be important if we have converted directly to zero frequency. We discuss DC offsets in section 3.3 of this chapter. We could also be in for some trouble in an ultrawideband radio, where the distortion products at twice the input frequency are still in our operating band. (Avoiding this situation is a good reason to partition bands into subbands less than an octave [factor of 2 in frequency] wide: see for example the subbands defined in Figure 2.29.)

However, as we discussed in Chapter 1, real signals are modulated and therefore contain a range of frequencies. What happens if an input with more than one frequency is presented to an amplifier with distortion? The smallest integer that isn't 1 is 2: let's see what happens if we insert sinusoidal signals at two (neighboring) frequencies into our distorting amplifier. The result, after some uninteresting algebra again exploiting the trigonometric identities of equations [3.4] and [3.5] and the additional product identity [3.6], is summarized in Figure 3.18. It's a mess. To make any sense of it we need to split up the groups of terms and examine them separately.

$$\sin(x)\sin(y) = \frac{1}{2}\cos(x - y) - \frac{1}{2}\cos(x + y) \tag{3.6}$$

$$\begin{aligned}
&\text{Linear} \\
v_o = {} & a_1 v_{i1}\sin(\omega_1 t) + a_1 v_{i2}\sin(\omega_2 t) \\
&\text{2nd order} \\
& - a_2\left[v_{i1}^2\left(\tfrac{1}{2} - \tfrac{1}{2}\cos(2\omega_1 t)\right) + v_{i2}^2\left(\tfrac{1}{2} - \tfrac{1}{2}\cos(2\omega_2 t)\right) + 2v_{i1}v_{i2}\left(\frac{1}{2}\cos(\omega_1 t - \omega_2 t) - \frac{1}{2}\cos(\omega_1 t + \omega_2 t)\right)\right] \\
&\text{3rd order} \\
& - a_3\left[\begin{array}{l} v_{i1}^3\left(\tfrac{3}{4}\sin\omega_1 t - \tfrac{1}{4}\sin(3\omega_1 t)\right) + v_{i2}^3\left(\tfrac{3}{4}\sin(\omega_2 t) - \tfrac{1}{4}\sin(3\omega_2 t)\right) \\ + 3v_{i2}^2 v_{i2}\left(\tfrac{1}{2} - \tfrac{1}{2}\cos(2\omega_1 t)\right)\sin(\omega_2 t) + 3v_{i1}v_{i2}^2\sin(\omega_1 t)\left(\tfrac{1}{2} - \tfrac{1}{2}\cos(2\omega_2 t)\right) \end{array}\right]
\end{aligned}$$

Figure 3.18: The Result of Distorted Amplification of a Two-Tone Input Signal

The second-order terms alone are shown in Figure 3.19. They consist of three components. First, there's a DC term and a double-frequency term from the first input tone, just like the single-frequency case. Input tone 2 similarly contributes its own single-tone distortion terms. The new and interesting part is the third block, composed of a term at the sum of the two input frequencies and another term at their difference frequency. Note that all the terms have an amplitude proportional to the square of the input, assuming the two tones to be of equal size.

$$-a_2\left[v_{i1}^2\left(\tfrac{1}{2}-\tfrac{1}{2}\cos(2\omega_1 t)\right)+v_{i2}^2\left(\tfrac{1}{2}-\tfrac{1}{2}\cos(2\omega_2 t)\right)+2v_{i1}v_{i2}\left(\frac{1}{2}\cos(\omega_1 t-\omega_2 t)-\frac{1}{2}\cos(\omega_1 t+\omega_2 t)\right)\right]$$

Single-tone, f_1 | Single-tone, f_2 | Sum and difference frequencies

Figure 3.19: Second-Order Two-Tone Distortion Terms

The resulting spectrum is depicted in Figure 3.20. So far we're still not much worse off than we were. Recall that the bandwidth of a typical 802.11 signal is around 16 MHz, so the biggest difference in frequency that can result is about 16 MHz: still close to DC from the point of view of an RF signal at 2.4 GHz. The sum frequency is still essentially twice the input and so easily filtered.

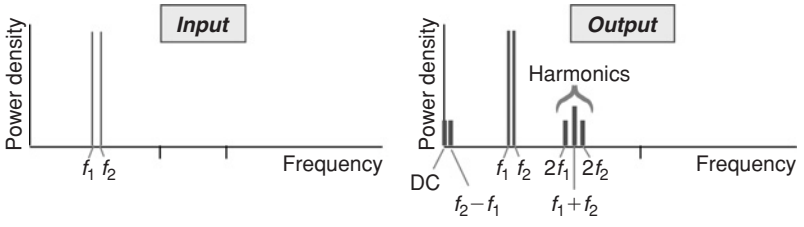


Figure 3.20: Spectrum of a Two-Tone Signal After Second-Order Distortion

This happy circumstance comes to an end when we allow for third-order two-tone distortion (Figure 3.21). Once again we see that each tone contributes the same individual distortion terms, at the fundamental and the third harmonic, that were observed in isolation. All the output terms are proportional to the cube of the input if the two tones are equal in magnitude.

However, the two-tone terms now contain both additional distortion of the fundamental frequencies and new tones at all the possible third-order combinations ($2f_1 + f_2$), ($2f_2 + f_1$), ($2f_1 - f_2$), and ($2f_2 - f_1$). This is significant because if the two input tones are closely spaced, the differences ($2f_1 - f_2$) and ($2f_2 - f_1$) *are close to the input tones and cannot be filtered.* This result is depicted graphically in Figure 3.22, where we show the whole two-tone output spectrum. Third-order distortion increases the apparent bandwidth of signals. Interferers grow wings in frequency that may overlap neighboring wanted signals. Transmitted signals extend into neighboring channels and bands.

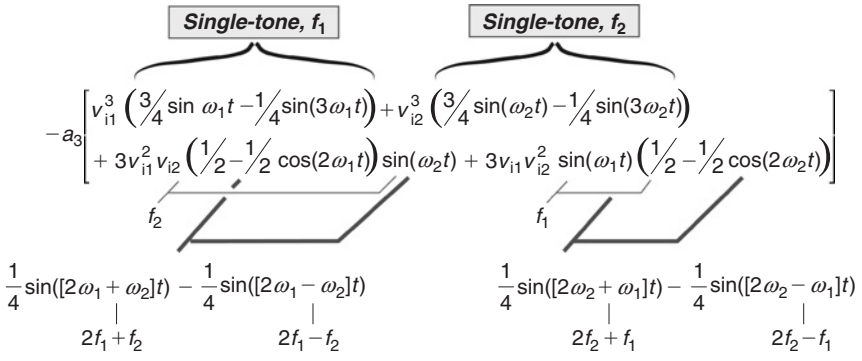


Figure 3.21: Third-Order Two-Tone Distortion Terms

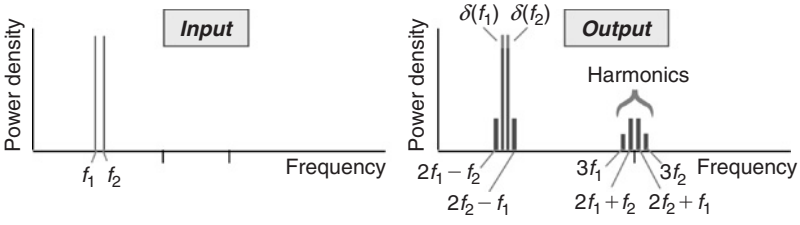


Figure 3.22: Spectrum of a Two-Tone Signal After Third-Order Distortion

We should note that the behavior cited here is characteristic of all even-and odd-order distortion products. Even-order (fourth, sixth, eighth, etc.) distortion terms always create DC or DC-like components and components at all even harmonics of the fundamental frequency; odd-order (fifth, seventh, ninth, etc.) distortion terms always create components at the fundamental frequency, odd harmonics, and, in the case of multiple tone, inputs at the possible difference terms near the fundamental. Some authors have pursued extensive investigations of the results of higher order polynomial distortion (which, when complex values are allowed, are known as Volterra series). However, in real devices it seems more sensible to treat high-order distortion as being due to an effectively abrupt encounter with the limitations of the device output: *clipping* of the input signal, which we discuss momentarily.

Given that it is not normally possible to specify the behavior of an amplifier in terms of a known polynomial, how shall we measure distortion? One method is to simply measure the power in one of the distortion products—an *intermodulation* (IM) product—at a particular input power and report this as the extent of the relevant distortion. Although this method is common and valuable, it requires that we measure the IM product at each power of interest. Recall that all second-order distortion products scale as the square of the input power and all third-order products scale as the cube. If the distortion products are truly of known order, it ought to be possible to measure the IM power at any input power and infer what the IM

would have been at a different input power. Because the order of the distortion associated with a given IM product may be inferred from its frequency (hoping that only the lowest order term contributes significantly), this prescription only requires that we adopt some convention about the input power to be used. The most common convention is to specify the input power at which the linearly amplified ideal output and the distortion product would be of equal magnitude: the intercept of lines drawn through the linear output and the IM power. This procedure is illustrated in Figure 3.23 for second-and third-order distortion.

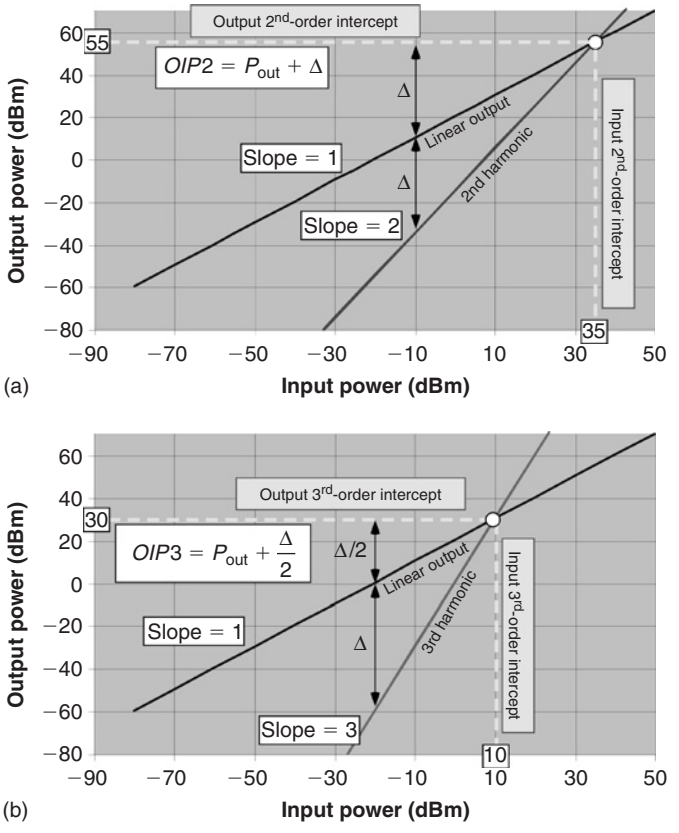


Figure 3.23: (a) Definition of the Second-Order Intercept (b) Definition of the Third-Order Intercept

It is important to note that intercepts are obtained by extrapolation, not direct measurement. Any device intended for use as a linear amplifier generally has such small distortion terms that long before we could turn the power up enough to make them equal to the undistorted fundamental, some other limitation of the device—typically its maximum output power—would be encountered, invalidating the idea of simple second- or third-order scaling of the distortion product.

The concept of the intercept point is only valid and useful in the region where the IM products scale appropriately: referring to Figure 3.23, where the IM power is in fact linear in decibels with input power, with slope of either 2 or 3 as appropriate. The definition of an intercept is often abused in literature and practice by measuring an IM product at a given power level and extrapolating from that single point to an intercept, without verifying the scaling of the distortion to be correct. When higher order distortion becomes significant, such extrapolations can be misleading: a fifth-order term may fortuitously cancel the third-order distortion over some limited range, resulting in an increase in the naively defined intercept point that misrepresents performance of the device except in the narrow power range in which it was characterized.

As a final caution, one must be alert to what intercept is being referred to, because many definitions are possible. First, referring to Figure 3.23, either the input or output intercept can be cited, the two being different by (approximately) the linear gain of the device. Common usage uses the output intercept for output-power–oriented applications such as a power amplifier and input intercept for an LNA. These are often abbreviated as *OIP2* and *IIP2* for second-order and *OIP3* and *IIP3* for third-order distortion, respectively. Second, there are several possible intercepts depending on which spurious product is selected for examination. One can define an intercept based on the harmonics that result when a single tone is placed at the input or based on the two-tone mixing products when two tones are present. Although there is no conceptual distinction between these methods and both provide sensible estimates of the nonlinearity of the amplifier in question, they do not produce the same numbers: IP2 from a second harmonic is 6 dB higher than IP2 obtained from a two-tone IM product. Similarly, IP3 from a third harmonic is 9 dB higher than IP3 from a $(2f_x \pm f_y)$ output power. (These distinctions may be derived from a thoughtful perusal of Figures 4.14 and 4.18; they are just the result of the differing coefficients multiplying the respective distortion terms.) Although third-order distortion is usually reported based on the $(2f_x \pm f_y)$, because of its practical importance and ease of measurement, conventions for reporting second-order distortion are less consistent. Finally, it is possible to define a two-tone intercept as occurring either at the point at which a distortion term is equal in power to the output power in one linear tone or equal to the total linear output power, 3 dB larger in the case of equal tones. The most common convention seems to be to define the intercept when one distortion tone is equal in power to one linear output,but this may not be universal. The distinction here is either (3/2 dB) for second-order or (3/3 = 1 dB) for third-order distortion and therefore is of modest practical importance.

Knowing the intercept for a given amplifier, we may obtain values of the IM products expected for any input power where scaling applies. This operation is depicted graphically in Figure 3.24. The operation is conveniently described in terms of the backoff from the intercept point: in Figure 3.24(a) the input power is backed off 40 dB from the intercept, and so the second-order IM product is decreased by twice as much—2(40) = 80 dB—relative to the power at

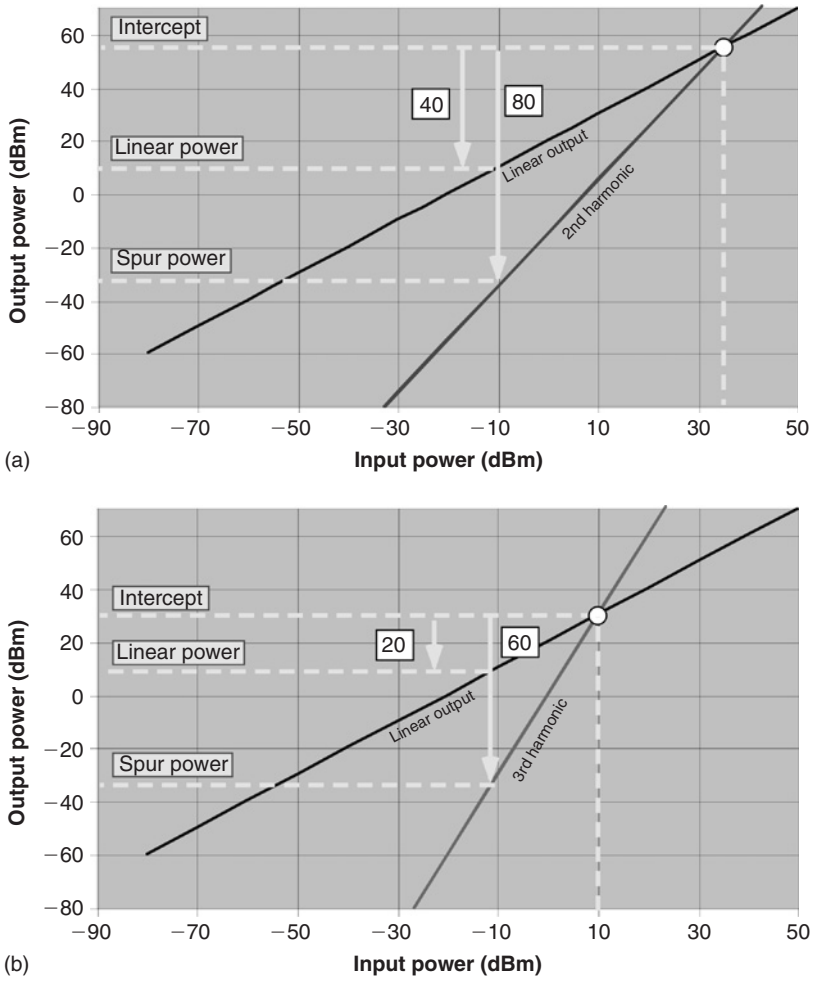


Figure 3.24: (a) Calculating Spurious Output Power Due to Second-Order Distortion Given OIP2 (b) Calculating Spurious Output Power Due to Third-Order Distortion Given OIP3

the intercept. Similarly, in Figure 3.24(b) 20 dB of backoff in linear power produces 60 dB reduction in third-order spurious output power. Spurious output powers are frequently measured by reference to the linear output power, the carrier in the case of a single-tone signal: thus, the spur is measured as decibels from the carrier, or *dBc*, because it is after all the size of the distortion with respect to the desired linear output that matters. Thus, at the indicated power in Figure 3.24(a) and (b), both the second- and third-order spurs are at −40 dBc.

Now that the reader is either thoroughly conversant with the terminology and concepts used in describing distortion or thoroughly confused, we (finally) present some typical values of

Table 3.5: Comparison of Third-Order Distortion for Devices/Technologies From Table 3.4

Technology	Gain (2 GHz)	NF (2 GHz)	OIP3 (2 GHz)	Reference
GaAs pHEMT	16 dB	1.2 dB	28 dBm	Agilent 9/03
GaAs MESFET	14 dB	3 dB	40 dBm	WJ Comm 9/03
SiGe HBT	17 dB	3 dB	14 dBm	Sirenza 9/03
Si CMOS, 0:8 mm	15 dB	3 dB	2 dBm	Kim MTT '98

third-order intercept, for the same devices we examined previously, in Table 3.5. Different technologies produce wildly different third-order intercepts for otherwise similar values of gain and noise figure. The reader should not conclude from this very simplified comparison that every CMOS amplifier suffers a 38-dB inferiority in OIP3 to every GaAs MESFET device, but it is valid to point out that compound-semiconductor devices in general are capable of higher output power and better linearity for otherwise similar performance because of their higher electron mobility and generally better resistance to electrical breakdown compared with silicon devices. These benefits must be measured against the expensive starting materials and relatively primitive and high-cost process technology characteristic of even mature GaAs fabrication. In most WLAN radios, the tremendous cost and integration benefits of silicon CMOS result in compound-semiconductor devices being relegated to the power amplifier and RF switches, if they are present at all.

The noise figure of a device represents some measure of the smallest signal it can usefully amplify. Similarly, the intercept represents a measure of the largest signal that can be linearly amplified. The two values may be combined to estimate the device's dynamic range, the difference between the smallest and largest useful signals. The exact definition of dynamic range can vary depending on what is considered useful. A fairly general but rather demanding approach is to define the *spur-free dynamic range* as the range of input power in which the amplified signal is both greater than the amplifier noise and no distortion is detectable (i.e., the spurious output is less than the amplifier noise floor). For a third-order spur to be undetectable, we must be backed off by one-third of the difference between the intercept and the noise floor (Figure 3.24(b)): we find the awkward formula

$$SFDR = \frac{2}{3}\{OIP3 - [-174\,\text{dBm} + BW + NF]\} \tag{3.7}$$

where all quantities are measured in dB or dBm as appropriate. The spur-free dynamic range (SFDR) depends on the bandwidth of interest and cannot be defined independent of the application.

We have so far concentrated all our efforts in describing distortion of signals composed of at most two input tones. Real digital signals are the result of an effectively random sequence of bits defining a time series of single-carrier modulation symbols (802.11 classic or b, 802.15.1,

or 802.15.3) or the subcarriers of an OFDM symbol (802.11a,g). The spectrum of such a signal is generally rather broadly distributed, containing power in a wide range of frequencies, as depicted for example in Figure 2-20. What happens to such a more-complex signal in the presence of nonlinear distortion?

The effect of third-order distortion on a realistic digital signal spectrum is qualitatively depicted in Figure 3.25. Replicas of the ideal spectrum—wings—appear on the sides of the spectrum, extending roughly as far as the original spectrum width in each direction. (Just rewrite the expression: $2f_2 - f_1 = f_2 + (f_2 - f_1)$. Images of the spectrum at the third harmonic and sum tones are also present but are generally easily filtered. The distorted spectrum is wider in frequency than the original spectrum and may extend into neighboring channels or bands.

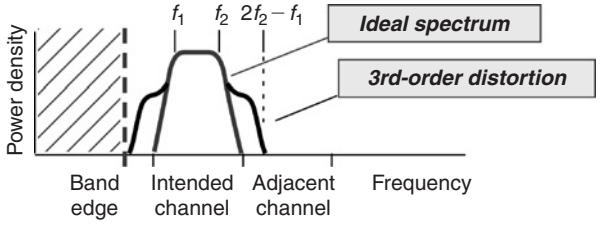


Figure 3.25: Effect of Third-Order Distortion on the Spectrum of a Typical Digital Signal

Power that extends into a neighboring channel is known, sensibly enough, as adjacent-channel power, and the ratio between the power in the intended channel and the adjacent one is known as the *adjacent-channel power ratio* (ACPR). Such spreading is important, even if the ACPR is quite small, because of the near-far problem: a nearby powerful radiator on one channel may, if its spectrum is broadened, block a distant wanted radio on a neighboring channel. Standards generally contain provisions to limit such unwanted interference in the form of the spectral masks defining the allowed shape of the emitted signal. For example, a comparison of the 802.11a spectral mask to the spectrum of a distorted signal (Figure 3.26) may clarify the origin of the heretofore mysterious mask shape: it is designed with the expectation of some third-order (and higher) distortion in the signal, providing a compromise between minimizing adjacent-channel interference and cost of the resulting spec-compliant transmitter.

It is not a trivial matter to evaluate quantitatively how a given degree of distortion will affect a complex digital signal. Fortunately, a rough but useful estimate of the resulting ACPR can be obtained directly from the two-tone intercept: Pedro and de Carvalho showed that for a fairly general digital-like signal composed of a number of equally spaced randomly phased tones, the ACPR, measured as the ratio of the power in one of the third-order "wings" of the

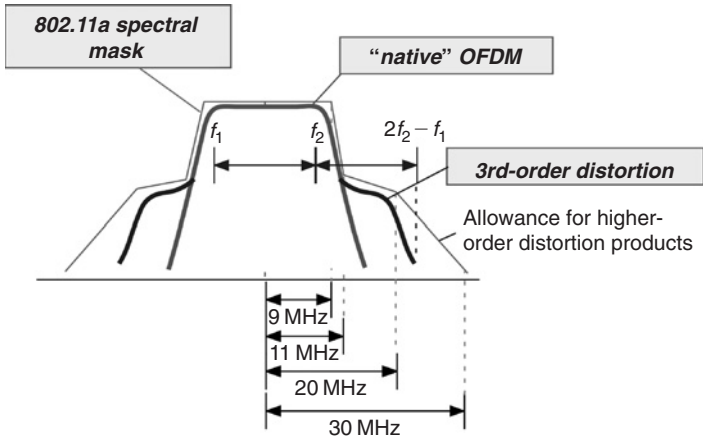


Figure 3.26: Third-Order Distortion vs. the 802.11a Spectral Mask

spectrum to the power in the linearly amplified original spectrum, is roughly the same as the ratio of one of the two-tone difference spurs to one of the corresponding linear tones, to within a decibel or two. It is this partially accidental agreement that accounts for the popularity and utility of the simple two-tone measurement of IP3 (Figure 3.27).

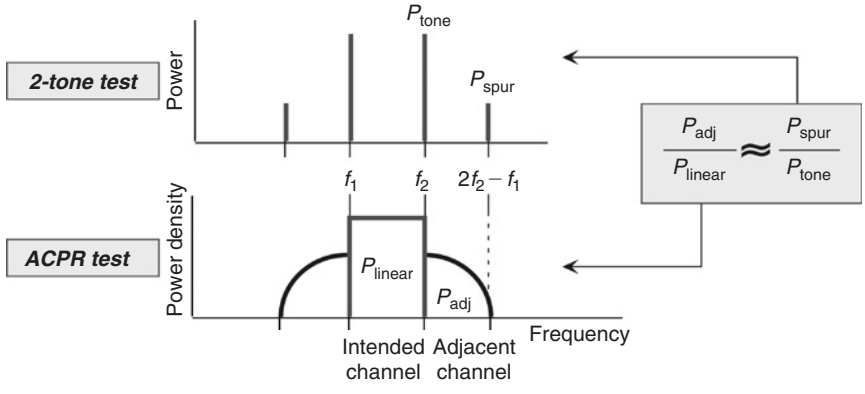


Figure 3.27: Correspondence Between ACPR of a Distorted Digital Signal and Spurious Power From a Two-Tone Test

A more subtle consequence of third-order distortion, in this case within the receiver, is of some consequence in 802.11 WLAN applications, where the three equally spaced nonoverlapping channels (1, 6, and 11) available in the United States allow for the IM product of two interferers to lie within a wanted channel. An example is shown in Figure 3.28: two nearby transmitters on channels 6 and 11 transmit simultaneously with a wanted signal on channel 1.

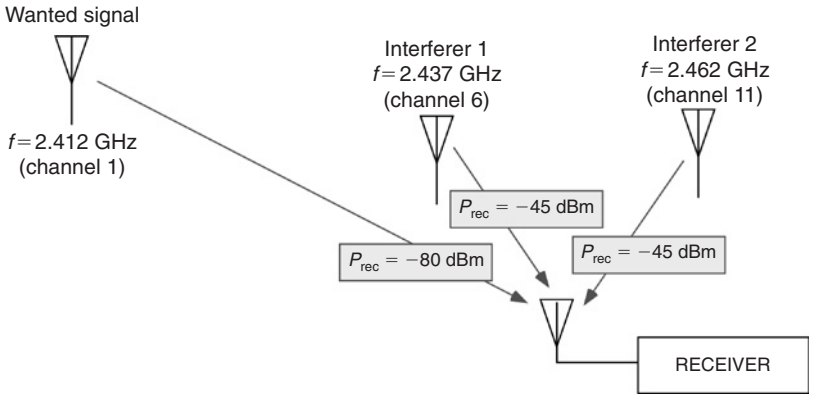


Figure 3.28: Simultaneous Transmitters on 802.11 U.S. ISM Channels

The receiver contains a first filter that removes signals outside the ISM band (Figure 3.29), but within this band all three signals are amplified by the LNA and mixed in the first mixer. It is only after mixing that a filter is used to remove channels 6 and 11 from the total. (Note that we have symbolically indicated a radio in a "high-gain" state by switching the LNA into the chain; a real radio would likely have variable-gain amplifiers or multiple stages of amplification with switching.)

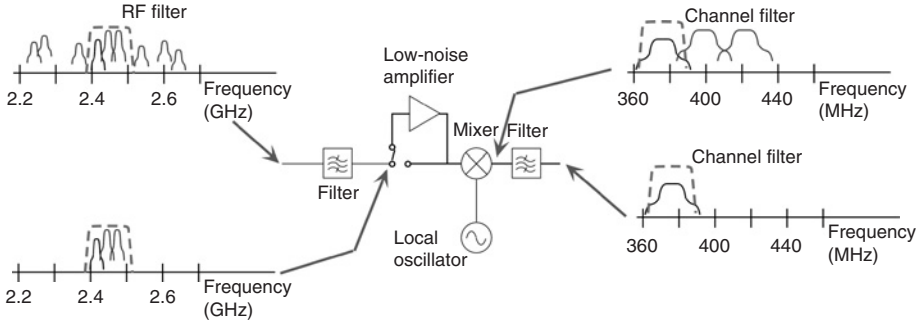


Figure 3.29: Progressive Filtering of Band and Channel Within a Superheterodyne Receiver

Any third-order distortion of the (channel 6 + channel 11) signal present in the LNA or mixer will result in an IM component at (2.437 – (2.462 – 2.437)) = 2.412 GHz: right on top of the wanted channel 1 (Figure 3.30). We can estimate the consequent requirements placed on the radio: we want the IM product to be less than or comparable with the noise, which for a QPSK signal is around 10 dB less than the signal for reliable reception. An IM product at −95 dBm seems unobjectionable. This is 50 dBc from the interfering tones at −45 dBm, so the input intercept required to produce this result is (−45 + (50/2)) = −20 dBm. This estimate is in reasonable agreement with actual chipset IIP3, which varies from around −18 to −8 dBm (see Table 3.10 later in this chapter).

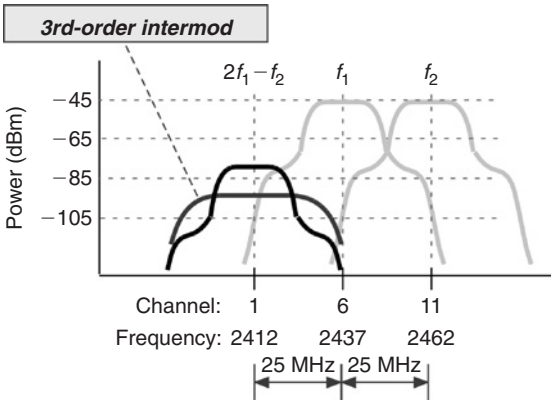


Figure 3.30: IM Product of Channels 6 and 11 at Channel 1

Obviously, the same problem could arise for interferers on channels 1 and 6 and a wanted channel 11. Because only three nonoverlapping ISM channels are available in the United States, this unfortunate occurrence must be regarded as reasonably probable and taken into account in radio design. An 802.11a WLAN deployment has a great deal more flexibility in frequency planning because of the expanded number of available channels; although equally spaced interferers remain possible and should be accounted for in design, they are less likely to be encountered in the field.

The astute reader, in perusing Figure 3.21, may have noted with some concern that both the single-tone and two-tone terms of a third-order–distorted signal produce components at the fundamental. For the rest of us, the calculation is reproduced with this fact highlighted in Figure 3.31: the change in the fundamental power due to distortion is three times larger in

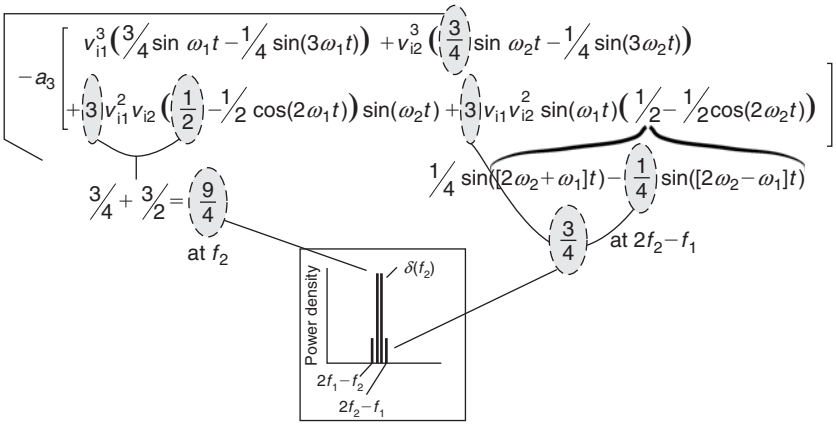


Figure 3.31: Repetition of the Calculation of a Third-Order Two-Tone Distortion; the Distortion Product at f_2 Is 3 Times Larger in Magnitude Than That at $(2f_2 - f_1)$

amplitude than that at the spur frequency. Third-order distortion has a much larger absolute effect at the fundamental frequency than at the spurs, even though the effect is less noticeable on, for example, a spectrum analyzer because of the presence of the large linearly amplified fundamental. Pedro and de Carvalho showed that for a representative multitone signal, the corresponding result is a discrepancy of about 12 dB between the large effect on the linear signal and the smaller adjacent-channel power. In other words, the presence of third-order distortion in the signal chain results in—dare we say it?—nonlinear distortion of the input signal and, consequently, a difference between the signal that was to be transmitted or received and is. How do we measure this difference and when do we care that it exists?

Distortion of the received signal is quantified as error vector magnitude (EVM). EVM measures the difference between the transmitted or received constellation point on the phase-amplitude plane and the intended symbol, normalized by the amplitude corresponding to the average transmitted power (Figure 3.32). The measure is averaged over a pseudo-random sequence of input data to obtain a statistically valid estimate of the accuracy of the system.

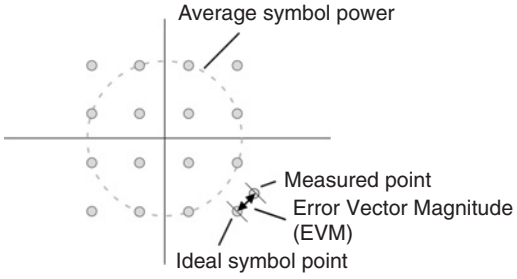


Figure 3.32: Definition of EVM for a 16QAM Signal

Requirements on EVM for the 802.11a signal versus the data rate are summarized in Table 3.6. For higher data rates, EVM requirements are more demanding; this should not be surprising once we recall that the higher data rates use QAM constellations with more points and thus

Table 3.6: EVM for 802.11a

Data Rate (Mbps)	<EVM> (dB)
6	−5
9	−8
12	−10
18	−13
24	−16
36	−19
48	−22
54	−25

less space between the points (see Table 2.2). How much third-order distortion is allowed by these specifications? For the 54 megabits/second (Mbps) case, we need distortion less than −25 dB. Remember that third-order distortion of the fundamental is actually about 9 dB larger than the adjacent-channel spurious output (Figure 3.21). To get third-order distortion in the same channel (*cochannel* distortion) to be down by 25 dB, we require the adjacent-channel distortion to be at least 34 dB less than the signal. To have a two-tone distortion 34 dB below the signal (Figure 3.23(b), D), we need the intercept to be at least 17 dB above the signal power (i.e., $\Delta = 2$).

On the receive side, it's not hard to provide a sufficiently large intercept: the maximum input power we expect to see is around −30 to −40 dBm, and the input third-order intercept IIP3 is typically around 10 dBm even in the high-gain state (which we would not be using to receive such a large signal). On transmit it's another story. A typical amplifier might achieve an output intercept point OIP3 that's 10 dB higher than its nominal output power. To keep EVM under control, we'd need to reduce the output power—back off—by another 8 dB. Every dB of backoff means less output power for the same DC power. Achieving EVM specifications is often the limitation on output power for high–data rate OFDM signals.

So far we've dealt only with low-order polynomial distortions. As output power is increased, every amplifier reaches its limit: the input power at which output power can no longer increase. The maximum output power is known as the saturated output power, P_{sat}. Any signal larger than the limit is *clipped*: the output power becomes substantially independent of the input power (Figure 3.33). Folks often use the power at which the gain has decreased by 1 dB, P1dB, as a somewhat less disastrously distorted limit on the output. In an ideal amplifier that is perfectly linear until it clips, the input 1-dB-compressed power would be 1 dB greater than the input power for saturation; however, real amplifiers aren't quite so abrupt, and P1 dB is a decent substitute for P_{sat} for many purposes.

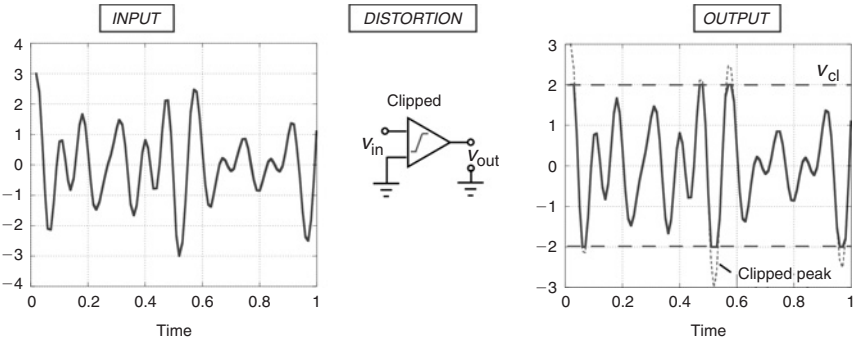


Figure 3.33: Clipping Distortion

Clipping can be analyzed by adding higher order odd distortion terms (fifth, seventh, etc.) to the polynomial transfer function (Figure 3.13). However, for many purposes, Cripps showed

that it is simpler and more profitable to use an ideal polynomial transfer function, as shown in Figure 3.33: the amplifier is either linear or fixed in output power (see section 3.6). In either case the treatment is complex and unenlightening for our purposes; the net result, in addition to the obvious distortion of the desired signal, is a spectrum with wings that extend farther and fall off more slowly than those due to third-order distortion.

Dealing with clipping appears at first sight to be quite simple: turn the power down until there isn't any. However, recall that many signals have a peak-to-average ratio P/A > 1. (This quantity is also frequently referred to as the *crest factor*.) The largest value that the signal can take is much larger than the average power for a random input stream. To avoid clipping, it's not enough to keep the power at the saturated power; we need to back off by (roughly) the peak-to-average ratio to keep from clipping. How much is that?

The peak-to-average ratio depends not just on the modulation scheme but also on the path taken between constellation points. Some modulations, such as Gaussian minimum-shift keying, are chosen specifically because the path between constellation points is along a circle of constant power (Figure 3.34), though at the sacrifice of some bandwidth. QPSK is a bandwidth-efficient modulation, but the path between points passes close to the origin (depending on the previous symbols), so the average power is less than the constellation point power.

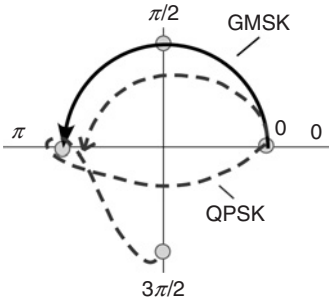


Figure 3.34: Trajectories in Phase-Amplitude Space for Two Modulations

The actual peak-to-average ratio for a particular modulation scheme depends on the exact approach to making the transitions between symbols and is generally obtained numerically. An example for a QPSK system is depicted in Figure 3.35: the calculated (P/A) = 2.7:1 or about 4.3 dB. An alternative approach, *offset QPSK*, is sometimes used in which the I-branch transition and the Q-branch transition are offset by 45 degrees; in this case, the trajectories never pass through the origin and the (P/A) is somewhat reduced.

For a complex signal like OFDM, the peak-to-average ratio is a statistical quantity: what is of interest is how often a given peak power is encountered in a random data stream, because if a certain level is sufficiently improbable, it is no longer important. The likelihood of a given power level can be succinctly displayed using a cumulative distribution function,

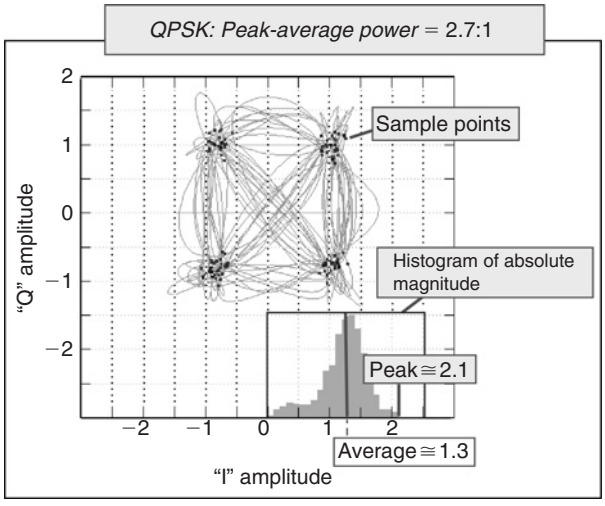
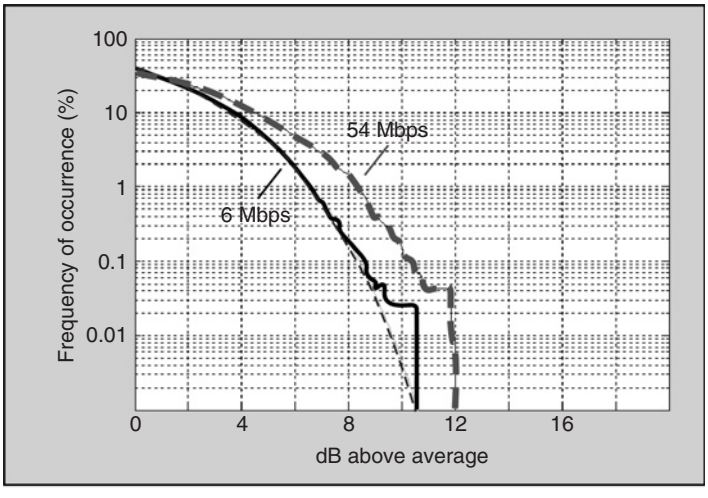


Figure 3.35: Numerical Simulation of Phase-Amplitude Trajectory of QPSK Signal for a Pseudo-Random Input (Simulation by Walter Strifler, Image Courtesy of WJ Communications)



Calculations from agilent signal studio 802.11

Figure 3.36: Combined Cumulative Distribution Functions for 802.11 OFDM Signals at Low and High Data Rate

the combined likelihood that the power in a sequence is less than a given power level. We encountered some of these before in Figure 2.20; they are reproduced below as Figure 3.36 for convenience.

If we were to choose a frequency of 0.1% (10^{-3}) as our threshold, we see that a 6 Mbps signal has a (P/A) ≈ 8.3 dB; the 54 Mbps signal (P/A) is around 10 dB. To transmit an OFDM

signal that is clipped less than 0.1% of the time, it appears we would have to back off about 10 dB—that is, we can only transmit 10% of the power that our output amplifier can deliver as useful OFDM power. Fortunately, things aren't quite this bad. In practice, around 5–8 dB of backoff is sufficient to achieve compliant output spectra. For high data rates (>32 Mbps), it is usually the EVM specification that limits output power rather than the spurious emission due to distortion or clipping.

Let us pause for a moment to briefly summarize this rather long subsection. Amplifiers must provide enough gain on the receive side to deliver a signal the ADC can read and enough power on the transmit side to reach the receiver. The noise in the receiver is dominated by the LNA noise if the gain of the LNA is reasonably large. The noise floor sets the lowest signals that can be received; the upper limit is set by distortion. Typically, we are most concerned with third-order distortion, because the resulting spurious outputs are in and near the original frequency and cannot be filtered. Third-order distortion can be avoided by using more linear amplifiers (which usually costs either current or dollars or both) or by reducing the signal amplitude. Complex signals also have a significant (P/A), and average power must be backed off enough to avoid undue disturbances of the signal resulting from clipping distortion during the relatively rare excursions to high power.

3.2.3 Mixers and Frequency Conversion

Frequency conversion plays a key role in analog radios, and mixers are the means by which this conversion is accomplished. Several are used in a typical radio (Figure 3.37).

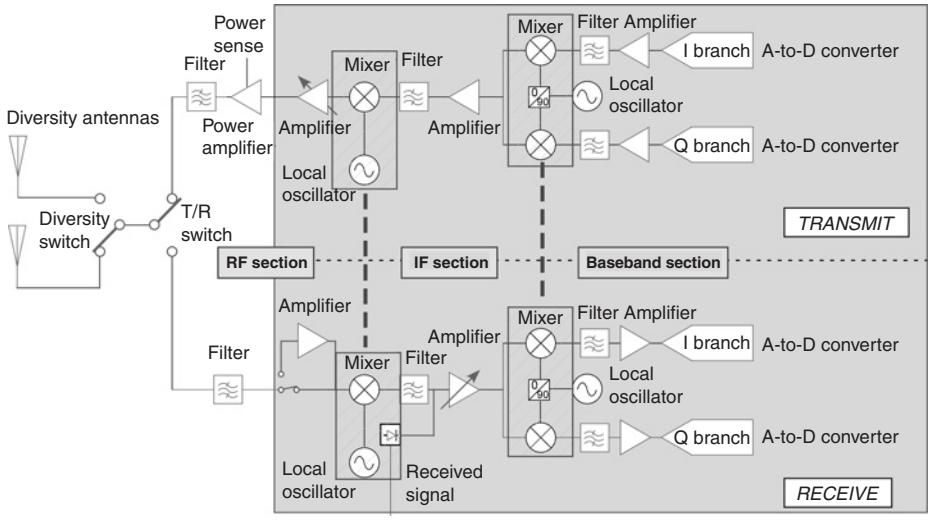


Figure 3.37: Block Diagram of aWLAN Radio Emphasizing Mixer/LO Functions

A mixer combines an incoming transmitted or received signal with a nearly pure sinusoid from a LO to create new frequencies. On the transmit side, these new frequencies are typically much higher than the incoming signal: the mixer is used for *up-conversion*. On the receive side, the output frequency is reduced: *down-conversion*. The efficiency of conversion is normally described as the *conversion loss*: the ratio of the output power at the converted frequency to the input power at the original frequency. (This can be a conversion gain in an active mixer such as the Gilbert cell, discussed below.) Mixers are always nonlinear devices, but good mixers are highly linear. This puzzling apparent contradiction will hopefully be resolved shortly as we examine mixer operation.

How does a mixer work? Recall from Chapter 1 (Figure 1.14 to be exact) that when we modulate a signal, new frequencies appear in its spectrum. A mixer is a modulator in disguise: it modulates a sinusoidal input to produce new frequencies, one of which is typically selected by filtering the output (Figure 3.38).

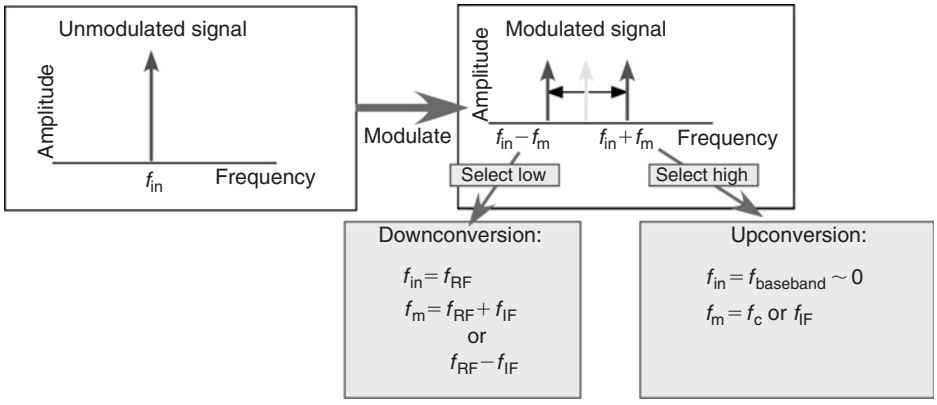


Figure 3.38: Modulation as Frequency Conversion

How are we to produce this modulated signal? Recall that the simplest modulation scheme is on–off keying (see Figure 1.12); similarly, the simplest mixer is just a switch that turns the incoming signal on or off. Such a simple mixer is depicted schematically in Figure 3.39; its use as an up-converter and down-converter is shown in Figure 3.40. The input signal is effectively multiplied by a square wave whose value varies between 1 and 0. Because the square wave and the input are at differing frequencies, their relative phase varies with time,

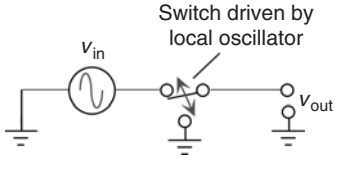


Figure 3.39: Simple Switch Mixer

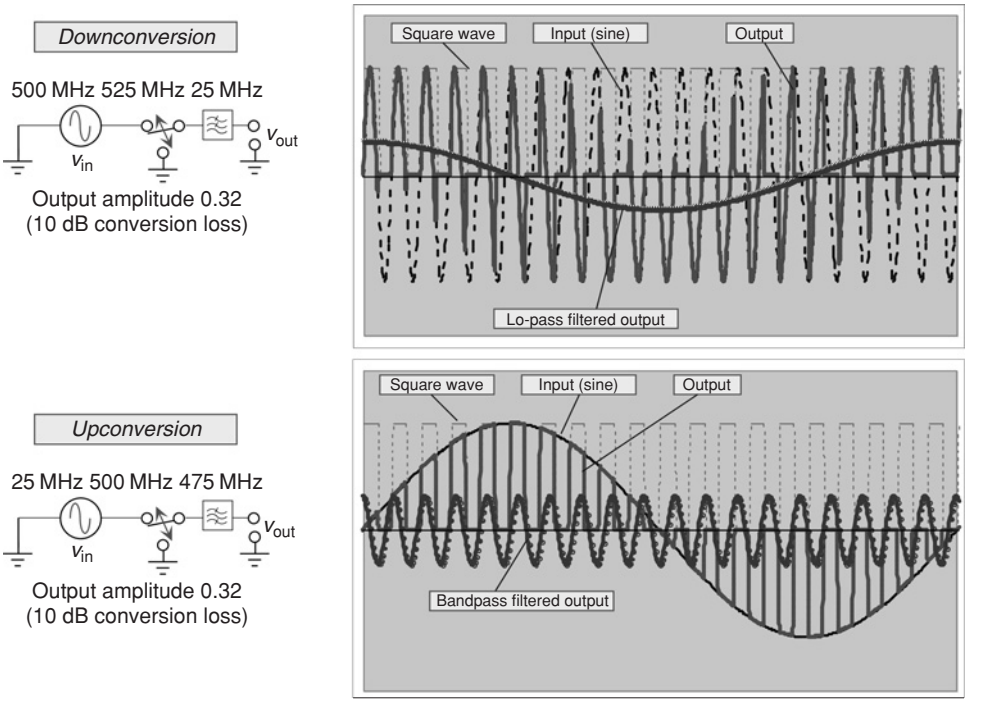


Figure 3.40: Input, Control, and Output Signals for Switch Mixer

producing a time-varying output with components at the sum and difference frequencies. It is this phase variation that is key to the operation of the mixer as a converter.

In addition to its manifest simplicity, such a mixer has significant advantages. Our ideal switch is either on or off and nothing else. When it is off, there is no output, no matter what the input. When it is on, the output is equal to the input. Thus, even though the operation of the device is highly nonlinear in the sense that the two possible states have very different output–input relationships, within each state the output is a perfect replica of the input (admittedly with a gain of 0 in the "off" state). Once the sharp transitions are removed by filtering, the output signal is a faithful replica of (part of) the input signal: there is no distortion. The mixer is a linear nonlinear device. Further, a reasonable approximation of the switch mixer can be implemented with any active device (FET, BJT, etc.), as long as the LO input is large enough to switch the device rapidly from its "off" state to a saturated "on" state.

The price for this simplicity is apparent in Figure 3.40. The switch is off half the time, so half the input signal power is wasted: it is immediately apparent that the conversion loss must be at least 3 dB. The actual situation is worse because many harmonics are present in the output signal in addition to the desired frequency; when the output is filtered to obtain the wanted signal only, 10 dB of conversion loss has resulted.

Another limitation of the simple switched mixer is that it does not reject *images* of the input or suppress images on the output. The down-converter will also convert a 550-MHz input to 25 MHz; the up-converter will also produce a 525-MHz output signal. A simple mixer of this sort must be combined with filtering to remove undesired image products.

An improved mixer design can be obtained using two switches, and *baluns* to convert the input signal from a *single-ended* voltage referenced to ground to a differential or *balanced*[1] signal symmetrically placed around the ground potential (Figure 3.41).

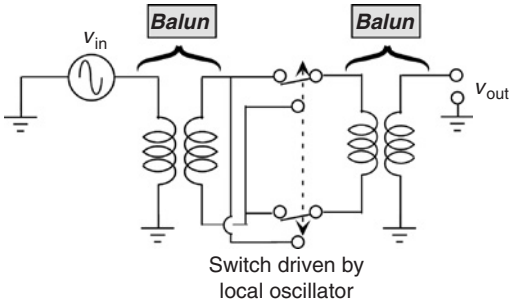


Figure 3.41: Balanced Switch Mixer

This mixer effectively multiplies the input signal by a square wave with values (+1, −1) instead of (1, 0) (Figure 3.42). Thus, the input is always connected to the output, albeit with alternating polarity, and the conversion loss is reduced. Because of the balanced design, no power is wasted in a DC offset, so the realized conversion loss is only about 4 dB: a 6-dB improvement. As long as the switches are ideal, the mixer is also linear in the sense described above. The trade-off is that complexity has increased considerably. An additional switching element has been added, and (if the input signal isn't already differential) two baluns are required to convert the single-ended signal to differential and back again. In Figure 3.41, the baluns are shown as transformers. Broadband baluns of this type for frequencies up to a few gigahertz can be constructed by wrapping a few turns of trifilar (three-filament) wire around a donut-shaped ferrite core a few millimeters in diameter, at a cost of some tens of cents; such devices are obviously huge relative to a chip and must be placed separately on the circuit board. Narrowband baluns can also be constructed using lines on the circuit board as transmission lines, with one branch delayed to produce a 180-degree phase shift, again a large structure requiring space on a board. Active baluns can be fabricated using an amplifier but add current consumption and distortion. Baluns are an inconvenient but often inevitable fact of life for radio designers. The balanced switch mixer is also subject to the same limitations with regard to image rejection noted for the simple (unbalanced) switch mixer.

[1] Unfortunately it is also common practice in the microwave world to use the term *balanced* to describe aquite distinct trick in which two signals are offset from each other by 90 rather than 180 degrees, done to produce an improved input and output match.

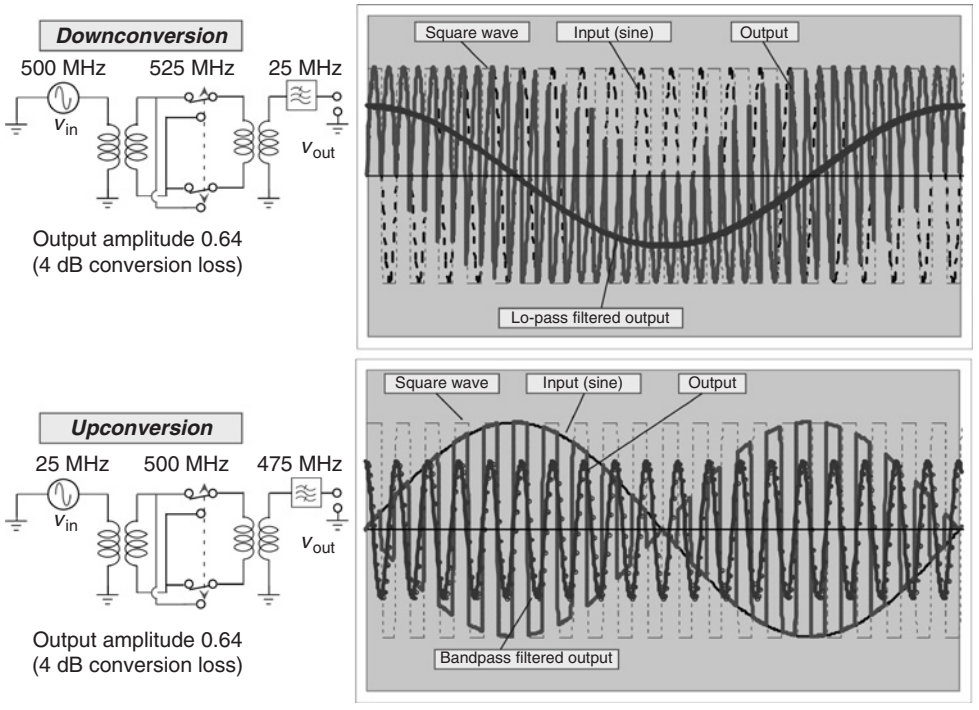


Figure 3.42: Balanced Switch Mixer Operation

Both mixers we describe here are known (despite the use of apparently active switch elements) as *passive mixers*. The nomenclature emphasizes the fact that the output is at best equal to the input; no amplification accompanies the mixing operation. We have already seen that balanced mixers have lower conversion loss. This fact is doubly important because the noise figure of a good passive mixer is essentially equal to the conversion loss (plus a slight correction for excess noise in the devices), so a balanced mixer requires less amplification before the mixer to get good overall noise performance.

To achieve good linearity between the input and the converted output, it is necessary that the transitions between switch states occur as rapidly as possible. This is because real switching devices (FETs, BJTs, or diodes) are highly nonlinear during the transition. For example, a FET device typically has high distortion near the pinch-off voltage where the current goes to zero; if the mixer spends a lot of time turning off, considerable distortion in the filtered output signal will result. A linear mixer requires that the switching devices be fast compared with the signal frequency, and the best performance is usually obtained at a LO voltage significantly larger than the minimum necessary to switch states, so that the time spent in the transition is only a small fraction of the LO cycle.

Another important characteristic of mixers is *isolation*, referring not to the existential loneliness of an unwanted image frequency but to the extent to which the output voltage is isolated from the input voltages. Only the converted products should appear at the output: the LO and input signals should not. The LO is usually the biggest problem because it is much higher in power than the other inputs. High LO powers may be used, as noted above, to ensure fast switching of the active devices and thus good linearity. Leakage of the LO into the transmitted signal during up-conversion results in an undesired emitted frequency, or for direct upconversion a peak in the center of the transmitted spectrum that may confuse the demodulation process; allowed leakage is generally specified in the standards. LO leakage into the IF stages on the down-conversion side may be further converted by IF nonlinearities into a DC offset. Balanced mixers generally provide better isolation than unbalanced mixers; for example, a small DC offset in the input signal in Figure 3.40 would be converted by the switch into an output signal at the LO frequency, whereas the same offset in Figure 3.41 would be rejected by the balun and have no effect. Isolation is normally limited by more subtle effects such as parasitic capacitive or inductive coupling within the device; balanced signals are also beneficial in this case, because they contain no net voltage to ground and are thus less likely to couple over long distances.

Unwanted images are not the only undesirable outputs of a mixer. An ideal switch mixer effectively multiplies the input signal by a square wave, which contains not only the fundamental frequency but all odd harmonics (Figure 3.43). The amplitude of higher harmonics decreases rather slowly (as $1/n$ for the nth harmonic). Each of these harmonics multiplies the input signal to produce spurious outputs (spurs) at, for example $3f_{LO} + f_{IF}$. Any nonlinearities present in the device will produce mixing of these spurs with whatever other frequencies are lying around to produce essentially every imaginable integer combination of frequencies: $3f_{LO} - 2f_{IF}$, $2f_{LO} - 3f_{IF}$, and so on. Spurs can sneak through filters and accidentally lie on or near wanted signals and cause all sorts of mischief. Balanced mixers are usually better than unbalanced mixers, but all are subject to spurious outputs. The intensity of all the possible low-numbered spurs of a mixer is often measured or simulated and summarized in a spur table for a given set of input conditions.

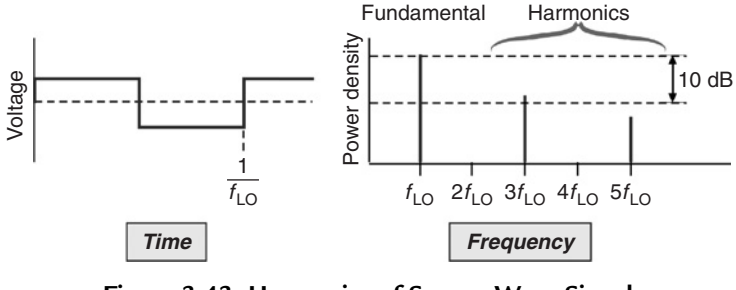


Figure 3.43: Harmonics of Square Wave Signal

We've mentioned a couple of times that images of the wanted frequency are a problem. A simple mixer with a LO frequency of 2.5 GHz will convert a signal at 2.43 GHz to a 700-MHz IF, but a signal at 2.57 GHz will also be converted to the same IF (Figure 3.44) and interfere with the wanted signal. What measures can be taken to avoid this problem?

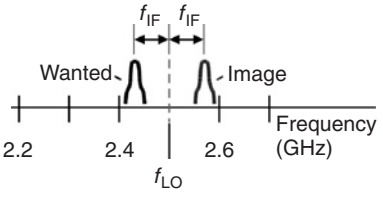


Figure 3.44: Image Frequency

One obvious approach is to filter out the offending signal before it reaches the mixer (Figure 3.45). Remember the near–far problem: an interfering signal could be much larger than the wanted signal, so an image filter has to do a very good job of filtering the image. Obviously, the larger the IF frequency, the greater the separation between image and wanted signal and the easier it is to filter the image. On the other hand, use of a very high IF sacrifices some of the benefits of converting in the first place: gain is harder to come by and channel filtering is rendered more difficult. The choice of the IF (*frequency planning*) is an important aspect of radio design. In the limit of an NZIF, radio filtering is impossible, because the image is immediately adjacent to the wanted signal.

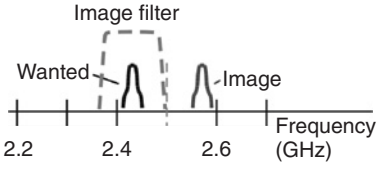


Figure 3.45: Image Filter

A different and much more subtle approach is to construct a mixer that converts only the wanted signal and rejects the image frequency: an *image-reject mixer* (IRM). A conceptually identical trick is to produce only the wanted sideband on up-conversion: a *single-sideband* mixer. IRMs are somewhat more complex than conventional mixers but are key enablers for NZIF architecture and are often useful in other radio designs. To explain how they work, it is helpful to expand upon the idea of complex exponential signals introduced in Chapter 1.

We can write a sinusoidal input voltage as the sum of two exponentials:

$$\cos(\omega t) = \frac{e^{i\omega t} + e^{-i\omega t}}{2}; \quad \sin(\omega t) = \frac{e^{i\omega t} - e^{-i\omega t}}{2i} \tag{3.8}$$

We can construct a graphic depiction of these operations if we admit the concept of a negative frequency (Figure 3.46). Exponentials of positive frequency are depicted as arrows on the frequency axis to the right of zero and are to be imagined as spinning counterclockwise around the axis at the frequency v. Negative-frequency terms lie on the negative axis to the left of $f = 0$ and spin clockwise in the complex plane, perpendicular to the frequency axis. A phase shift a displaces positive frequencies and negative frequencies in opposite directions. (We apologize for the rotation of the conventional complex plane to place the real axis vertically, but this is convenient as real signals then point up or down.) The justification of equation [4.7] can be easily constructed: for a cosine, the positive-frequency and negative-frequency arrows both start vertical at time $t = 0$ and counterrotate so that their sum has no imaginary part. A similar construction takes place for the sine, which, however, is rotated onto the imaginary axis.

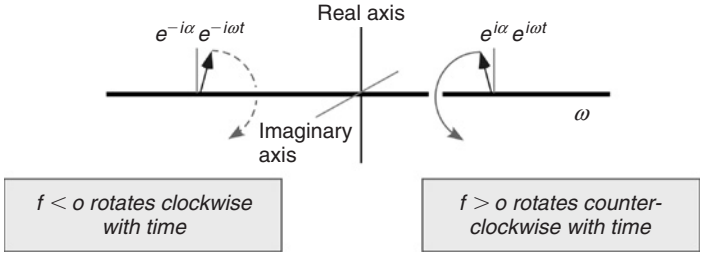


Figure 3.46: Exponentials in Complex–Frequency Space

Because mixing is a multiplicative process, it is necessary to go a bit further and figure out what happens to our picture when we multiply two signals together. Mathematically, this is very simple (which is the reason to use exponential signals)—the product of two exponentials is a third exponential of the sum of the arguments. In the case of harmonic signals, the frequencies add

$$e^{i\omega_1 t} \cdot e^{i\omega_2 t} = e^{i(\omega_1+\omega_2)t} \tag{3.9}$$

In our spatial picture, the product of two arrows is another arrow whose angle (phase) is the sum of the constituent phases and whose frequency is the sum of the constituent frequencies. Assume for the present we are interested in a down-conversion application and ignore the terms with the same sign of frequency, which when summed will produce high-frequency components to be filtered out. The product of a positive-frequency term from, for example, a signal at $(\omega + \delta)$ and a negative-frequency term from a LO at frequency $(-\omega)$ will be at positive frequency (Figure 3.47, top set of dashed lines). A negative-frequency signal term will combine with a positive-frequency LO of lesser absolute frequency to produce a negative-frequency term (bottom set of dashed lines). Here the negative-frequency signal is shown as phase shifted by 180 degrees, and the product term inherits this phase shift.

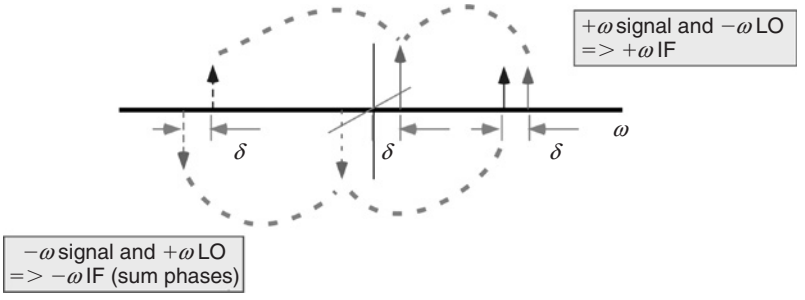


Figure 3.47: The (Low-Frequency) Products of Cosine and Sine

With these preliminaries in mind, let us now consider the problem at hand: we wish to down-convert a wanted signal at some frequency $f = f_{LO} + f_{IF}$ but reject the image at $f_{LO} - f_{IF}$. First, let's mix the signal (regarded as a pair of cosines) with a cosine LO—an in-phase conversion. The result is shown in Figure 3.48. At positive frequency we find both the positive-frequency component of the wanted signal and the negative-frequency component of the image (dashed lines); the negative-frequency terms arise from the negative-frequency wanted signal and the positive-frequency image.

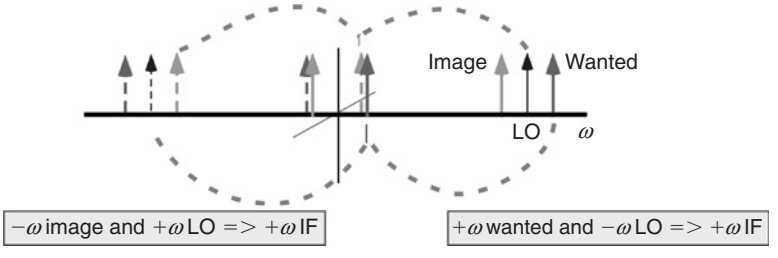


Figure 3.48: I-Channel Mix of Image and Wanted Input

What happens if we also mix the signal with a sine—the *quadrature* channel? As shown in Figure 3.49, the components at each frequency acquire a different phase depending on whether they originate from the positive-or negative-frequency component of the original signal (as

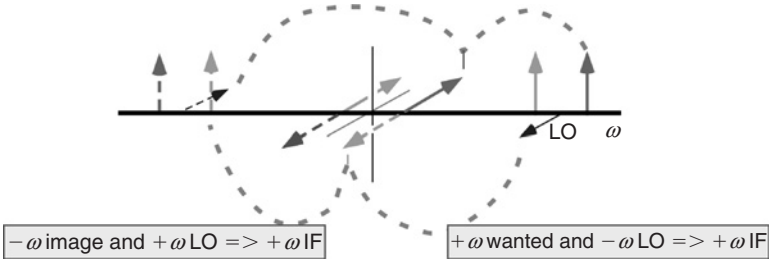


Figure 3.49: Q-Channel Mix of Image and Wanted Input

indicated by the dashed lines). Two constituents that end up at the same frequency are not the same in phase: it is this fact that enables us to separate out image and wanted signals.

To exploit the phase difference between the components of the signal, we apply a 90-degree phase shift to one of the branches, for example, the Q branch. This can be done using an R-C network or by simply delaying the branch by a quarter-cycle with a bit of extra transmission line. Recall that a positive phase shift rotates vectors counterclockwise on the positive frequency axis and clockwise on the negative frequency axis. The resulting signal is shown in Figure 3.50, where the rotation operation is explicitly shown on only the negative frequency components for clarity.

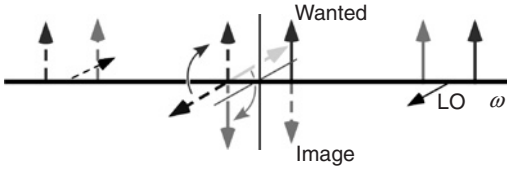


Figure 3.50: Q Branch After Phase Shift

The components deriving from the wanted signal are positive at positive frequency and negative at negative frequency. The components deriving from the image are of opposite sign. If we now add this signal to the I branch signal (Figure 3.48), the components from the wanted signal will add, whereas the components which came from the image will subtract and cancel: the image has been rejected. If on the other hand we subtract the two signals, the image will survive and the formerly wanted signal will be rejected, that is, we can choose to accept either the upper or lower signal relative to the LO (Figure 3.51).

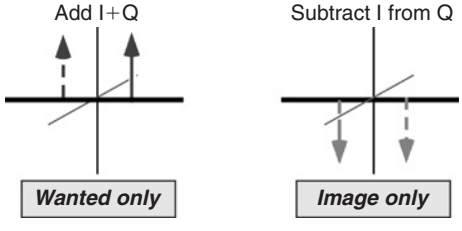


Figure 3.51: Alternatives for IRM Output

A block diagram depicting an IRM is shown in Figure 3.52. The price we pay for achieving rejection without filtering is additional complexity: we not only need two mixers instead of one, but also an accurate means of providing phase shifts both to the LO (to produce the I and Q branches) and the mixed signal to produce the phase-shifted results.

It is worth pausing for a moment to reflect upon Figures 3.48 and 3.49 in the context of a single digital signal. The digital signal generally contains sidebands at frequencies both above and below the carrier, playing the role of the image and wanted frequencies above. If

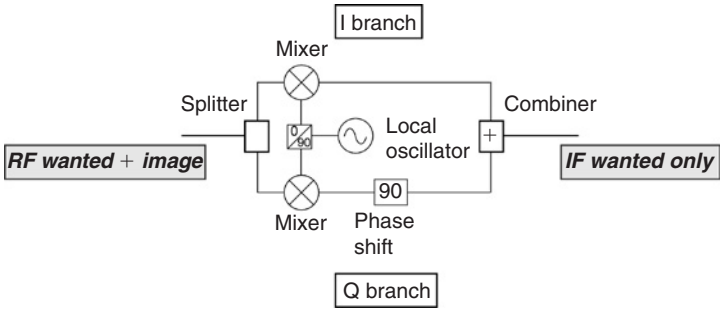


Figure 3.52: Block Diagram of an Image-Reject Mixer

we convert the carrier to DC in only one branch (I here, Figure 3.48), the upper and lower sidebands sum to form the mixed signal and cannot be separated out. If they are the same (an AM or on–off keying signal) we don't care, but if they are not (which is the case with, e.g., QPSK or QAM), we have lost the information needed to demodulate the signal. If we also perform a conversion with a phase-shifted LO signal (Q branch, Figure 3.49), the upper and lower sidebands subtract to produce the mixed result. By having both I and Q branches (the sum and difference of the upper and lower sidebands), we can separately extract the upper and lower sideband amplitudes and phases and thus the original signal. An I-Q or complex mixing operation is necessary to preserve the full information in a signal if the carrier is converted to DC; the same requirement applies in up-converting the transmitted signal. This is why we see I and Q mixers and modulators in our standard block diagram (e.g., Figure 3.5 or 3.37).

To provide an example of how a mixer is actually constructed, we take a quick look at a very popular arrangement, the Gilbert cell mixer. We examine an implementation based on bipolar transistors, but Gilbert cells can also be constructed using FETs as active devices. A schematic diagram of a simple Gilbert cell is shown in Figure 3.53. Figure 3.53 assumes a down-conversion application, but the circuit is equally applicable to a modulator or up-converter.

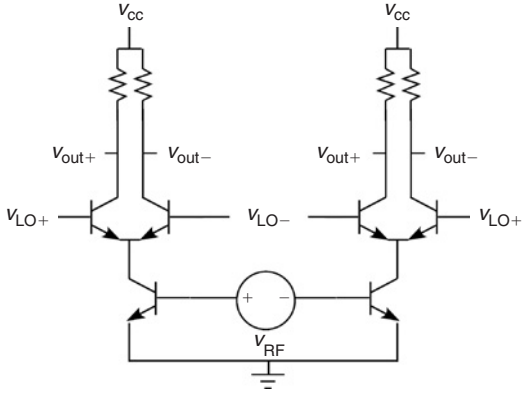


Figure 3.53: Gilbert Cell Mixer

A cell of this type is considered to be a triple-balanced mixer, because the input, voltage-controlled oscillator (VCO), and output connections are all differential. Such an arrangement is very helpful in achieving good isolation. As is apparent with the aid of Figure 3.54, the output voltage is zero if the RF voltage is zero, because in this case equal currents flow in the left and right branches of the mixer (assuming the devices are ideal and symmetric), and the output voltage is the sum of a positive LO contribution (the left branch) and a negative LO contribution (the right branch). Similarly, the output voltage is zero if the LO voltage is zero, because in this case equal currents flow in the upper pairs of the mixer and the left and right branch voltages are individually equal to 0. A fully balanced Gilbert cell will provide good LO-IF and RF-IF isolation if care is taken to ensure symmetric devices and layout.

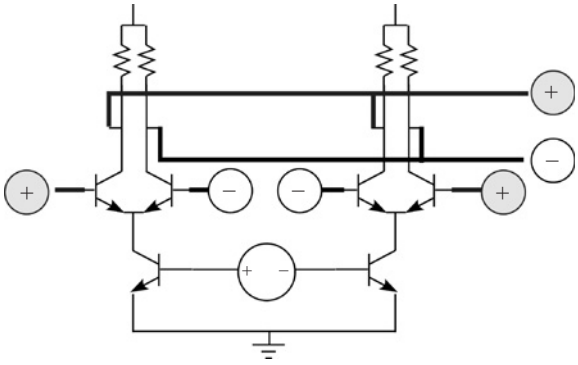


Figure 3.54: Balanced Input and Output Connections

The bottom pair of transistors converts the RF voltage into a differential current between the left and right branches of the mixer. The upper two pairs of transistors act as switches to direct the differential currents into either the positive or negative output connections (Figure 3.55).

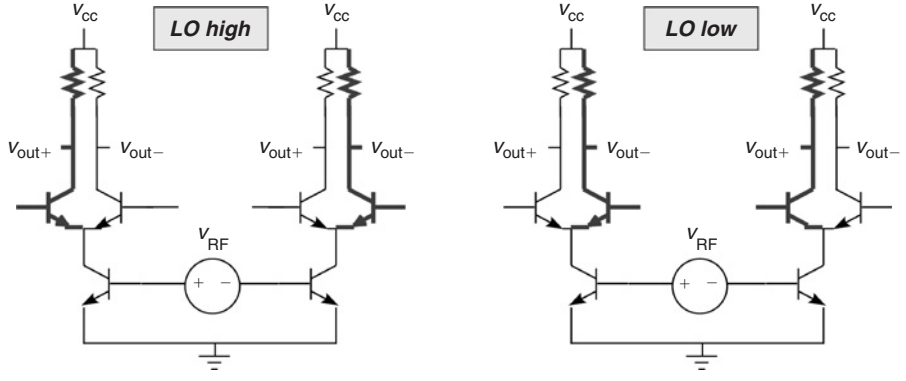


Figure 3.55: Gilbert Cell States

When the LO is high, the current is directed to the outermost branches and the output is taken in the same sense as the RF voltage. When the LO is low, the current is directed to the inner branches and the output is taken in the opposite sense to the RF current. The mixer in effect multiplies the input signal by a square wave with values of (+1, –1), performing the function of the balanced switch mixer of Figure 3.41.

Reported performance of some representative commercial mixers is summarized in Table 3.7. Performance of mixers incorporated within radio chips is not always reported separately; to focus only on the mixing function we included only discrete mixers in the table. The MESFET and MOSFET are passive mixers and display sizable conversion loss; the SiGe HBT is a Gilbert cell and thus benefits from the gain of the RF (bottom) transistors to provide conversion gain. However, noise figure is comparable between the three technologies and is quite large. As we discussed in connection with Figure 3.12, in a receiver it is important to precede the first RF mixer with enough gain to ensure that the noise figure of the receiver is preserved despite the mediocre noise performance of the mixing operation.

Table 3.7: Mixer Performance

Technology	Conversion Gain (Loss)	NF (2 GHz)	IIP3 (2 GHz)	LO-RF Isolation	Reference
GaAs MESFET	−9.0 dB	9.5 dB	30 dBm	37 dB	WJ Comm 03
Si MOSFET	−7.5 dB	7.5 dB	31 dBm	30 dB	Sirenza 03
SiGe HBT	+10.4 dB	10.2 dB	4.4 dBm	32 dB	Maxim 03

Mixer linearity, as measured by the input third-order intercept, can vary over a wide range. The difference between the IIP3 and the noise figure is a measure of the mixer's dynamic range. If the IIP3 is low, one must limit the gain before the mixer to avoid the spurious output of a large distorted interferer from blocking reception of the small wanted signal. This isn't going to work very well if the noise figure of the mixer is high. A high noise figure can be dealt with by additional gain before the mixer if the IIP3 of the mixer is large.

Isolation of 30–40 dB is achievable; this is normally sufficient to avoid any serious problems from leakage of the LO signal into the wanted signal. Isolation is dependent on the details of the mixer construction and packaging as much as on the active device technology. Full specification of a mixer is more complex than an amplifier, involving selection of three input frequency bands (RF, IF, LO), three isolations, three power levels, and innumerable possible spurious products.

3.2.4 Synthesizers

It is perhaps apparent from the discussions of the previous section that mixers don't work very well without a LO. The quality of the LO signal plays an important role in determining

several key parameters of the radio performance. As a transmitter, the LO must have sufficient absolute accuracy to deliver the proper transmitted carrier frequency within a precision typically specified in the standards; the receiver needs an accurate LO to convert the incoming RF signal accurately onto the IF bandwidth. The LO must be sufficiently tunable to address all the usable transmit or receive channels, and in the case of frequency-hopping protocols such as Bluetooth, the LO must tune rapidly enough to be ready to receive the next packet at the new frequency. Finally, the random variation in frequency—the *phase noise* of the oscillator—must be small enough to preserve the phase of the transmitted signal and allow accurate detection of the phase of the received signal, because as the reader will recall from Chapter 2 most WLAN modulations are phase sensitive.

Essentially all modern radios use a synthesizer architecture to produce the LO signals. A block diagram of a synthesizer is shown in Figure 3.56. At the heart of a synthesizer is a *voltage-controlled oscillator* (VCO), which produces a signal whose frequency varies monotonically with a control voltage over the desired band. The VCO is embedded in a *phase-locked loop* to accurately set its frequency.

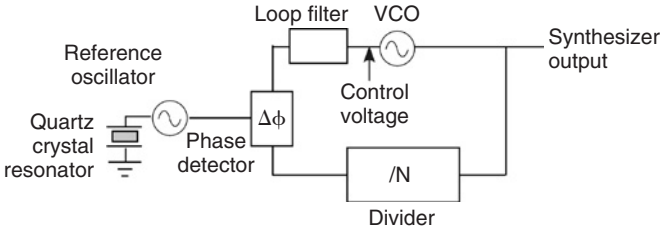


Figure 3.56: Synthesized Local Oscillator Block Diagram

The output of the synthesizer is split, and a portion of the signal drives a divider to produce a signal at much lower frequency, whose zero crossings are accurately synchronized to the LO signal. The divided signal is compared with a reference signal, typically generated by an oscillator whose frequency is set by a quartz crystal resonator. Quartz crystals are piezoelectric; in a resonator this property is exploited to convert an electrical signal to an acoustic displacement within a slice of bulk quartz. The crystal then acts as an acoustic resonator. Quartz resonators have very high quality factors (the reader who is unfamiliar with this concept should refer to section 3.2.5 below) and, if cut along the appropriate planes, have an acoustic velocity and therefore resonant frequency that is nearly temperature independent. A quartz crystal reference oscillator provides an inexpensive, reliable, accurate reference source. Typical frequencies for these oscillators are around 10 MHz, so the IF LO divisor is a few tens and the RF requires a larger divisor in the hundreds.

A phase detector determines the difference between the phase of the divided signal and the reference; this information is used to supply the control voltage for the VCO. A loop filter is provided to optimize the trade-off between responsiveness and stability, as with any feedback system. Roughly speaking, the LO control voltage is adjusted by the feedback loop until the phase error becomes small, ensuring that the synthesizer is locked to an integer multiple of the reference frequency.

The synthesizer in Figure 3.56 is not very flexible: for a fixed N, the output frequency will always be $N \cdot f_{\text{ref}}$. A more versatile synthesizer results if the divisor can be easily varied to allow differing output frequencies. One option is the *integer-N* synthesizer, in which two divisors can be used, differing by 1: for example, 16 and 17. The first modulus N is used for, for example, S cycles and then $(N + 1)$ for $(S - F)$ cycles, for a total of F cycles. After all the cycles are done, the divider outputs one rising or falling edge. The net effect is to divide by $N_{\text{eff}} = (N + 1)S + N(F - S) = NS + S + NF - NS = NF + S$. Thus, by adjusting S (which just involves setting a counter), the overall divisor and thus the output frequency can be adjusted over a wide range, with a resolution of f_{ref}. If higher resolution is desired, the reference frequency can also be divided by some other integer M. However, using larger and larger divisors has a penalty: because the divider outputs an edge only after every N_{eff} cycles of the VCO output, information about the phase of the VCO signal becomes increasingly sparse as the divisor grows. The result is that the variation of phase that can occur without being suppressed by the feedback loop—the phase noise—increases with increasing N.

An integer-N synthesizer is typically adequate for WLAN-type applications, where channels are separated by 5 MHz or some other convenient value, but when better resolution is required a different architecture, the *fractional-N* synthesizer, may be used. In this case the divisor is again dithered between two values differing by 1, but in this case an edge is output after each divide cycle. If the fraction of time spent at divisor N is K, then the output frequencies are $f_{\text{ref}}(N + 1/K)$: the resolution can be made arbitrarily fine by increasing the total number of cycles to achieve accurate control over K. The trade-off is that the VCO frequency is actually varying on an instantaneous basis: it is a frequency-modulated signal. The frequency modulation results in undesired low-level spurious output at a range of frequencies determined in part by the timing of the index dither. Because these spurs are predictable based on the known timing of the divider modulus, they can be corrected for by downstream adjustments. Considerable progress has been made in recent years to minimize spurious outputs of fractional-N synthesizers, but they remain significantly more complex than integer-N architectures.

The absolute accuracy of transmitted signals must meet regulatory restrictions and compliance with standards requirements. For example, the 802.11 classic PHY requires that the transmitted center frequency be accurate to 25 ppm, which works out to 61 kHz for ISM channel 6. The receiver typically uses the same synthesizer used on transmit.

The difference between the transmitted signal frequency and that to which the receive LO tunes constitutes an effective error in the frequency of the received signal when converted to nominal zero frequency; the error is on the order of 100 kHz. For two successive DQPSK signals at 11 Mchips/second, the error adds a phase difference of $2\pi(10^5)(9 \times 10^{-8}) \approx 0.06$ radians, which is much less than the 1.6 radians between QPSK constellation points.

OFDM signals have special requirements for frequency and phase control. An error in the received frequency results in the displacement of the apparent position of the subcarriers. The orthogonality between subcarriers only applies when the sample time is an integer number of cycles; an error in subcarrier frequency causes inter-subcarrier interference, the mistaken assignment of some of a given subcarrier's signal to its neighbors (Figure 3.57). A 100-kHz frequency offset is obviously a large fraction of the 312.5-kHz separation between subcarriers in 802.11a/g and would cause significant interference between neighboring subcarriers. Fortunately, a simple offset of this type is easily corrected in digital signal processing after demodulation.

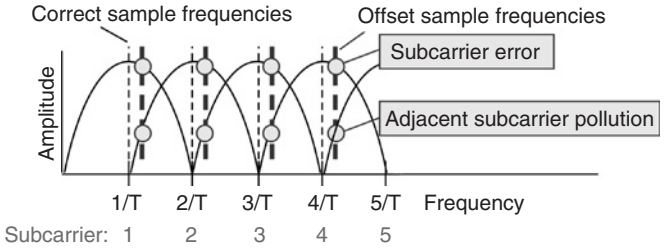


Figure 3.57: Offset Frequency and OFDM Subcarriers

Instantaneous random variations in frequency—phase noise—cause the measured constellation points to be smeared out slightly and can be approximately regarded as another noise source in addition to the usual thermal noise. Thus, the total phase noise integrated over a frequency range corresponding to a symbol time degrades the (S/N) of the receiver and the EVM of the transmitter. Phase noise can also have a more subtle deleterious effect by broadening a large interferer on a nearby channel so that part of the interfering signal appears on the wanted channel (Figure 3.58).

An example of synthesizer phase noise performance in a WLAN chipset is given in Figure 3.59. The phase noise is normally measured as power relative to the carrier power (dBc) in a given bandwidth. Recall that one 802.11 OFDM symbol lasts 4 μsec, so phase noise at frequencies less than about 250 kHz corresponds to phase variation from one symbol to another; phase noise at frequencies above about 1 MHz corresponds to phase variation during a single OFDM symbol.

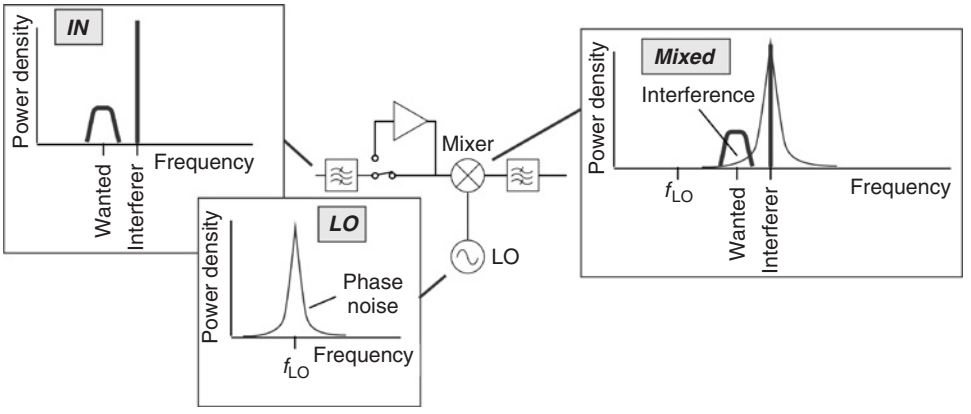


Figure 3.58: LO Phase Noise Broadens Interferer to Block Wanted Signal

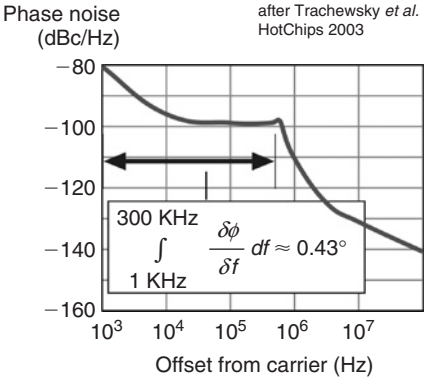


Figure 3.59: Example of Synthesizer Phase Noise for an 802.11a Radio Chip

The total phase error from 1 kHz (corresponding to the length of a complete packet) to 300 kHz is less than 1 degree, so that the radio can use the preamble of the packet to establish frequency offset and phase synchronization and thereafter preserve reasonable phase coherency over the remainder of the packet. Simulations have shown that this level of phase noise corresponds to less than 1 dB degradation in (S/N) required for a given bit error rate.

In early WLAN architectures the VCO was often a separate chip, with the synthesizer logic (the phase-locked loop part) integrated onto the converter chip. Modern chipsets normally include the necessary synthesizers integrated onto the radio chip.

3.2.5 Filters

Filters are circuits that reject some frequencies and transmit others. They are generally classified into three types: low pass, high pass, and bandpass (Figure 3.60).

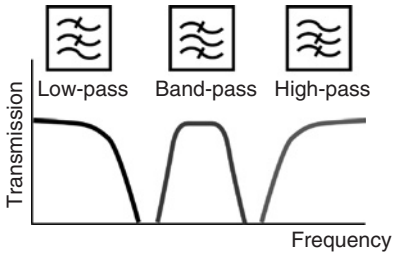


Figure 3.60: Three Types of Filters With Common Schematic Symbol

Filters play a key role in rejecting undesired signals in a radio (Figure 3.61). The first filter in the receiver, the band or image reject filter, is designed to pass the RF band of interest (e.g., the 2.4- to 2.48-GHz ISM band) and reject signals from nearby bands.

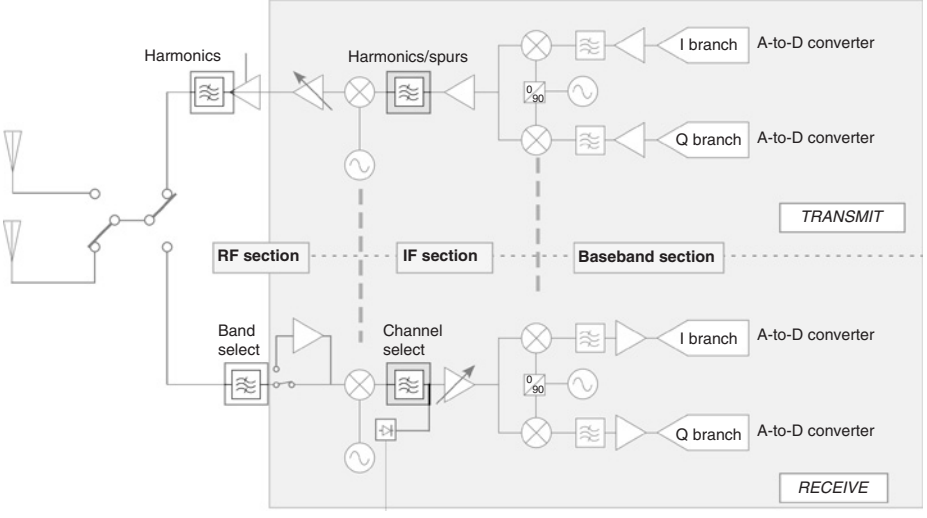


Figure 3.61: RF Filters in a Superheterodyne Radio

After mixing, a superhet radio passes the signal through a channel filter that, in combination with the LO frequency, selects the desired channel (e.g., channel 1 at 2412 MHz). After the demodulation step, or after the conversion step in a direct conversion receiver, additional low-pass filtering is used, although the frequencies of interest in this case are a few megahertz to perhaps 20 MHz, and discrete components or active filters can be used. Filters are also used in the corresponding stages of the transmit operation; in a WLAN superhet radio, a common IF is chosen so that the IF channel filter can be used both for transmit and receive to minimize cost.

The simplest bandpass filter structure is a resonator composed of an inductor and a capacitor (Figure 3.62). The resistor represents unintentional parasitic losses, typically concentrated in the inductor wiring. Resonance occurs at the frequency at which the inductive reactance and

capacitive susceptance (i.e., the positive and negative imaginary parts of the conductance) are exactly equal. During each cycle, energy is alternately stored in the inductor (at the moment of peak current flow) and the capacitor (at the moment of peak voltage). An ideal resonator with no loss would appear as a perfect open circuit at resonance, so that all the input current would be transferred to the output. Away from resonance, the current would be shorted through the inductor (low frequency) or capacitor (high frequency). Thus, the resonator acts like a bandpass filter.

The filter can be characterized by a characteristic impedance Z_o and a quality factor Q. The quality factor is the ratio of the characteristic impedance to the parasitic resistance; it expresses the ratio of the energy stored in the inductor and capacitor to the energy lost per cycle by the resistor. As shown in the right half of the figure, the quality factor also determines the width of the passband of the filter. Narrow passband filters must have very high Q. For example, an RF band filter for the 2.4-GHz 80-MHz wide ISM band must have a bandwidth on the order of 3% of the center frequency, requiring a Q of at least 30. (In practice, considerably higher Q is needed to make a good filter: the filter should have a fairly flat transmission in the passband and a sharp transition to very low transmission in the stopbands, rather than the peaked behavior shown in Figure 3.62). On-chip filters constructed with inductors and capacitors in Si CMOS processes are generally limited to Qs of about 10, mostly due to loss in the inductors. Discrete components offer Qs as high as 30 up to a few gigahertz, but complex filters with many elements constructed using discrete components will become physically large and are inappropriate for gigahertz frequencies. Better technologies with high Q, small physical size, and low cost are needed to provide image and channel filtering.

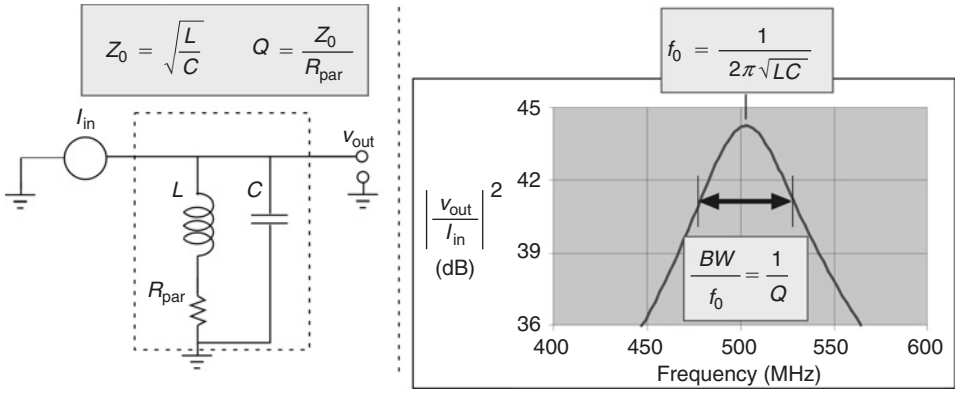


Figure 3.62: Simple L‖C Filter

One of the most commonly used technologies for high frequencies is the *surface acoustic wave* (SAW) filter. SAW filters achieve Qs in the hundreds, are available in surface-mount packages, and can pass hundreds of milliwatts without damage. SAW filters are relatively expensive ($1 to $10) and, on the order of 1 cm square, are large enough that the number

of filters must be minimized both to conserve board space and minimize cost. The resonant frequency of a SAW filter is somewhat temperature dependent; quartz filters are better in this respect than most other piezoelectric materials but are more expensive to fabricate.

A simplified SAW filter structure is shown in Figure 3.63. The device is constructed on a piezoelectric substrate such as quartz, $LiNbO_4$, or ZnO. Electrical transducers are constructed of a layer of a conductive metal such as aluminum, deposited and patterned on the surface of the piezoelectric using techniques similar to those used in integrated circuit fabrication.

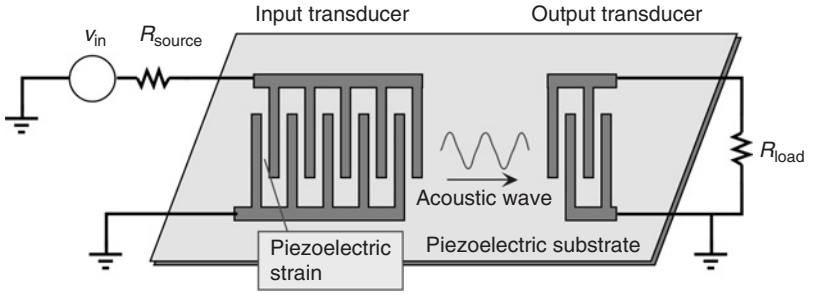


Figure 3.63: Simplified SAW Filter

The input transducer consists of on the order of 100 interdigitated fingers, driven at alternating polarity from an RF source. Between each pair of fingers an electric field is formed within the piezoelectric material, inducing a time-dependent strain that creates an acoustic wave. For an input frequency such that the spacing between fingers is half of the acoustic wavelength, a resonant enhancement of the wave will occur as it propagates along the transducer, as each alternating region of strain will be in phase with the wave and add to the displacement. The resulting strong acoustic wave propagates to the smaller output transducer, where the acoustic strain induces an electric field between the electrodes, resulting in an output voltage. The slice of piezoelectric is often cut at an angle to the propagation axis, as shown in Figure 3.63, so that the acoustic energy that is not converted back to electrical energy is reflected off the edges of the substrate at an odd angle and dissipated before it can interfere with the desired operation of the filter. Because the acoustic wave propagates about 10,000 times more slowly than electromagnetic radiation, wavelengths for microwave frequencies are on the order of 1 mm, making it possible to create compact high-Q filter designs.

The performance of a fairly typical RF band (or image-reject) filter is shown in Figure 3.64, as the transmission in decibels through the filter versus frequency. Within the ISM band, the loss through the filter is only about 3 ± 1 dB: this transmission is known as *insertion loss*, because it is the loss in band due to insertion of the filter in the circuit. Low insertion loss is important on both transmit and receive. The insertion loss of the transmit filter comes directly out of the signal power, so lossy filters mean bigger transmit power amplifiers that cost more and

consume more DC power. On the receive side, the filter loss is essentially equal to its noise figure, and because it is typically placed before the LNA, the filter noise figure must be added directly to the noise figure of the chain.

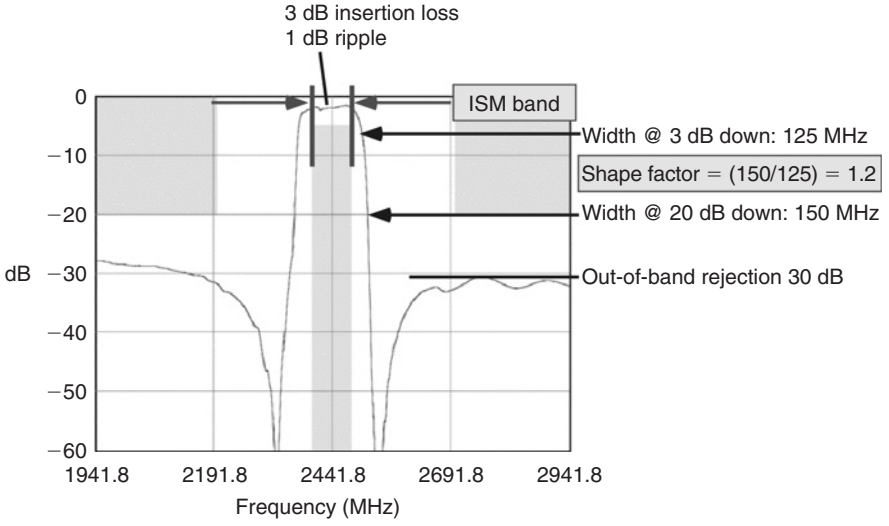


Figure 3.64: SAWTEK 855916 ISM-Band Filter Transmission vs. Frequency (Based on data plot © 2004 TriQuint Semiconductor; used by permission of TriQuint Semiconductor, SAWTEK Division)

Other important properties of a filter are the sharpness with which transmission cuts off once the frequency is beyond the edges of the band and the transmission (hopefully small, thus rejection) of out-of-band signals. The bandwidth in this case at a 3-dB decrease in transmission versus the center frequency is about 125 MHz, rather noticeably wider than the 83-MHz ISM band: one cannot expect perfect rejection of a signal a couple of megahertz out of band. However, transmission falls quite rapidly thereafter: the *shape factor*, the ratio of bandwidth at 20 dB rejection to 3 dB rejection, is only 1.2. The rejection of signals far from the band edges—for example, the third-generation cell phone (UMTS) downlink frequency at 2170 MHz or the image frequency for a high IF superhet—is a substantial 30 dB.

Performance of an IF (channel) filter is shown in Figure 3.65. Higher insertion loss is tolerable for an IF filter as the transmit power at this point is modest and the receive-side gain before the filter ensures that the loss has little effect on noise figure. The benefit of accepting higher loss is the improved out-of-band rejection: the upper adjacent channel suffers about 43 dB loss and the lower adjacent channel 55 dB. High adjacent channel rejection is necessary for a receive channel filter likely to be exposed to simultaneous transmissions at the wanted and neighboring channels.

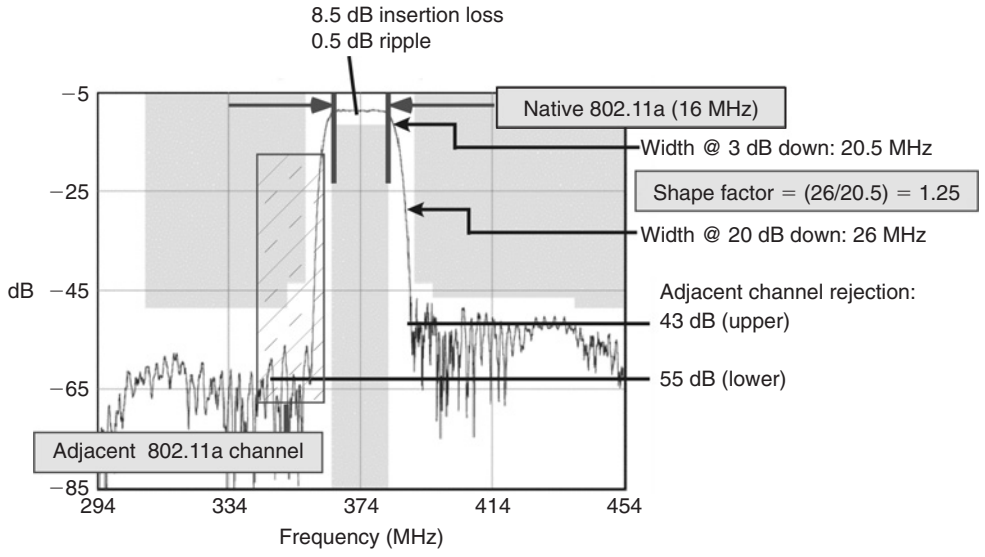


Figure 3.65: SAWTEK 855898 374-MHz IF Filter (Based on data plot © 2004 TriQuint Semiconductor; used by permission of TriQuint Semiconductor, SAWTEK Division)

A related but distinct filter technology is the *bulk acoustic wave* (BAW) filter, sometimes referred to as a film bulk acoustic resonator. In place of the relatively complex lateral structure of a SAW device, a BAW substitutes a simple sandwich of a piezoelectric thin film between two metal electrodes. If the thickness of the piezoelectric is equal to a half wavelength, a resonant acoustic wave is created within the film. The key problem in fabricating a BAW device based on thin film techniques is to avoid leakage of the acoustic wave into the substrate, which would represent a loss mechanism and decrease the quality factor of the filter. Two common methods of solving this problem are shown schematically in Figure 3.66. The core BAW device, consisting of a bottom metal electrode, piezoelectric layer such as AlN or ZnO, and a metal top electrode, is the same in both cases. The difference is the means of isolation from the substrate. One approach uses a mirror constructed of layers of alternating high and low acoustic impedance, each one-fourth of an acoustic wavelength thick. The acoustic wave is reflected at each successive interface between the mirror layers with opposite

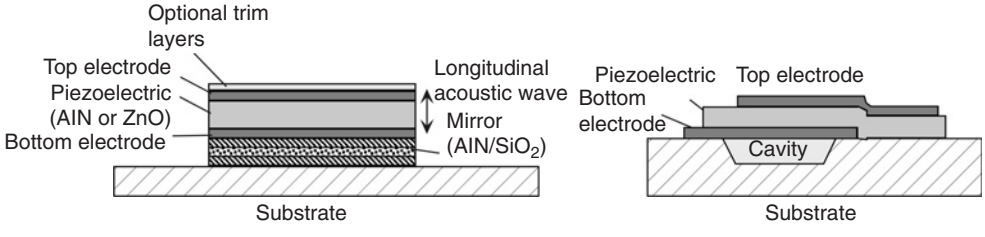


Figure 3.66: Cross-Sectional View of Two Alternative Approaches to BAW Fabrication

sign. The net phase shift of a half-wave (there and back again through a quarter-wave layer) means that the reflections from all the interfaces add in phase at the bottom electrode to produce a very high reflection coefficient and little transmission to the substrate.

A conceptually simpler but practically challenging method is to fabricate the active part of the piezoelectric layer over a cavity filled with air or vacuum, which conducts little acoustic radiation into the substrate. Such structures are normally formed by subtractive techniques. The cavity is etched into the substrate, filled with a material that can later be selectively etched such as silicon dioxide, and the resulting flat surface used to fabricate the remaining BAW layers. The device can then be immersed in an etchant that removes the cavity fill layer, leaving a void behind.

BAW filters have some advantages over SAW filters. The key dimension is the thickness of the layer rather than the lateral separation of electrode lines as in the SAW device. This thickness can usually be controlled more easily than a linewidth, so that good precision is possible. Typically, piezoelectric thickness are one to several microns for frequencies on the order of one to a few gigahertz. Thin trim layers of metal or dielectric may be added to compensate for any nonuniformity or imprecision in the piezoelectric layers to target precisely the correct resonant frequency. Power handling capability of BAW filters is higher than a similar SAW filter, both because the acoustic wave is distributed uniformly over the piezoelectric material, whereas in a SAW filter the energy is confined within a wavelength of the surface, and because in a SAW filter significant area is consumed by the periphery of the electrodes, edge reflection suppression, and other acoustically inactive regions. BAW filters, being thin-film devices, can in principle be directly integrated in a standard silicon fabrication process, though cost and complexity are significantly increased. High-performance BAW filters are currently commercially available for cell phone duplexers at PCS (1.8 GHz) band in the United States and similar (GSM 1900) frequencies in Europe and Asia.

High-performance filters can also be constructed using ceramic dielectric resonators. The resonator is simply a typically cylindrical slab of high-dielectric-constant material; a coil or other simple electrode structure can excite electrically resonant modes, like those within a metal cavity, which are confined mostly within the dielectric because of the high dielectric susceptibility. The basic challenge is that the size of the filter is on the order of half a wavelength of the electrical radiation in the dielectric; for low frequencies (for example, for IF channel filtering) such a structure is typically impractically large. Ceramic resonator filters become increasingly sensible at higher frequencies: at 5 GHz, the free-space wavelength is only about 5 cm, so a half-wave resonator with a dielectric constant of 10 would require a minimum dimension of about $5/\sqrt{(10)} \approx 1.5\,\text{cm},$ just at the boundary of a practical surface-mount device. Ceramic resonator construction is fairly simple; the main challenges are in the production of an accurately dimensioned, high-dielectric-constant, low-loss ceramic puck.

Design of a radio, particularly a superhet architecture, is often a trade-off between radio performance and filter capability. A good band filter protects the LNA from out-of-band interferers and thus reduces the requirements for third-order distortion in the LNA and mixer; a cheaper filter will require a better amplifier. On transmit, harmonic filters can remove spurs generated by distortion in the power amplifier, reducing the linearity required of this expensive and power-hungry component. Direct-conversion and NZIF architectures are desirable in part because they place the channel filtering operation at low frequencies where integrated solutions are available. Filter technology influences radio design; as it progresses, it may be expected that optimal design choices for low-cost radios will also evolve.

3.2.6 Switches

Common WLAN protocols such as all 802.11 standards, Bluetooth and 802.15.3, and HiperLAN are all half-duplex: the radio switches between transmit and receive states but does not occupy both at once. Some client cards and most access points also make use of either one of two *diversity antennas*, depending on which produces the best received signal strength at a given moment. The choice of antenna changes from one packet to the next and thus a switching capability between antennas must be provided. The net result is that several switches are usually interposed between the radio input (filter/LNA) and output (power amplifier/filter) and the antenna(s).

The key performance issues in switch design and selection are as follows:

- *Insertion loss*: Is on on? The signal loss encountered in traversing the switch in the ''on'' state must be as small as practical. Insertion loss subtracts directly from the output power on transmit and adds directly to noise figure on the receive side. Typical values are 1–2 dB at 5 GHz. To minimize insertion loss, the series devices in the switch are made larger to reduce their on-state resistance.

- *Isolation*: Is off off? The isolation is the amount of power that sneaks through from the input to the output despite the switch being nominally in the ''off'' state. Typical 802.11 applications are not very demanding in this respect. During transmit, the power to the receive side LNA is typically disabled, so that fairly high input power can be tolerated without damage. The transmitter is also powered down during receive but would not be damaged by the small received RF power in any case: the issue here is residual transmitted power or noise interfering with the received signal. Small amounts of power coupled from the unused diversity antenna have little effect on the received power. Isolation values of 20–30 dB are relatively easy to obtain and are sufficient for WLAN radios.

- *Speed*: The 802.11 standards require switching from transmit to receive in roughly 1 μsec to ensure that the transmitter can respond after the short interframe space of

10 μsec. Choice of the diversity antenna must be made during the preamble, which varies somewhat in the various protocols but is much longer than 1 μsec. A slower switch could be used, but it is usually cheaper and simpler to use the same part for both roles.

- *Power handling*: A switch is ideally a passive part, having no effect on the signal when it is in the "on" state. Real switches are constructed with active devices that can introduce additional distortion to the transmitted signal; the output intercept power and compressed power must be sufficiently in excess of the actual transmitted power to ensure that the transmitted signal remains compliant with the spectral mask and regulatory requirements.

Switches can be implemented in any technology that supports active devices. FETs or BJTs in silicon or compound semiconductors are used in this function. FETs typically provide better isolation than BJTs. Compound semiconductors have superior on resistance due to higher electron mobility and better output power ratings due to higher saturated current and breakdown fields but cannot be integrated into the silicon CMOS radio. Separately packaged switches implemented with GaAs FETs are a common option.

A specialized versatile switching technology commonly encountered in microwave applications is the p-intrinsic-n or PIN diode. PIN diodes have a relatively large intrinsic (not intentionally doped) semiconductor region that makes them very resistive at both low and high frequencies when no DC current is flowing. However, application of DC forward bias causes electrons to be injected from the n region and holes from the p region, increasing the conductivity of the intrinsic region. PIN diode microwave resistance can be varied from $>1000\,\Omega$ to $<5\,\Omega$ by varying the DC current from 0 to 10–20 mA. PIN diodes are inexpensive and have excellent power handling capability, but their use requires provisions for separating the variable DC bias from the RF signal.

Switches can also be integrated in Si CMOS processing. An example of such an implementation, specifically designed as an 802.11a T/R switch, is shown in Figure 3.67.

In the TRANSMIT state, the series transistor (250 μm wide) in the TX side is turned on. Note that the substrate side of this transistor is connected to a parallel L‖C filter whose resonant frequency is about 5.2 GHz. Recall that such a filter presents a large input impedance at resonance; thus, the back connection of the MOSFET is essentially isolated from the rest of the (grounded) substrate and is able to follow the imposed RF voltage. The bias resistor on the gate serves a similar purpose; it can be a large value because little charge is required to switch the FET state. The net result is that the FET gate and back gate are free to follow the instantaneous RF potential. Therefore, there is no RF voltage between the channel and gate and consequently little variation in the channel conductance during the RF cycle. Without

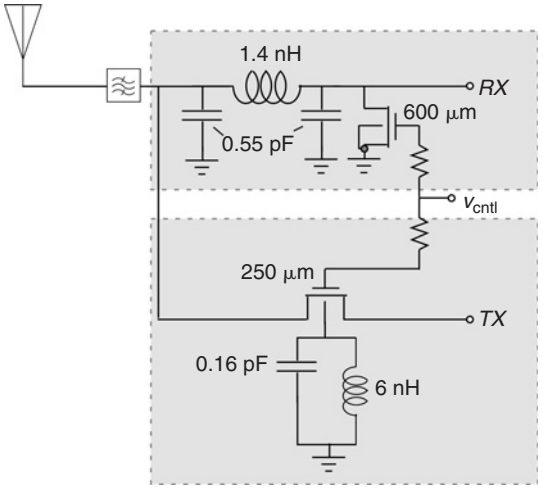


Figure 3.67: Simplified Schematic of Integrated CMOS Switch (After Talwalker et al., ISSCC 2003)

this provision, the gate and/or substrate voltages would be fixed, causing the channel to be partially pinched off when a large RF voltage is present and leading to signal-dependent conductance—that is, distortion—of the RF signal.

Isolation of the receiver is achieved by turning the large parallel MOSFET (600 μm wide) on as well. This transistor shorts out the second capacitor; the first 0.55 pF capacitor and the 1.4 nH inductor together form a parallel resonator at 5 GHz, which again provides a large impedance and prevents the signal from reaching the FET. The *Q* of the filter is about 13, providing about 20 dB of isolation. Whatever signal does reach the FET is then shorted to ground by its small impedance, realizing an additional 10 dB of isolation.

In the RECEIVE state, the FETs are both switched off. The C—L—C structure then provides matching of the receiver to the (presumed 50V) filter/antenna. The TX FET, being off, provides about 20 dB of isolation. Performance of the switch is compared with a typical GaAs MESFET switch in Table 3.8. The CMOS switch achieves comparable TX output power and isolation, though insertion loss is slightly inferior to the compound semiconductor approach. The design requires about 0:6 mm^2 of silicon area, which is about 10% of the total area of a radio chip: a significant though not catastrophic increase in chip area.

It is thus possible to integrate the switching function with the radio chip. Whether architectures will move in this direction or not is a function of the cost benefits (if any) of doing so and whether they outweigh the loss of flexibility the designer gains from using external instead of integrated components.

Table 3.8: Performance of Typical MESFET Switch and CMOS T/R Switch Compared

Parameter	MESFET	CMOS TX	CMOS RX
Insertion loss (dB)	1.2	1.5	1.4
Isolation (dB)	26	30	15
P1dB (dBm)	31	28	11.5

3.3 Radio System Design

Now that we are familiar with the various components used in constructing a radio, we can take a look at the overall problem of designing a radio for a given set of requirements.

Because of the half-duplex nature of a WLAN radio and the importance of component cost, it is generally the case that key choices in architecture are the same for transmit and receive. In a receiver, a key choice is the frequency of first conversion. The choice of the conversion frequency or frequencies, and the corresponding LO frequency or frequencies, is generally referred to as frequency planning. In a direct conversion architecture frequency planning is simple: there is no IF, and the LO frequency must equal the desired RF channel.

In a superhet design, there are several possible choices of IF (Figure 3.68). In an NZIF radio, the IF is chosen to be comparable with the bandwidth of the baseband signal, typically a few megahertz. Channel filtering is done at megahertz frequencies and can be accomplished with on-chip filters or inexpensive discrete components. However, the image frequency is very close to the wanted frequency and cannot be rejected by filtering. Instead, an IRM design must be used.

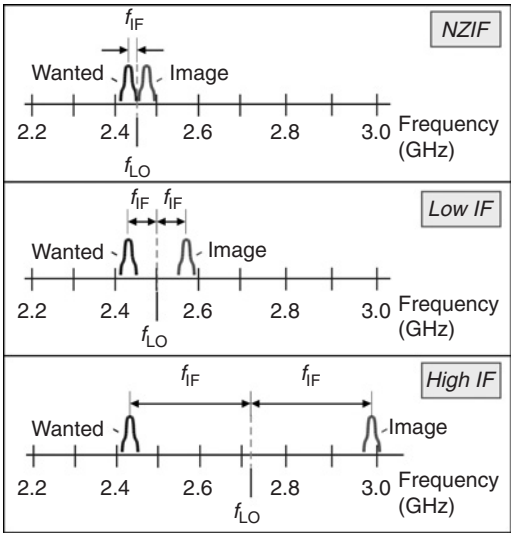


Figure 3.68: Choices for IF in a Superhet Design

Low-IF designs, where the IF is typically chosen between a few 10s of MHz and 200–300 MHz, represent the "classic" approach to superhet radios. The band filter can filter the image frequency effectively if the IF is greater than about 80 MHz (recall from Figure 3.64 that a typical SAW filter has a 20-dB bandwidth of about 160 MHz). Choice of a high value of the IF makes the image filtering easier but requires a high-performance SAW or similar filter for the IF. In such a design, the IF is common for transmit and receive so that the relatively expensive channel filter can be used in both directions.

In high-IF designs, the IF is several hundred megahertz to a gigahertz. In a high-IF design, image filtering is relatively easy and can be done with discrete components or on chip; however, some of the advantages of a superhet design are vitiated, in that IF gain is more difficult to obtain and channel filtering is difficult.

The designer also needs to choose whether the LO frequency is larger than the wanted RF frequency (*high-side injection*) or less than the wanted frequency (*low-side injection*). High-side injection requires higher absolute LO frequency but less percentage tunability. Low-side injection may make it more difficult to filter spurious outputs of the LO, because they are closer to the RF frequency.

Once the architectural decisions are made, the actual design uses chain analysis to follow the signal through the radio, keeping track of the signal strength, noise floor, IM distortion, and maximum unclipped power at each stage. The analysis enables the designer to estimate the receiver sensitivity, gain control requirements, and interferer rejection. On transmit the analysis is simpler, because there are no interferers.

A generic superhet radio chain is shown in Figure 3.69; the performance parameters are loosely based on the Intersil/Virata Prism 2.5, which was a very popular early 802.11b chipset. The radio is shown in the high-gain (best sensitivity) configuration. The signal passes through two switches (diversity antenna select and T/R), each of which contributes 1 dB of loss. The

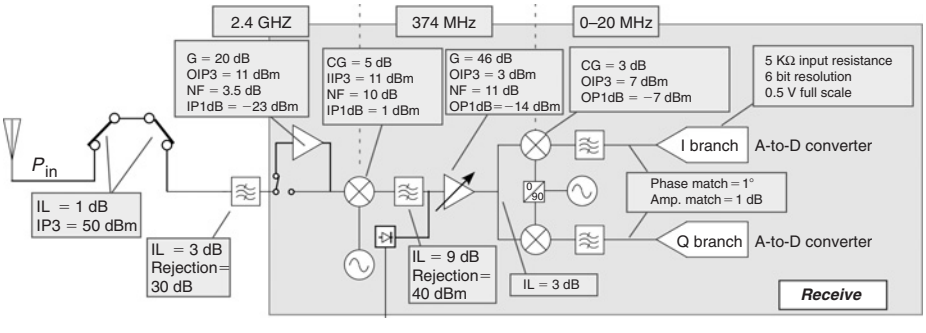


Figure 3.69: Simplified Superheterodyne Receive Chain (IL = Insertion Loss; IP3 = Third-Order Intercept; NF = Noise Figure; P1db = 1-dB-Compressed Power)

switch IP3 is taken as 50 dBm, so they contribute little distortion for the tiny RF signals likely to be encountered on receive. The band select (image reject) filter is taken to contribute 3 dB additional loss and provide 30 dB image rejection. The LNA has 20 dB of gain, enough so that subsequent stages will contribute only modestly to the overall noise figure. The output IP3 is 11 dBm, so the input IP3 is (11 – 20) = −9 dBm. The first mixer is presumed an active mixer with some conversion gain. Interference rejection is set mostly by the linearity of these two stages, because after conversion the signal must pass through the channel filter, which will reject most interferers. After filtering, a variable-gain IF amplifier boosts the signal before final I/Q demodulation, low-pass filtering, and ADC.

The chain analysis for sensitivity is shown in Figure 3.70, assuming the minimum signal required by the 802.11b standard. The switches and band filter do not change the noise floor but decrease the signal, so (S/N) is degraded in the first few stages.

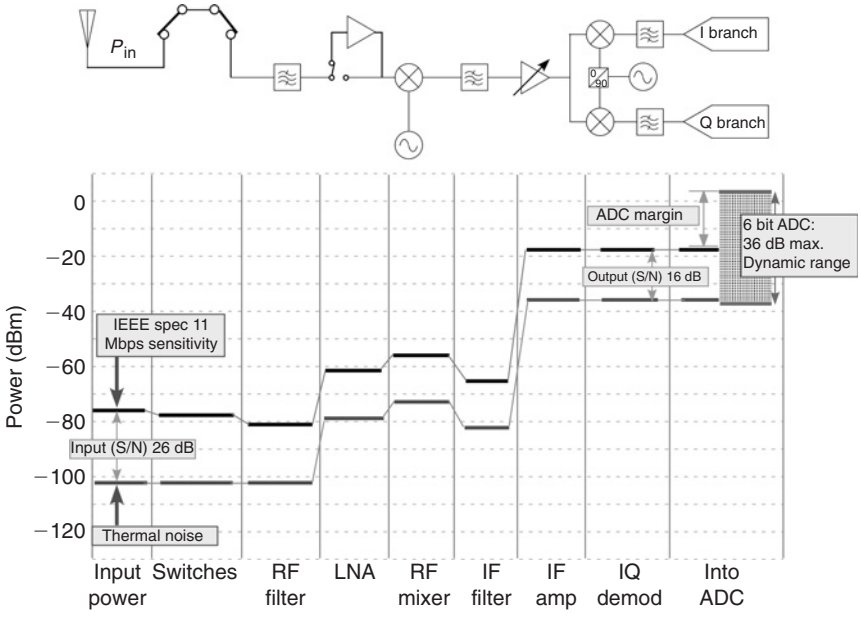


Figure 3.70: Chain Analysis for Noise, Parameters From Figure 4.69

Note that after the LNA, the signal level is far above the thermal noise level: excess noise in subsequent stages has only a modest effect on (S/N). In this maximum gain configuration, the output (S/N) is 16 dB for an input (S/N) of 26 dB, corresponding to an overall noise figure for the receiver of 10 dB, of which 8.5 dB comes from the input switches, filters, and LNA. The resulting signal can be placed roughly in the middle of the ADC dynamic range, allowing for minimal (S/N) degradation due to quantization noise in the converter and leaving some room for the signal strength to increase without saturating the ADC. The final (S/N) of 16 dB

is quite sufficient to demodulate QPSK with good error rate, in view of the 2 dB gain from complementary code keying.

Now that we have established that the design meets minimum requirements for sensitivity, let's examine selectivity. Figure 3.71 shows what happens when interfering signals are present on adjacent channels. In this case the hatched lines depict the IP3 of the relevant stage; recall that the IM product in dBc is suppressed by twice the separation of the signal level and the intercept. The RF mixer nonlinearity turns out to contribute most of the distortion. The resulting IM product (shown as a solid line) is only a few decibels below the wanted signal. A comparable interfering signal results from the limited rejection of the channel filter (dotted lines).

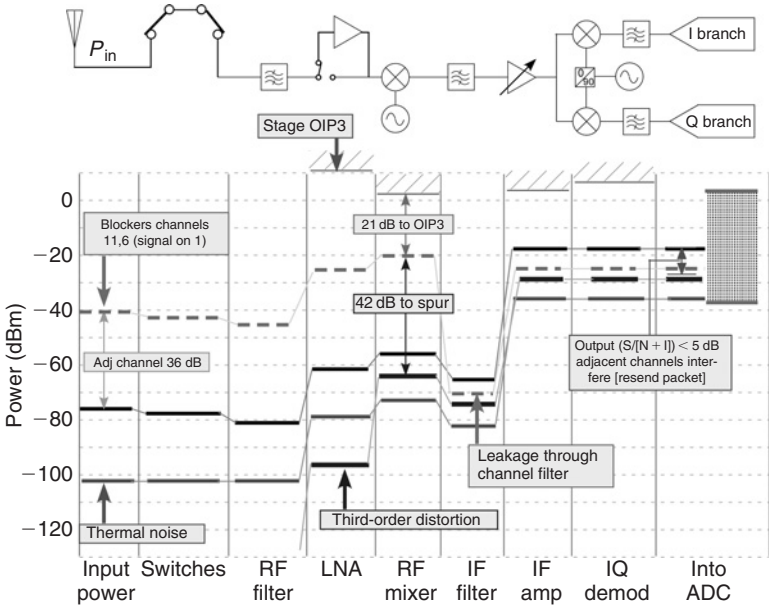


Figure 3.71: Chain Analysis With Interferers Present

The combination of the two interfering signals gives rise to a (signal/(noise + interference)) ratio (S/(N + I)) of only about 5 dB, insufficient for 802.11b transmission (but sufficient for 1 Mbps BPSK with spreading gain). The radio will not be able to communicate at 11 Mbps under these conditions but will work at 1 Mbps. Retransmission of the packet is also very likely to solve the problem, because the interfering stations are unlikely to both be present at the next transmission slot.

This simple example shows that a good design must trade off sensitivity and linearity. If we had added another 10 dB of LNA gain, the sensitivity would be improved but the distortion of the interferers in the RF mixer would have made the selectivity poor. If the wanted signal is much larger than the sensitivity threshold, significant distortion of the wanted signal could occur in the analog stages or through saturation of the ADC; we need to turn the overall gain

down for larger wanted signals. Gain adjustment is usually performed during the preamble to the packet. Improved linearity would give better selectivity but at the cost of increased DC power and chip area. Fortunately, in an 802.11 environment interference is accounted for by retransmission, with graceful degradation in performance. HiperLAN and Bluetooth are time-division multiplexed and somewhat less tolerant of interferers.

We mentioned direct conversion as an alternative to superhet radios. Direct conversion radios use fewer stages and thus fewer components than superhet radios, shown in Figure 3.72. Direct transmission is of particular interest because there are no interfering signals to deal with and thus less need for IF filtering. Because there is no image frequency to remove, transmit filtering need merely be concerned with eliminating harmonics, which is generally much easier.

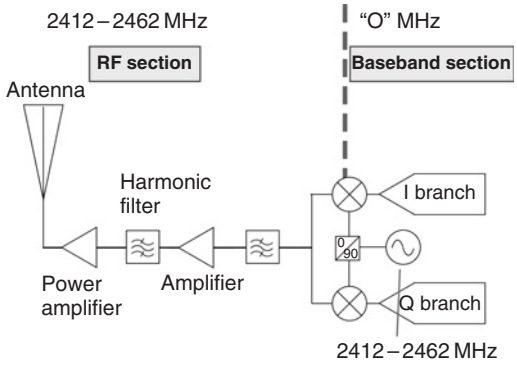


Figure 3.72: Direct-Conversion Transmitter

Direct transmission encounters some special difficulties. The transmitted symbols are points in amplitude-phase space. Errors in amplitude or phase give rise to errors in the transmitted signals, which show up as increased EVM; examples of distorted constellations are depicted in Figure 3.73. Phase match between the I and Q branches is much more difficult when the

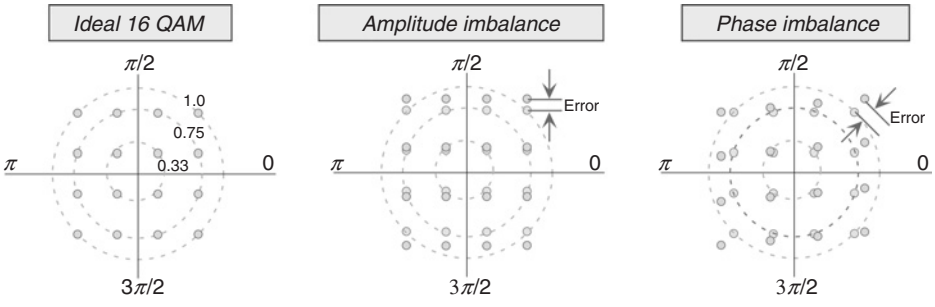


Figure 3.73: Effect of Imbalances in Amplitude and Phase Between I and Q Branches on the Transmitted Signal

transmitted signal is at RF frequency rather than IF, as shown in Table 3.9. Direct transmission is only practical with an integrated modulator and requires very careful attention to symmetry in circuit layout to ensure good phase matching. Modern implementations are also incorporating increasing amounts of on-chip calibration of phase and amplitude match.

Table 3.9: Effect of Increasing Modulation Frequency on Imbalance Requirements

Frequency (MHz)	Phase Offset (degrees)	Time Offset (psec)	Line Offset (μm)	Capacitance Offset (pF)	Implementation
90	1	30	4500	0.15	Hybrid
374	1	7.5	1130	0.04	Multiple ICs
2440	1	1.1	165	0.005	Single-chip

Direct transmission at high-output power also encounters another somewhat more intractable challenge. Because there is no channel filter, any noise originating in the modulator is converted into broadband noise in the transmitter. The RF band filter has a much sloppier cutoff than an IF channel filter. For example, by reference to Figures 3.64 and 3.65, the IF filter might achieve 20 dB of rejection by 5 MHz beyond the signal edges, whereas the RF filter requires about 80 MHz outside the band edge to reach the same rejection. Thus, the RF filter doesn't effectively filter any spurious transmitter output near the edges of the band. If the signal power is limited at the modulator to ensure good linearity, the amount of RF gain can be considerable, resulting in broadband noise 20–30 MHz from the nominal carrier exceeding regulatory limits for out-of-band radiation. Broadband noise from the DACs or mixers is mainly a problem for high-output-power (1 W for U.S. ISM band) access points. (Broadband noise is also a major problem for licensed radiators such as cellular telephone base stations, where powers of 10s or even 100s of watts per channel are encountered, and multiple channels may be transmitted with a single radio to reduce component costs.)

Direct conversion receivers present unique challenges to the designer. Because there is no IF gain, the converted signal into the baseband section may be very small: for example, −50 dBm into 200 Ω is 1 mV. If any DC offset is present, it may grow large enough after further amplification to drive the downstream amplifiers to saturation, swamping the tiny wanted signal. Recall (Figure 3.20) that second-order distortion creates IM products at DC. Thus, DC offsets can arise in the radio from second-order distortion of the large LO signal. Offsets can also result from the LO signal escaping into the receive chain due to the finite LO-RF isolation of the mixer (Figure 3.74); the LO signal may reflect from the filter or antenna back to the mixer, where it is mixed down to DC. These reflections can be time dependent because they depend on the near-antenna environment.

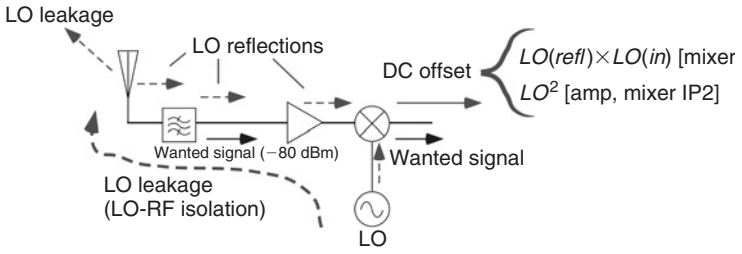


Figure 3.74: Sources of DC Offset in a Direct Conversion Receiver

If the signal spectrum is predominantly away from DC, a simple solution to the DC offset problem is to filter the DC out of the signal. For example, the 802.11a and g OFDM signals do not use the innermost pair of subcarriers. A filter that removes DC but passes the third subcarrier frequency would be sufficient to eliminate DC offsets without affecting the signal. However, this approach cannot be used in protocols like Bluetooth, where significant signal power is at DC and very low frequencies. A second approach is to engineer the radio for minimal offset. The mixer must be designed for very good isolation, the amplifiers and filters carefully matched, and all components must have a high second-order intercept. Such an approach is effective but expensive in component cost and power consumption. Finally, the offset may be measured (for example, during the packet preamble) and corrected with an intentional offset from a DAC. Active calibration and correction is becoming more common as digital capability of low-cost radio and baseband chips increases.

Let us conclude with a few general remarks. In deciding how to implement a radio, it is important to note that *analog does not scale*: 1 W is 1 W no matter how small the gate length of the transistor amplifying it. This is in strong contradistinction to digital circuitry, where reductions in lithographic dimensions produce simultaneous increases in speed, density, and reduction in power consumption. Thus, the advantages of integrating many functions in a single chip are much reduced in analog systems compared with digital systems. On the other hand, if we construct a radio of discrete components, we are free to choose the best technology for each application: GaAs pHEMT for the LNA, SiGe or InGaP bipolar transistors for gain and linearity, GaAs for switches and power amplifiers. Discrete passive components also offer significant performance advantages: wirewound inductors can achieve Q of 30 at 2 GHz and don't consume expensive semiconductor area. Very broadband baluns can be fabricated in this fashion. Integrated implementations are ideal for complex functions where branch matching is important. Examples are IRMs, direct modulators, and differential amplifiers with high IP2.

Integrated radio chip design involves many trade-offs related to what to put on the chip and how to get to external components. For example, linearity of an amplifier can be improved by placing an inductor between the source of a transistor and the ground connection, where it provides positive feedback without consuming any DC voltage (a key issue for chips that

need to operate from 3.3 V or less). On-chip inductors take a lot of space (see Figures 3.84 and 3.85 later in this chapter) and are lossy. A wire bond can be used as an inductor with better performance and no chip area, but a bond pad is required to get to the inductor, and bond pads can be in short supply, considering the many signals that need to get on and off the chip. The inductor can be eliminated, but then the degraded linearity of the amplifiers must be dealt with, by increasing the size of the amplifier transistors and thus their power consumption, redesigning other parts of the radio chain, or accepting reduced radio performance.

Real radios are a constantly evolving compromise of size, performance, and cost. Today's WLAN radios use an integrated radio chip for core analog functions, which may or may not include the analog-to-digital conversion. A separate predominantly or purely baseband chip is usually though not always assigned the complex tasks of managing the PHY and MAC functions of the protocol. The chips are placed in surface-mount packages and combined with inexpensive, high-performance, discrete components, such as inductors, capacitors, balun transformers, RF switches, and power amplifiers. The components are all mounted with a reflowed solder onto a composite fiberglass-plastic circuit board, which may also contain antennas or antenna connectors for the RF interface, and a bus or wired protocol such as Ethernet for communicating with the host system.

3.4 Examples of Radio Chips and Chipsets

Radios for WLANs have been undergoing rapid evolution over the last few years. The Harris/Intersil Prism product line (now owned by Globespan/Virata) was a very popular 802.11b chipset, appearing in products from companies such as D-Link and Linksys. The Prism 2 product line featured separate chips for the RF converter, IF converter, baseband including the ADC/DAC and MAC, VCOs, and power amplifier (Figure 3.75). The RF converter and IF

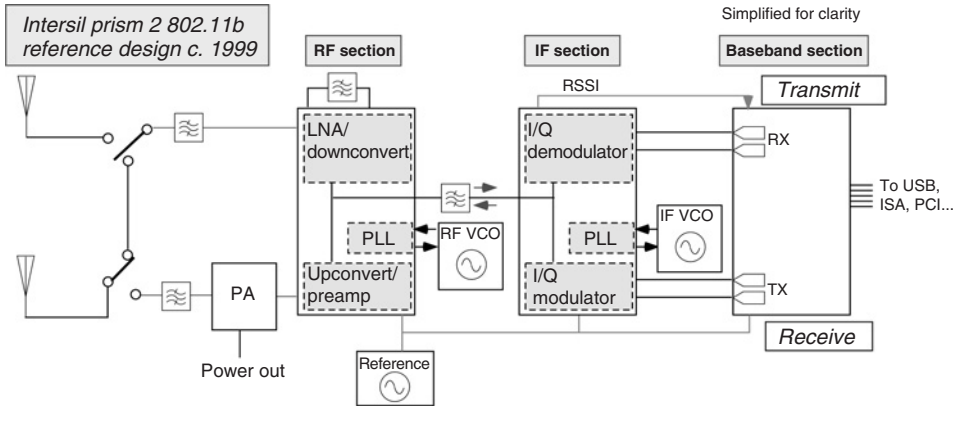


Figure 3.75: Simplified Prism 2 Block Diagram Based on Published Datasheets and Reference Design

converter chips incorporated phase-locked loops to control the VCO frequency. Switches were also external to the chipset. Thus, as many as eight analog active components (not including the reference oscillator) and three SAW filters were required.

A similarly simplified view of a radio based on a more recent Broadcom 802.11g chipset is shown in Figure 3.76. This radio chipset is also capable of dual-band operation using a second radio chip; Figure 3.76 depicts only ISM band operation for clarity. This radio design uses only a single radio chip, which incorporates the complete synthesizer function as well as down-convert/demodulate and direct modulation. Only four active analog components and a single filter are required.

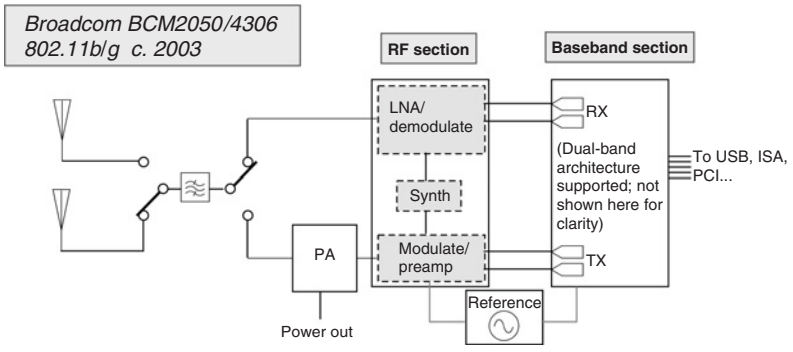


Figure 3.76: Simplified 802.11b/g Radio Block Diagram Using Direct Conversion Chipset

Let us take a closer look at a couple of representative WLAN chipsets. The first is a relatively early 802.11a design from Atheros, described at the 2002 International Solid-State Circuits Conference (ISSCC). A simplified functional diagram of the radio chip is provided in Figure 3.77. The design is a high-IF superhet radio using a 1 GHz IF to make image filtering very easy: simple discrete filters can be used to extract the wanted 5.2-GHz signal from the image at 3.2 GHz. Gain adjustment is provided in the RF, IF, and baseband sections. Active compensation of DC offsets and branch mismatch is provided. The transmit function uses an interesting stacked image-reject up-converting mixer to minimize filtering requirements; an off-chip balun is required to convert the output differential signal to a single-ended signal on the circuit board. An on-chip integer-N synthesizer provides the LO signal for the RF conversions. Digital-analog conversion uses 9-bit ADCs and DACs and oversampling.

A typical transmitted signal is shown in Figure 3.78. The baseband signal is a 6-MHz (BPSK modulation on the subcarriers) signal, which has the highest peak-to-average ratio and is most demanding of the linearity of the amplifiers. The chip is able to produce 17 dBm out while remaining compliant with the 802.11a output mask; this power level may be sufficient

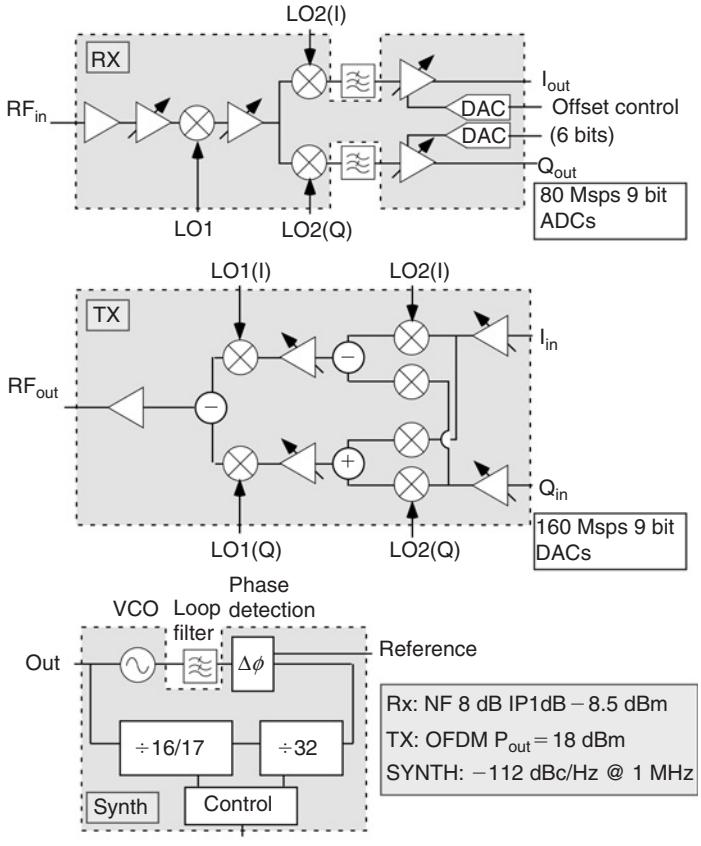


Figure 3.77: Simplified Block Diagram of 802.11a Radio Chip (After Su et al., ISSCC 2002)

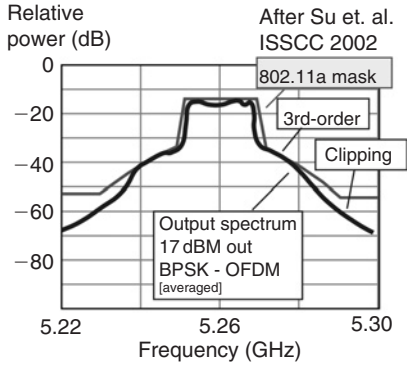


Figure 3.78: Example of 802.11a OFDM Output Spectrum (After Su et al.)

for a client card application without an external power amplifier. The price of this very high output power is substantial power consumption: the radio chip consumes 0.8 W in transmit mode.

A more recent radio chip, supporting 802.11b and g, was described by Trachewsky et al. of Broadcom at the 2003 IEEE Hot Chips Conference. A simplified block diagram of the chip is shown in Figure 3.79. This chip uses direct conversion on both receive and transmit. Channel filtering is thus simple low-pass filtering; all the channel filters are active filters whose cutoff frequency can be adjusted to optimize signal fidelity. Two received signal strength indicators are provided, before and after channel filtering, to aid in detection of interferers. Gain adjustment at both RF and baseband is provided; active compensation and calibration is used during idle periods to remove offsets and ensure good I/Q matching.

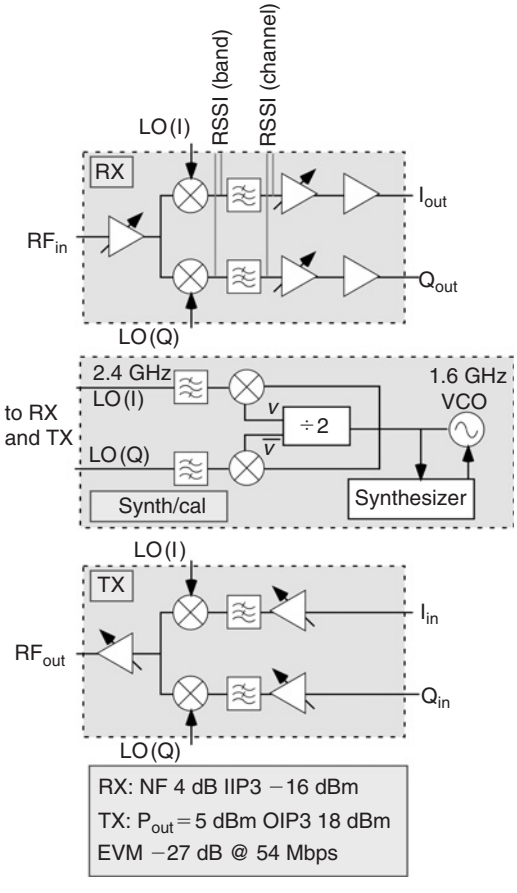


Figure 3.79: Simplified Block Diagram of 802.11b/g Radio Chip (After Trachewsky et al., Hot Chips 2003)

The on-chip synthesizer uses a VCO at two-thirds of the final frequency: the VCO output is divided by 2, and the two complementary outputs of the divider are mixed with the original VCO signal to produce I and Q synthesizer outputs. This displacement of the VCO frequency from the intended RF frequency is a common technique used to ensure that the VCO frequency is not displaced or pulled away from its intended value by exposure to a leaked high-amplitude RF signal, such as a fraction of the transmitted signal that may be conducted through the substrate. Receive noise figure is excellent; IP3 is adequate for typical WLAN applications. The transmitted power is modest, and thus an external power amplifier is needed for most applications; however, DC power consumption is correspondingly small.

Performance parameters for these and a number of other reported 802.11 chipsets are summarized in Table 3.10. Certain trends are clear: single-chip radios are the rule, direct conversion is becoming increasingly common, and higher ADC resolution and sampling rate are used as more complex signals become common. Note, on the other hand, that no strong trend in RF performance is evident: receive noise figure is within a range of 4–8 dB, and input intercept varies over the range around 10 to 20 dBm appropriate for WLAN applications. The large transmit power variations are primarily a function of whether an external power amplifier is intended as part of the chipset or not and do not reflect any strong trend in performance scaling. The same can be said for synthesizer performance, save that the complete synthesizer function is integrated in the radio chip in modern chipsets, unlike the older separate VCO solutions.

Some recently reported results suggest the direction of future chipset development. At the 2004 ISSCC, Ahola and coworkers described a single-chip radio that supports both the 2.4- and 5-GHz bands. They use a dual-conversion high-IF architecture, with the first fixed LO1 frequency chosen to be roughly between the ISM and UNII band ranges (3840 MHz), so that both bands can be mapped onto about the same IF (1310–1510 MHz) using the same input mixer, saving a bunch of inductors (see Figure 3.84 to get an idea of how much space inductors consume in a chip!). A second but still fixed LO1 of 4320 MHz is used to convert the upper part of the 5-GHz band. The variable second LO2 frequency then directly down-converts the IF signal to baseband. The same approach is used in transmit, although separate output mixers are needed for the two bands. The fixed LO frequencies are relatively easy to provide with low phase noise, and the variable LO is at a reduced frequency where again good performance is achievable. Because of the 2X separation between the input bands, no special image rejection provisions are needed. The radio chip achieves noise figure of about 5.3 dB, though IIP3 is a rather modest 23 to 26 dBm at maximum gain, apparently due to removal of inductors on the input mixer to allow multiband operation. Transmit output easily meets the EVM and spectral mask requirements at around 0 to 3 dBm out. A similar approach using two separate synthesizers to convert the ISM and UNII bands to a 1.8-GHz IF was described by Zargari et al.

Table 3.10: Summary of Reported WLAN Chipset Performance

Vendor	Intersil – Virata	Atheros	Broadcom	Broadcom	Resonext	Marvell	Athena	AMD	Thomson
Part #(s)	Prism 2	? 2002	BCM2050, 4306	BCM 2060, 4306	Unknown	Unknown	Unknown	Unknown	Unknown
Protocols supported	802.11 classic, b	802.11a	802.11b,g	802.11a	802.11a	802.11b	802.11a	802.11b	802.11a
Chips in radio	5?	1	1	1	1	1	1	1	1
Chips in MAC/ baseband	2?	1	1	1		1	1	1	
Radio chip area	?	22 mm^2		11.7 mm^2	13 mm^2	16 mm^2	18.5 mm^2	10 mm^2	17 mm^2
Architecture	Superhet	Superhet	Direct conversion	Direct conversion	Direct conversion	Superhet	Direct conversion	Direct conversion	Superhet dual conv
Technology	0.35 μm SiGe BiCMOS	0.25 μm CMOS		0.18 μm CMOS	0.18 μm CMOS	0.25 μm CMOS	0.18 μm CMOS	0.25 μm CMOS	0.5 μm SiGe BiCMOS
IF (MHz)		1000							1225, 60
TX P1dB		22 dBm		19 dBm			0 dBm		15 dBm
TX OIP3			18 dBm		15 dBm				
TX P (CCK)			5 dBm			20 dBm		0 dBm	
TX P (OFDM)		18 dBm	5 dBm	15 dBm	5 dBm				
TX EVM			−27dB @ 54 Mbps		−28 dB @ 54 Mbps				
RX NF		8 dB	4 dB	4 dB	7 dB		5.5 dB	5 dB	5 dB

RX IP1dB (max gain)		−8.5 dBm					−20 dBm		
RX IIP3 (max gain)			−16 dBm		−18 dBm	−10 dBm	−17 dBm	−8.5 dBm	
RX sensitivity, lowest rate			−97dBm @ 1 Mbps	−94 dBm @ 6 Mbps		−95 dBm@ 1 Mbps		−96 dBm @1 Mbps	
Phase noise		−112 dBc/ Hz @ 1 MHz		−100 dBc/ Hz @ 30 KHz	−110 dBc/ Hz @ 1 MHz	−110 dBc/ Hz @ 1 MHz	−115 dBc/ Hz @ 1 MHz	−111 dBc/ Hz @ 1 MHz	−88 dBc/ Hz @10 KHz
Integrated phase noise					1.5° 10 KHz to 10 MHz		-37dBc 1 KHz- 10 MHz (1.6°?)	1b	
DAC resolution			8b	8b		9b 88 Msps			8b 160 Msps
ADC resolution			8b	8b		6b 44 Msps			8b 80 Msps
DC power: TX		0.8 W	144 mW	380 mW	138 mW	1250 mW	302 mW	290 mW	920 mW
DC power: RX		0.4 W	200 mW	150 mW	171 mW	350 mW	248 mW	322 mW	200 mW
Reference	Published datasheets	Su et al. ISSCC 2002 paper 5.4	Trachewsky et al. HotChips 2003	Trachewsky et al. HotChips 2003	Zhang et al. ISSCC 2003 paper 20.3	Chien et al. ISSCC 2003 paper 20.5	Bouras et al. ISSCC 2003 paper 20.2	Kluge et al. ISSCC 2003 paper 20.6	Schwanen-berger ISSCC et al. 2003 paper 20.1

A different approach to a dual-band chip was reported by Perraud and coworkers at this meeting. They used a tunable 9.6- to 11.8-GHz LO, which was then divided either by 2 or by 4 to produce the I and Q signals for direct conversion of the 5-GHz or 2.4-GHz signals. This chip uses a distributed active transformer output amplifier (see section 6) to achieve as much as 12 dBm output from a 1.8-V supply. A Cartesian feedback loop, a linearization technique heretofore only common in high-power amplifiers used for cellular telephone base stations, is used to actively correct distortion in the transmit power amplifier chain, increasing the effective third-order intercept by about 6 dB. The chip achieves less than 5 dB noise figure and good EVM and spectral mask performance.

These reports suggest that it is feasible to create single-chip dual-band radios. Therefore, it seems likely that within 1–2 years, dual-band tri-mode (802.11a/b/g) clients will be common at a cost comparable with today's single-band equipment, at which point the choice of band and protocol will be to some extent transparent to the system user.

IEEE 802.11 chipsets are the most common WLAN technology currently shipping, but by no means the only one. It is worth contrasting a representative Bluetooth design with the 802.11 chips we have examined. Recall that Bluetooth is a low-power low–data rate protocol with a simple Gaussian minimum-shift keying modulation. A simplified block diagram of the chip reported by Chang et al. at ISSCC 2002 is shown in Figure 3.80. The chip is a direct-conversion architecture using an on-chip synthesizer; it is implemented in a silicon-on-insulator technology, which provides good isolation between functional elements on the chip, simplifying such tasks as integrating the synthesizer. The design assumes an external matching circuit and balun to connect the differential interface of the chip to the single-ended antenna.

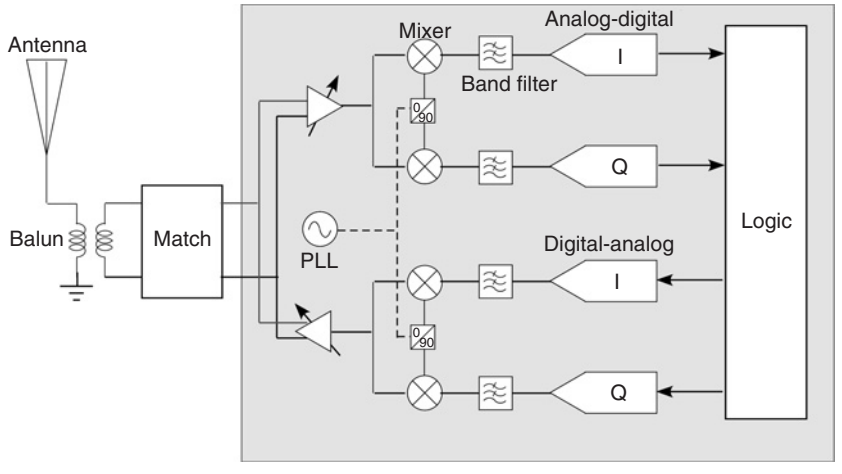


Figure 3.80: Simplified Block Diagram of Direct Conversion Bluetooth Radio Chip (After Chang et al., ISSCC 2002)

Recall that Bluetooth's spectrum peaks at zero frequency; DC offsets cannot be solved by filtering in this protocol. Instead, the design uses a wide-dynamic-range ADC to allow some digital DC offset correction. The external balun allows a fully differential chip design with excellent second-order distortion properties. The input second-order intercept IIP2 = 40 dBm. The isolation provided by the insulating substrate minimizes leakage of the LO where it isn't wanted, reducing DC offset generation (Figure 3.74).

The mixers on this chip, like many of the WLAN chips, are implemented as Gilbert cells; a simplified schematic is shown in Figure 3.81. The receive side mixers substitute inductors for the resistors to minimize noise generation. An image-reject topology is not needed because the direct-conversion architecture has no image frequency.

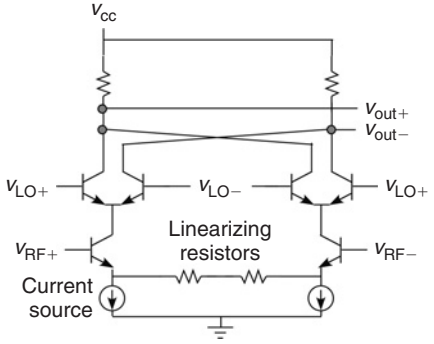


Figure 3.81: Gilbert Cell Mixer Topology (After Chang et al.)

Channel filtering is performed digitally. A Σ-Δ converter with 80 dB of dynamic range at 32 Msps is used in this case. Performance of the chip is summarized in Table 3.11. Noise performance is much better than required by the standard. Power consumption is reduced from that required for the 802.11 radio chips, albeit modestly; note, however, that ADC and DAC power are included here. Rejection of the adjacent channel may appear to be modest,

Table 3.11: Silicon Wave Bluetooth Radio Performance

Parameter	Value
Maximum TX output	3 dBm
Adjacent-channel rejection	−5 dB
RX noise figure	5 dB
RX IIP2	40 dBm
DC power: TX	0.12 W
DC power: RX	0.12 W

but recall that the narrow 1-MHz Bluetooth channels directly abut one another without guard bands, and that the frequency-hopping Bluetooth architecture means that adjacent-channel interference is generally transitory.

A completed radio is more than just a few silicon chips. Inexpensive radios of this type are generally fabricated on a board fabricated of a common fiberglass/plastic composite such as FR4. Most boards have four metal layers: RF ground, RF signal, DC power, and digital data. Along with the key radio chips, the boards carry on the order of 100 surface-mount discrete components. Surface-mount resistors, inductors, and capacitors typically cost from $0.01 to $0.10 each; SAW filters are from roughly $0.50 to $5 each. Power amplifiers and switches are also on the order of $1 to $2 each. A board for a client radio (a *network interface card*) will generally include one or two antennas mounted on the board; an access point will incorporate one or two spring-loaded connectors to allow cabling to remote antennas. Some laptop computers also have built-in antennas cabled to the remote radio card. Some interface to the host or network is also provided. Host interfaces, such as the PC-card bus or universal serial bus (USB), are generally built into the MAC/baseband chip; an Ethernet interface will use a separate chip.

Connections on radio boards are made with plated copper lines a few microns thick and some tens of microns wide. The loss of these lines is critical in ensuring good overall radio performance: losses between the antennas and LNA add directly to the receive noise figure, and losses between antennas and the power amplifier subtract directly from transmitted power. FR4 is the most common substrate used in board manufacture, but it is quite lossy at 5 GHz. A typical 50 Ω transmission line on FR4 has a loss of about 0.2 dB/cm at 5 GHz, so keeping lines short helps. However, on-board antennas need to be reasonably isolated from the rest of the circuit to avoid strong coupling between the antennas and circuit lines, setting a lower limit on the line lengths that can be used in this case. Unintended radiation must also be minimized: lines carrying a high-power LO signal may couple with other lines or radiate, in the case where the VCO is separate from the radio chip.

Some examples of recent radio boards are shown in Figures 3.82 and 3.83. (Note the names are trademarks of the respective vendors.) Figure 3.82 depicts an Airport Xtreme card used in an Apple Computer client card, purchased in 4Q 2003. This card uses a Broadcom 802.11b/g chipset. The high-frequency components are contained within sheet-metal enclosures (the tops have been removed to make the components visible) that ensure good shielding of the receiver from external interference. A separate enclosure is provided for the power amplifier to minimize unwanted coupling of the high-level signal back to the radio chip. The crystal reference oscillator is contained within the main shielding enclosure. Roughly 100 passive components support the operation of the active radio. The use of an off-board antenna and the shielded enclosures allow the connection to the antenna to be kept short.

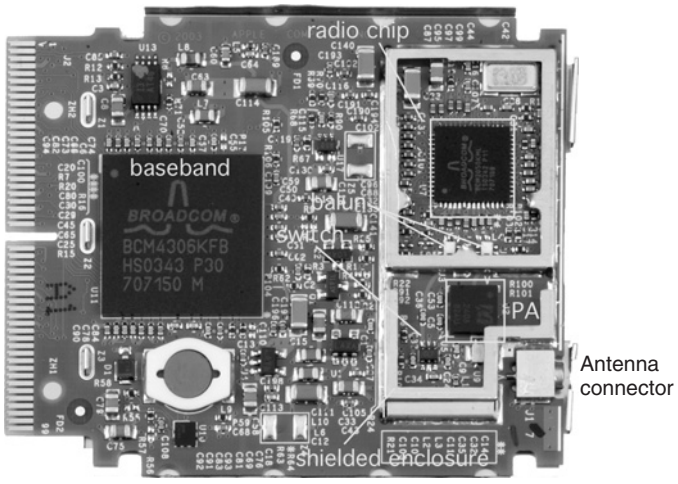


Figure 3.82: Apple Airport Xtreme 802.11b/g Radio Card; Plastic Covers and Shielded Enclosure Covers Removed (c. 12/03; Photo by Chuck Koehler)

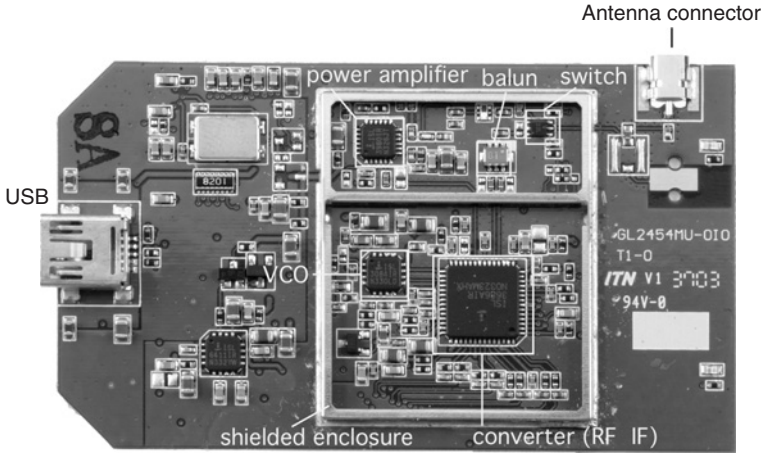


Figure 3.83: D-Link DWLG120 802.11b/g Radio Card; Plastic Covers and Shielded Enclosure Covers Removed; Backside Not Shown (c. 12/03; Photo by Chuck Koehler)

Figure 3.83 depicts a D-Link DWL G120 802.11b/g USB radio. Again, sheet-metal enclosures surround the radio chip and the front end containing the power amplifier. A Prism chipset is used (descended from Harris Semiconductor, then Intersil, through Globespan/Virata and part of Conexant at the time of this writing). A superhet architecture is used, with separate chips to convert between IF/RF and IF/ baseband. To keep the overall structure small, a two-sided board is used; the back side (not shown) contains the modulator and baseband/MAC functions. The radio and front end are again placed in shielded enclosures. The reference oscillator

is placed external to the shielded enclosure. Again, a connector to an external antenna is provided, so the transmission line to the antenna is short.

These two boards represent two different approaches to minimization of size and cost. The Apple board uses a small parts count direct-conversion radio, whereas the D-Link board uses a well-known superhet radio requiring more parts but with a compact dual-sided packaging arrangement. There's more than one way to accomplish a single task.

In Figure 3.84, we show the radio chip (BCM2050; see also Figure 3.79) from the Airport Xtreme card after removal of the plastic encapsulation. Recall that this is a direct-conversion architecture, so the whole radio function is contained in this chip. It is readily apparent that a great deal of the chip area is taken up by inductors and that although the chip is functionally analog, in practice this is a mixed-signal design with extensive logic area.

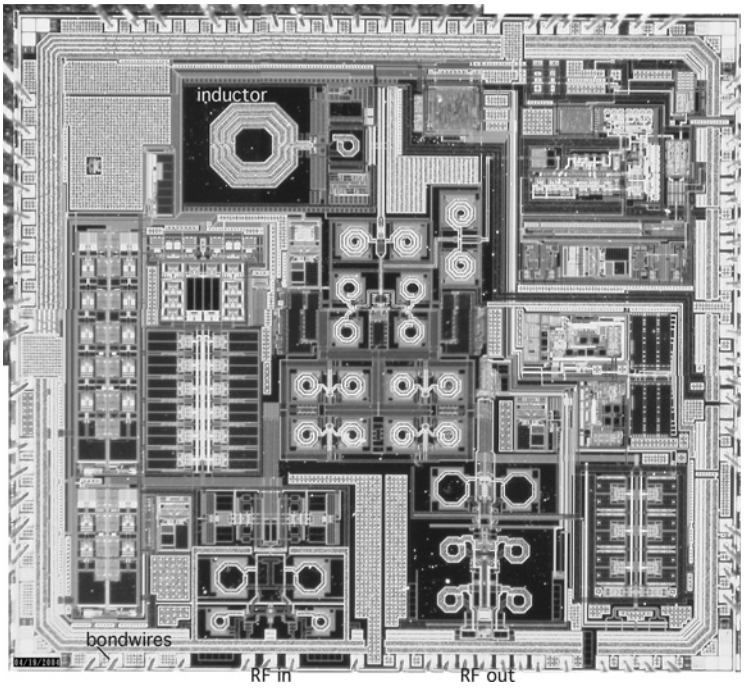


Figure 3.84: Broadcom BCM2050 From Apple Airport Xtreme Card, After Removal Of Plastic Encapsulation; Montage of Four Quadrant Photos (Image Courtesy of WJ Communications)

In Figure 3.85, the converter chip from the D-Link card shown in Figure 3.83 is similarly depicted after decapsulation. Inductors still consume significant area, though the large areas of logic show even more clearly the mixed-signal requirements of a modern WLAN radio chip.

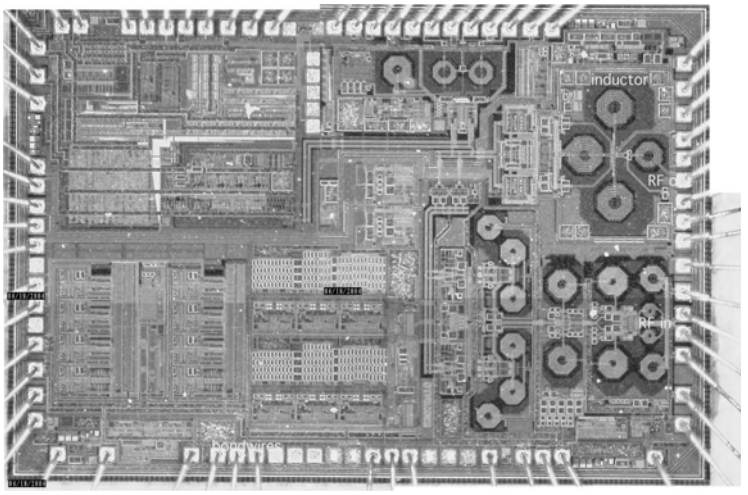


Figure 3.85: ISL3689 From D-Link USB Card, After Removal of Plastic Encapsulation; Montage of Four Quadrant Photos (Image Courtesy of WJ Communications)

3.5 Summary

Radios must detect signals of widely varying magnitude in the presence of interferers. A radio is required to minimize excess noise while tolerating large input signals (both wanted and interfering signals) without excessive distortion. Transmit distortion must be minimized to avoid spurious radiation outside the intended channel and band. Radios must provide good frequency selectivity and tunability and adapt to a huge range of input power.

Modern digital radios combine frequency conversion with analog-to-digital conversions. There are three popular architectures: superhet, NZIF, and direct conversion, each with its own advantages and pitfalls. Each analog component has certain key performance parameters. Amplifiers are characterized by gain, distortion, and noise. Mixers add isolation to distortion and noise. Synthesizers must deliver low phase noise and good frequency stability. Filters require narrow bandwidths, low insertion loss, and good rejection of unwanted signals. Switches need good insertion loss, fast actuation, and sufficient isolation. Radio boards are a mixture of discrete and integrated analog and digital components. The final radio chain is a compromise between gain, noise, distortion, DC power consumption, and cost of component acquisition and radio manufacture.

3.6 Further Reading RFIC

Design

The Design of CMOS Radio-Frequency Integrated Circuits, Thomas Lee, Cambridge, 1998: An encyclopedic introduction to the design of radio components, though the emphasis is

much broader than purely CMOS implementation (which was probably added to the title to increase sales). Includes treatments of synthesizer operation, oscillator phase noise, and feedback design.

Analog-to-Digital Conversion

"Delta-Sigma Data Conversion in Wireless Transceivers," Ian Galton, IEEE Transactions on Microwave Theory and Techniques, vol. 50, #1, p. 302 (2002)

"Analog-to-Digital Converter Survey and Analysis," R. Walden, IEEE Journal on Selected Areas in Communications, vol. 17, #4, p. 539 (1999)

Amplifiers

RF Power Amplifiers for Wireless Communications, Steve C. Cripps, Artech House, 1999: Cripps is bright, opinionated, and brings extensive practical experience to bear on abstruse topics in amplifier design.

Design of Amplifiers and Oscillators by the S-Parameter Method, George Vendelin, Wiley Interscience, 1982: Purely microwave-oriented, antedating modern CMOS and SiGe devices, but a useful reference and introduction to matching techniques, low-noise and broadband design.

"A Fully Integrated Integrated 1.9-GHz CMOS Low-Noise Amplifier," C. Kim et al., IEEE Microwave and Guided Wave Letters, vol. 8, #8, p. 293 (1998)

"On the Use of Multitone Techniques for Assessing RF Component's Intermodulation Distortion," J. Pedro and N. de Carvalho, IEEE Transactions on Microwave Theory and Techniques, vol . 47, p. 2393 (1999)

"Impact of Front-End Non-Idealities on Bit Error Rate Performance of WLANOFDM Transceivers," B. Côme et al., RAWCON 2000, p. 91

"Weigh Amplifier Dynamic-Range Requirements," D. Dobkin (that's me!), Walter Strifler, and Gleb Klimovitch, Microwaves and RF, December 2001, p. 59

Mixers

A great deal of useful introductory material on mixers was published over the course of about 15 years by Watkins-Johnson Company as TechNotes. These have been rescued from oblivion (in part by the current author) and are available on the web site of WJ Communications, Inc., www.wj.com. The material is focused on diode mixers but many issues are generic to all mixer designs. Of particular interest are the following:

"Mixers, Part 1: Characteristics and Performance," Bert Henderson, volume 8

"Mixers, Part 2: Theory and Technology," Bert Henderson, volume 8

"Predicting Intermodulation Suppression in Double-Balanced Mixers," Bert Henderson, volume 10

"Image-Reject and Single-Sideband Mixers," Bert Henderson and James Cook, volume 12

"Mixers in Microwave Systems, Part 1," Bert Henderson, volume 17

Switches

"An Integrated 5.2 GHz CMOS T/R Switch with LC-Tuned Substrate Bias,"

N. Talwalker, C. Yue, and S. Wong, International Solid-State Circuits Conference 2003, paper 20.7, p. 362

Chipsets

"An Integrated 802.11a Baseband and MAC Processor," J. Thomson et al., International Solid-State Circuits Conference 2002, paper 7.2

"A 5 GHz CMOS Transceiver for IEEE 802.11a Wireless LAN," D. Su et al., "An Integrated 802.11a Baseband and MAC Processor," J. Thomson et al. International Solid-State Circuits Conference 2002, paper 5.4

"Broadcom WLAN Chipset for 802.11a/b/g," J. Trachewsky et al., IEEE Hotchips Conference, Stanford University, 2003

"Direct-Conversion CMOS Transceiver with Automatic Frequency Control for 802.11a Wireless LANs," A. Behzad et al., International Solid-State Circuits Conference 2003, paper 20.4, p. 356

"A Multi-Standard Single-Chip Transceiver covering 5.15 to 5.85 GHz," T. Schwanenberger et al., International Solid-State Circuits Conference 2003, paper 20.1, p. 350

"A Digitally Calibrated 5.15–5.825 GHz Transceiver for 802.11a Wireless LANs in 0:18 mm CMOS," I. Bouras et al., International Solid-State Circuits Conference 2003, paper 20.2, p. 352

"A Direct Conversion CMOS Transceiver for IEEE 802.11a WLANs," P. Zhang et al., International Solid-State Circuits Conference 2003, paper 20.3, p. 354

"A 2.4 GHz CMOS Transceiver and Baseband Processor Chipset for 802.11b Wireless LAN Application," G. Chien et al., International Solid-State Circuits Conference 2003, paper 20.5, p. 358

"A Direct-Conversion Single-Chip Radio-Modem for Bluetooth," G. Chang et al., International Solid-State Circuits Conference 2002, paper 5.2

"A Single Chip CMOS Transceiver for 802.11a/b/g WLANs," R. Ahola et al., International Solid-State Circuits Conference 2004, paper 5.2, p. 64

"A Dual-Band 802.11a/b/g Radio in 0.18 mm CMOS," L. Perraud et al., International Solid-State Circuits Conference 2004, paper 5.3, p. 94

"A Single-Chip Dual-Band Tri-Mode CMOS Transceiver for IEEE 802.11a/b/g WLAN," M. Zargari et al., International Solid-State Circuits Conference 2004, paper 5.4, p. 96

Distributed Active Transformers

"Fully Integrated CMOS Power Amplifier Design Using the Distributed Active-Transformer Architecture," I. Aoki, S., Kee, D. Rutledge, and A. Hajimiri, IEEE J. Solid-State Circuits, vol. 37, # 3, p. 371 (2002)

CHAPTER 4

Radio Propagation

Alan Bensky

It is fitting to begin a book about wireless communication with a look at the phenomena that lets us transfer information from one point to another without any physical medium—the propagation of radio waves. If you want to design an efficient radio communication system, even for operation over relatively short distances, you should understand the behavior of the wireless channel in the various surroundings where this communication is to take place. While the use of "brute force"—increasing transmission power—could overcome inordinate path losses, limitations imposed on design by required battery life, or by regulatory authorities, make it imperative to develop and deploy short-range radio systems using solutions that a knowledge of radio propagation can give.

The overall behavior of radio waves is described by Maxwell's equations. In 1873, the British physicist James Clerk Maxwell published his *Treatise on Electricity and Magnetism* in which he presented a set of equations that describe the nature of electromagnetic fields in terms of space and time. Heinrich Rudolph Hertz performed experiments to confirm Maxwell's theory, which led to the development of wireless telegraph and radio. Maxwell's equations form the basis for describing the propagation of radio waves in space, as well as the nature of varying electric and magnetic fields in conducting and insulating materials, and the flow of waves in waveguides. From them, you can derive the skin effect equation and the electric and magnetic field relationships very close to antennas of all kinds. A number of computer programs on the market, based on the solution of Maxwell's equations, help in the design of antennas, anticipate electromagnetic radiation problems from circuit board layouts, calculate the effectiveness of shielding, and perform accurate simulation of ultrahigh-frequency and microwave circuits. While you don't have to be an expert in Maxwell's equations to use these programs (you do in order to write them!), having some familiarity with the equations may take the mystery out of the operation of the software and give an appreciation for its range of application and limitations.

4.1 Mechanisms of Radio Wave Propagation

Radio waves can propagate from transmitter to receiver in four ways: through ground waves, sky waves, free space waves, and open field waves.

Ground waves exist only for vertical polarization, produced by vertical antennas, when the transmitting and receiving antennas are close to the surface of the earth (see *Polarization*

under Section 5.2 in Chapter 5). The transmitted radiation induces currents in the earth, and the waves travel over the earth's surface, being attenuated according to the energy absorbed by the conducting earth. The reason that horizontal antennas are not effective for ground wave propagation is that the horizontal electric field that they create is short circuited by the earth. Ground wave propagation is dominant only at relatively low frequencies, up to a few megahertz, so it needn't concern us here.

Sky wave propagation is dependent on reflection from the ionosphere, a region of rarified air high above the earth's surface that is ionized by sunlight (primarily ultraviolet radiation). The ionosphere is responsible for long-distance communication in the high-frequency bands between 3 and 30 MHz. It is very dependent on time of day, season, longitude on the earth, and the multi-year cyclic production of sunspots on the sun. It makes possible long-range communication using very low-power transmitters. Most short-range communication applications that we deal with in this book use VHF, UHF, and microwave bands, generally above 40 MHz. There are times when ionospheric reflection occurs at the low end of this range, and then sky wave propagation can be responsible for interference from signals originating hundreds of kilometers away. However, in general, sky wave propagation does not affect the short-range radio applications that we are interested in.

The most important propagation mechanism for short-range communication on the VHF and UHF bands is that which occurs in an open field, where the received signal is a vector sum of a direct line-of-sight signal and a signal from the same source that is reflected off the earth. We discuss below the relationship between signal strength and range in line-of-sight and open-field topographies.

The range of line-of-sight signals, when there are no reflections from the earth or ionosphere, is a function of the dispersion of the waves from the transmitter antenna. In this free-space case the signal strength decreases in inverse proportion to the distance away from the transmitter antenna. When the radiated power is known, the field strength is given by equation [4.1]:

$$E = \frac{\sqrt{30 \cdot P_t \cdot G_t}}{d} \tag{4.1}$$

where P_t is the transmitted power, G_t is the antenna gain, and d is the distance. When P_t is in watts and d is in meters, E is volts/meter.

To find the power at the receiver (P_r) when the power into the transmitter antenna is known, use [4.2]:

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2} \tag{4.2}$$

G_t and G_r are the transmitter and receiver antenna gains, and λ is the wavelength.

Range can be calculated on this basis at high UHF and microwave frequencies when high-gain antennas are used, located many wavelengths above the ground. Signal strength between the earth and a satellite, and between satellites, also follows the inverse distance law, but this case isn't in the category of short-range communication! At microwave frequencies, signal strength is also reduced by atmospheric absorption caused by water vapor and other gases that constitute the air.

4.2 Open Field Propagation

Although the formulas in the previous section are useful in some circumstances, the actual range of a VHF or UHF signal is affected by reflections from the ground and surrounding objects. The path lengths of the reflected signals differ from that of the line-of-sight signal, so the receiver sees a combined signal with components having different amplitudes and phases. The reflection causes a phase reversal. A reflected signal having a path length exceeding the line-of-sight distance by exactly the signal wavelength or a multiple of it will almost cancel completely the desired signal ("almost" because its amplitude will be slightly less than the direct signal amplitude). On the other hand, if the path length of the reflected signal differs exactly by an odd multiple of half the wavelength, the total signal will be strengthened by "almost" two times the free space direct signal.

In an open field with flat terrain there will be no reflections except the unavoidable one from the ground. It is instructive and useful to examine in depth the field strength versus distance in this case. The mathematical details are given in the Mathcad worksheet "Open Field Range."

In Figure 4.1 we see transmitter and receiver antennas separated by distance d and situated at heights h_1 and h_2. Using trigonometry, we can find the line of sight and reflected signal path lengths d_1 and d_2. Just as in optics, the angle of incidence equals the angle of reflection θ. We get the relative strength of the direct signal and reflected signal using the inverse path length relationship. If the ground were a perfect mirror, the relative reflected signal strength would exactly equal the inverse of d_2. In this case, the reflected signal phase would shift 180 degrees at the point of reflection. However, the ground is not a perfect reflector. Its characteristics as a reflector depend on its conductivity, permittivity, the polarization of the signal and its angle of incidence. In the Mathcad worksheet we have accounted for polarization, angle of incidence,

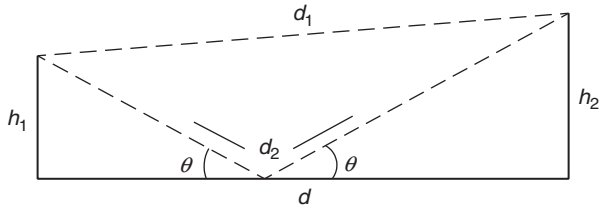


Figure 4.1: Open Field Signal Paths

and permittivity to find the reflection coefficient, which approaches −1 as the distance from the transmitter increases. The signals reaching the receiver are represented as complex numbers since they have both phase and amplitude. The phase is found by subtracting the largest interval of whole wavelength multiples from the total path length and multiplying the remaining fraction of a wavelength by 2π radians, or 360 degrees.

Figure 4.2 gives a plot of relative open field signal strength versus distance using the following parameters:

Polarity—horizontal

Frequency—300 MHz

Antenna heights—both 3 meters

Relative ground permittivity—15

Also shown is a plot of free space field strength versus distance (dotted line). In both plots, signal strength is referenced to the free space field strength at a range of 3 meters.

Notice in Figure 4.2 that, up to a range of around 50 meters, there are several sharp depressions of field strength, but the signal strength is mostly higher than it would be in free space. Beyond 100 meters, signal strength decreases more rapidly than for the free space

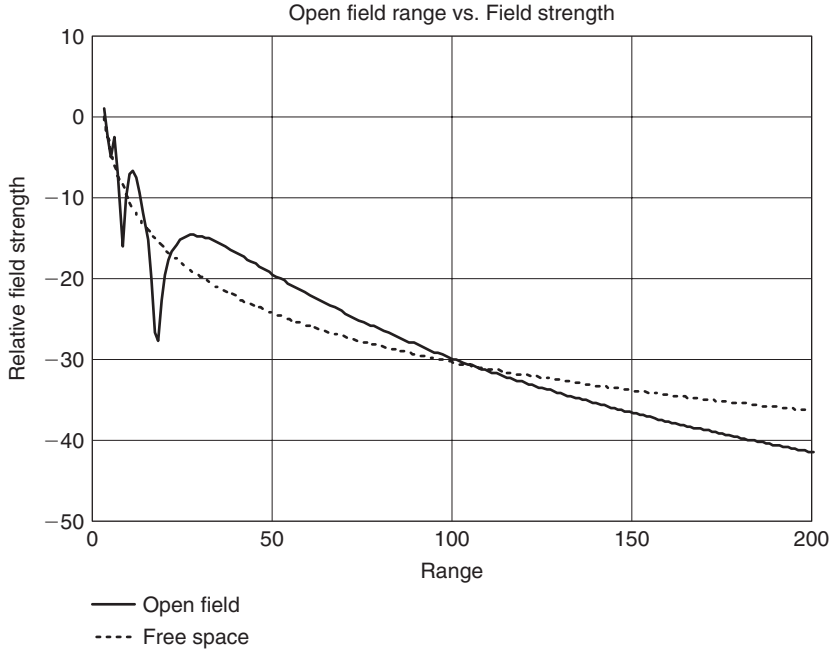


Figure 4.2: Field Strength vs. Range at 300 MHz

model. Whereas there is an inverse distance law for free space, in the open field beyond 100 meters (for these parameters) the signal strength follows an inverse square law. Increasing the antenna heights extends the distance at which the inverse square law starts to take effect. This distance, d_m, can be approximated by

$$d_m = (12 \times h_1 \times h_2)/\lambda \tag{4.3}$$

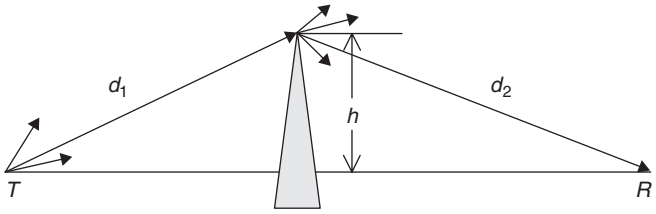
where h_1 and h_2 are the transmitting and receiving antenna heights above ground and λ is the wavelength, all in the same units as the distance d_m.

In plotting Figure 4.2, we assumed horizontal polarization. Both antenna heights, h_1 and h_2, are 3 meters. When vertical polarization is used, the extreme local variations of signal strengths up to around 50 meters are reduced, because the ground reflection coefficient is less at larger reflection angles. However, for both polarizations, the inverse square law comes into effect at approximately the same distance. This distance in Figure 4.2 where λ is 1 meter is, from equation [4.3]: $d_m = (12 \times 3 \times 3)/\lambda = 108$ meters. In Figure 4.2 we see that this is approximately the distance where the open-field field strength falls below the free-space field strength.

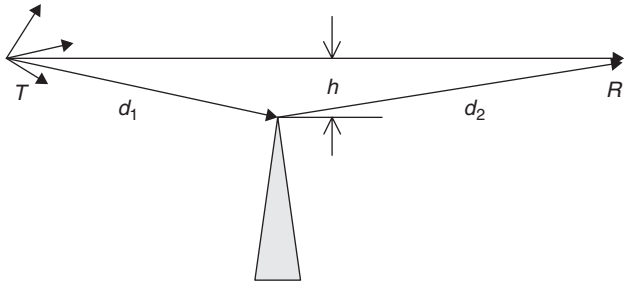
4.3 Diffraction

Diffraction is a propagation mechanism that permits wireless communication between points where there is no line-of-sight path due to obstacles that are opaque to radio waves. For example, diffraction makes it possible to communicate around the earth's curvature, as well as beyond hills and obstructions. It also fills in the spaces around obstacles when short-range radio is used inside buildings. Figure 4.3 is an illustration of diffraction geometries, showing an obstacle whose extremity has the shape of a knife edge. The obstacle should be seen as a half plane whose dimension is infinite into and out of the paper. The field strength at a receiving point relative to the free-space field strength without the obstacle is the diffraction gain. The phenomenon of diffraction is due to the fact that each point on the wave front emanating from the transmitter is a source of a secondary wave emission. Thus, at the knife edge of the obstacle, as shown in Figure 4.3a, there is radiation in all directions, including into the shadow.

The diffraction gain depends in a rather complicated way on a parameter that is a function of transmitter and receiver distances from the obstacle, d_1 and d_2, the obstacle dimension h, and the wavelength. Think of the effect of diffraction in an open space in a building where a wide metal barrier extending from floor to ceiling exists between the transmitter and the receiver. In our example, the space is 12 meters wide and the barrier is 6 meters wide, extending to the right side. When the transmitter and receiver locations are fixed on a line at a right angle to the barrier, the field strength at the receiver depends on the perpendicular distance from the line-of-sight path to the barrier's edge. Figure 4.4 is a plot of diffraction gain when transmitter



a) *R* is in shadow—*h* is positive



b) *T* and *R* are line of sight—*h* is negative

Figure 4.3: Knife-edge Diffraction Geometry

and receiver are each 10 meters from the edge of the obstruction and on either side of it. The dimension "*h*" varies between −6 meters and 6 meters—that is, from the left side of the space where the dimension "*h*" is considered negative, to the right side where "*h*" is positive and fully in the shadow of the barrier. Transmission frequency for the plot is 300 MHz. Note that the barrier affects the received signal strength even when there is a clear line of sight between the transmitter and receiver ("*h*" is negative as shown in Figure 4.3b). When the barrier edge is on the line of sight, diffraction gain is approximately −6 dB, and as the line-of-sight path gets farther from the barrier (to the left in this example), the signal strength varies in a cyclic manner around 0 dB gain. As the path from transmitter to receiver gets farther from the barrier edge into the shadow, the signal attenuation increases progressively.

Admittedly, the situation depicted in Figure 4.4 is idealistic, since it deals with only one barrier of very large extent. Normally there are several partitions and other obstacles near or between the line of sight path and a calculation of the diffraction gain would be very complicated, if not impossible. However, a knowledge of the idealistic behavior of the defraction gain and its dependence on distance and frequency can give qualitative insight. The Mathcad worksheet "Defraction" lets you see how the various parameters affect the defraction gain.

4.4 Scattering

A third mechanism affecting path loss, after reflection and diffraction, is scattering. Rough surfaces in the vicinity of the transmitter do not reflect the signal cleanly in the direction determined by the incident angle, but diffuse it, or scatter it in all directions. As a result, the

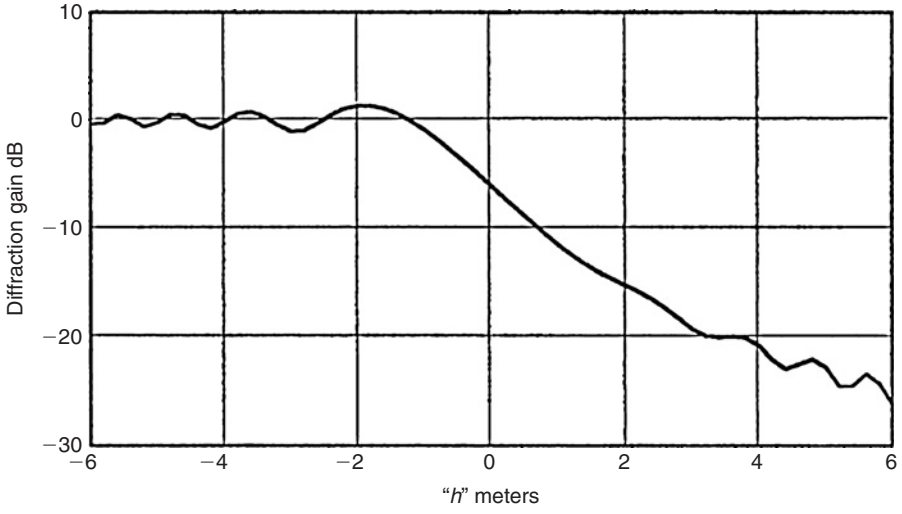


Figure 4.4: Example Plot of Diffraction Gain vs. "*h*"

receiver has access to additional radiation and path loss may be less than it would be from considering reflection and diffraction alone.

The degree of roughness of a surface and the amount of scattering it produces depends on the height of the protuberances on the surface compared to a function of the wavelength and the angle of incidence. The critical surface height h_c is given by [Gibson]:

$$h_c = \frac{\lambda}{8 \cos \theta_i} \tag{4.4}$$

where λ is the wavelength and θ_i is the angle of incidence. It is the dividing line between smooth and rough surfaces when applied to the difference between the maximum and the minimum protuberances.

4.5 Path Loss

The numerical path loss is the ratio of the total radiated power from a transmitter antenna times the numerical gain of the antenna in the direction of the receiver to the power available at the receiver antenna. This is the ratio of the transmitter power delivered to a lossless antenna with numerical gain of 1 (0 dB) to that at the output of a 0 dB gain receiver antenna. Sometimes, for clarity, the ratio is called the *isotropic* path loss. An isotropic radiator is an ideal antenna that radiates equally in all directions and therefore has a gain of 0 dB. The inverse path loss ratio is sometimes more convenient to use. It is called the path gain and when expressed in decibels is a negative quantity. In free space, the isotropic path gain *PG* is derived from equation [4.2], resulting in

$$PG = \frac{\lambda^2}{(4\pi d)^2} \tag{4.5}$$

We have just examined several factors that affect the path loss of VHF-UHF signals—ground reflection, diffraction, and scattering. For a given site, it would be very difficult to calculate the path loss between transmitters and receivers, but empirical observations have allowed some general conclusions to be drawn for different physical environments. These conclusions involve determining the exponent, or range of exponents, for the distance d related to a short reference distance d_0. We then can write the path gain as dependent on the exponent n:

$$PG = k\left(\frac{d_0}{d}\right)^n \tag{4.6}$$

where k is equal to the path gain when $d = d_0$. Table 4.1 shows path loss for different environments.

Table 4.1: Path Loss Exponents for Different Environments [Gibson]

Environment	Path gain exponent *n*
Free space	2
Open field (long distance)	4
Cellular radio—urban area	2.7–4
Shadowed urban cellular radio	5–6
In building line-of site	1.6–1.8
In building—obstructed	4–6

As an example of the use of the path gain exponent, let's assume the open field range of a security system transmitter and receiver is 300 meters. What range can we expect for their installation in a building?

Figure 4.5 shows curves of path gain versus distance for free-space propagation, open field propagation, and path gain with exponent 6, a worst case taken from Table 4.1 for "In building, obstructed." Transmitter and receiver heights are 2.5 meters, polarization is vertical, and the frequency is 915 MHz. The reference distance is 10 meters, and for all three curves the path gain at 10 meters is taken to be that of free space. For an open field distance of 300 meters, the path gain is −83 dB. The distance on the curve with exponent $n = 6$ that gives the same path gain is 34 meters. Thus, a wireless system that has an outdoor range of 300 meters may be effective only over a range of 34 meters, on the average, in an indoor installation.

The use of an empirically derived relative path loss exponent gives an estimate for average range, but fluctuations around this value should be expected. The next section shows the spread of values around the mean that occurs because of multipath radiation.

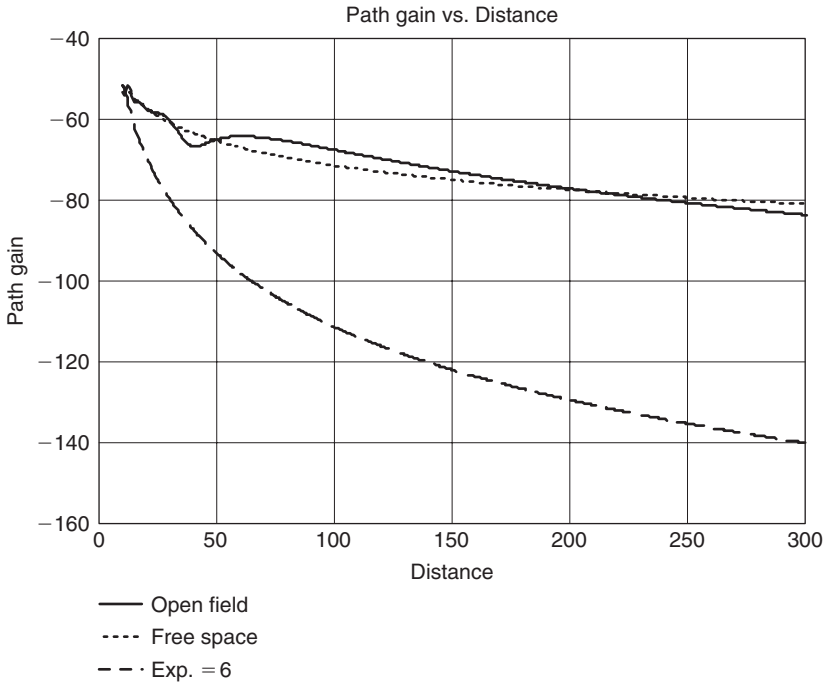


Figure 4.5: Path Gain

4.6 Multipath Phenomena

We have seen that reflection of a signal from the ground has a significant effect on the strength of the received signal. The nature of short-range radio links, which are very often installed indoors and use omnidirectional antennas, makes them accessible to a multitude of reflected rays, from floors, ceilings, walls, and the various furnishings and people that are invariably present near the transmitter and receiver. Thus, the total signal strength at the receiver is the vector sum of not just two signals, as we studied in section 4.2, but of many signals traveling over multiple paths. In most cases indoors, there is no direct line-of-sight path, and all signals are the result of reflection, diffraction and scattering.

From the point of view of the receiver, there are several consequences of the multipath phenomena.

a) *Variation of signal strength*. Phase cancellation and strengthening of the resultant received signal causes an uncertainty in signal strength as the range changes, and even at a fixed range when there are changes in furnishings or movement of people. The receiver must be able to handle the considerable variations in signal strength.

b) *Frequency distortion*. If the bandwidth of the signal is wide enough so that its various frequency components have different phase shifts on the various signal paths, then the

resultant signal amplitude and phase will be a function of sideband frequencies. This is called frequency selective fading.

c) *Time delay spread*. The differences in the path lengths of the various reflected signals causes a time delay spread between the shortest path and the longest path. The resulting distortion can be significant if the delay spread time is of the order of magnitude of the minimum pulse width contained in the transmitted digital signal. There is a close connection between frequency selective fading and time-delay distortion, since the shorter the pulses, the wider the signal bandwidth. Measurements in factories and other buildings have shown multipath delays ranging from 40 to 800 ns (Gibson).

d) *Fading*. When the transmitter or receiver is in motion, or when the physical environment is changing (tree leaves fluttering in the wind, people moving around), there will be slow or rapid fading, which can contain amplitude and frequency distortion, and time delay fluctuations. The receiver AGC and demodulation circuits must deal properly with these effects.

4.7 Flat Fading

In many of the short-range radio applications covered in this book, the signal bandwidth is narrow and frequency distortion is negligible. The multipath effect in this case is classified as *flat fading*. In describing the variation of the resultant signal amplitude in a multipath environment, we distinguish two cases: (1) there is no line-of-sight path and the signal is the resultant of a large number of randomly distributed reflections; (2) the random reflections are superimposed on a signal over a dominant constant path, usually the line of sight.

Short-range radio systems that are installed indoors or outdoors in built-up areas are subject to multipath fading essentially of the first case. Our aim in this section is to determine the signal strength margin that is needed to ensure that reliable communication can take place at a given probability. While in many situations there will be a dominant signal path in addition to the multipath fading, restricting ourselves to an analysis of the case where all paths are the result of random reflections gives us an upper bound on the required margin.

4.7.1 Rayleigh Fading

The first case can be described by a received signal $R(t)$, expressed as

$$R(t) = r \cdot \cos(2\pi \cdot f_c \cdot t + \theta) \tag{4.7}$$

where r and θ are random variables for the peak signal, or envelope, and phase. Their values may vary with time, when various reflecting objects are moving (people in a room, for example), or with changes in position of the transmitter or receiver which are small in respect

to the distance between them. We are not dealing here with the large-scale path gain that is expressed in equation [4.5] and [4.6]. For simplicity, equation [4.7] shows a CW (continuous wave) signal as the modulation terms are not needed to describe the fading statistics.

The envelope of the received signal, r, can be statistically described by the Rayleigh distribution whose probability density function is

$$p(r) = \frac{r}{\sigma^2} e^{\frac{-r^2}{2\sigma^2}} \tag{4.8}$$

where σ^2 represents the variance of $R(t)$ in equation [4.7], which is the average received signal power. This function is plotted in Figure 4.6. We normalized the curve with σ equal to 1. In this plot, the average value of the signal envelope, shown by a dotted vertical line, is 1.253. Note that it is not the most probable value, which is 1 (σ). The area of the curve between any two values of signal strength r represents the probability that the signal strength will be in that range. The average for the Rayleigh distribution, which is not symmetric, does not divide the curve area in half. The parameter that does this is the *median*, which in this case equals 1.1774. There is a 50% probability that a signal will be below the median and 50% that it will be above.

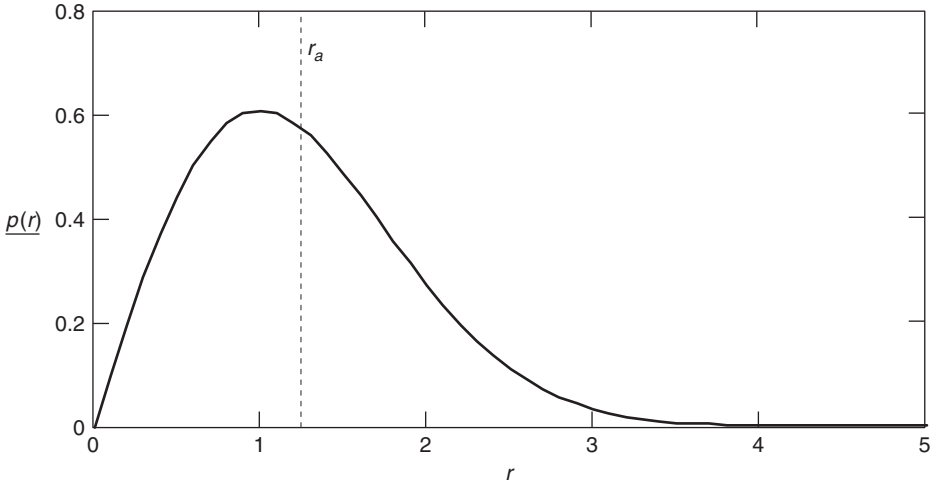


Figure 4.6: Rayleigh Probability Density Function

As stated above, the Rayleigh distribution is used to determine the signal margin required to give a desired communication reliability over a fading channel with no line of sight. The curve labeled "1 Channel" in Figure 4.7 is a cumulative distribution function with logarithmic axes. For any point on the curve, the probability of fading below the margin indicated on the

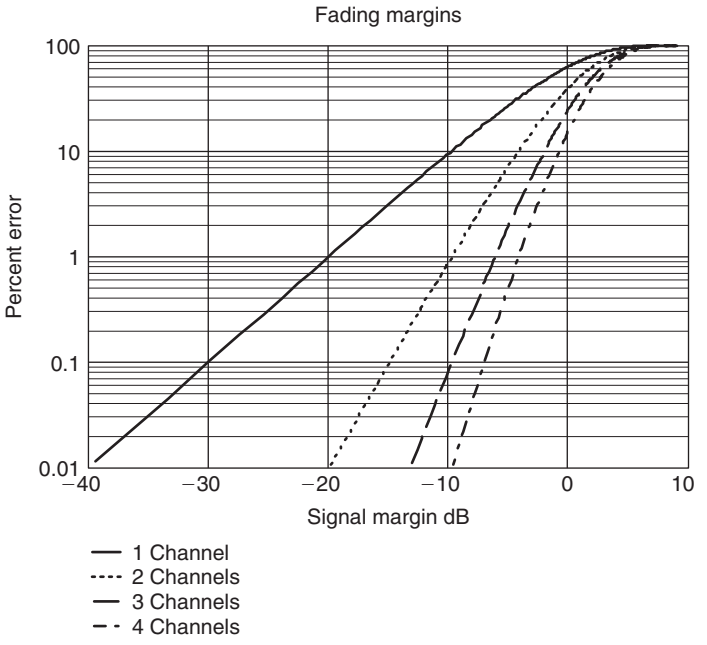


Figure 4.7: Fading Margins

abscissa is given as the ordinate. The curve is scaled such that "0 dB" signal margin represents the point where the received signal equals the mean power of the fading signal, σ^2, making the assumption that the received signal power with no fading equals the average power with fading. Some similar curves in the literature use the median power, or the power corresponding to the average envelope signal level, r_a, as the reference, "0 dB" value.

An example of using the curve is as follows. Say you require a communication reliability of 99%. Then the minimum usable signal level is that for which there is a 1% probability of fading below that level. On the curve, the margin corresponding to 1% is 20 dB. Thus, you need a signal strength 20 dB larger than the required signal if there was no fading. Assume you calculated path loss and found that you need to transmit 4 mW to allow reception at the receiver's sensitivity level. Then, to ensure that the signal will be received 99% of the time during fading, you'll need 20 dB more power or 6 dBm (4 mW) plus 20 dB equals 26 dBm or 400 mW. If you don't increase the power, you can expect loss of communication 63% of the time, corresponding to the "0 dB" margin point on the "Channel 1" curve of Figure 4.7.

Table 4.2 shows signal margins for different reliabilities.

4.8 Diversity Techniques

Communication reliability for a given signal power can be increased substantially in a multipath environment through diversity reception. If signals are received over multiple,

Table 4.2: Signal Margins for Different Reliabilities

Reliability, Percent	Fading Margin, dB
90	10
99	20
99.9	30
99.99	40

independent channels, the largest signal can be selected for subsequent processing and use. The key to this solution is the independence of the channels. The multipath effect of nulling and of strengthening a signal is dependent on transmitter and receiver spatial positions, on wavelength (or frequency) and on polarity. Let's see how we can use these parameters to create independent diverse channels.

4.8.1 Space Diversity

A signal that is transmitted over slightly different distances to a receiver may be received at very different signal strengths. For example, in Figure 4.2 the signal at 17 meters is at a null and at 11 meters at a peak. If we had two receivers, each located at one of those distances, we could choose the strongest signal and use it. In a true multipath environment, the source, receiver, or the reflectors may be constantly in motion, so the nulls and the peaks would occur at different times on each channel. Sometimes Receiver 1 has the strongest signal, at other times Receiver 2. Figure 4.8 illustrates the paths to two receivers from several reflectors. Although there may be circumstances where the signals at both receiver locations are at around the same level, when it doesn't matter which receiver output is chosen, most of the time one signal will be stronger than the other. By selecting the strongest output, the

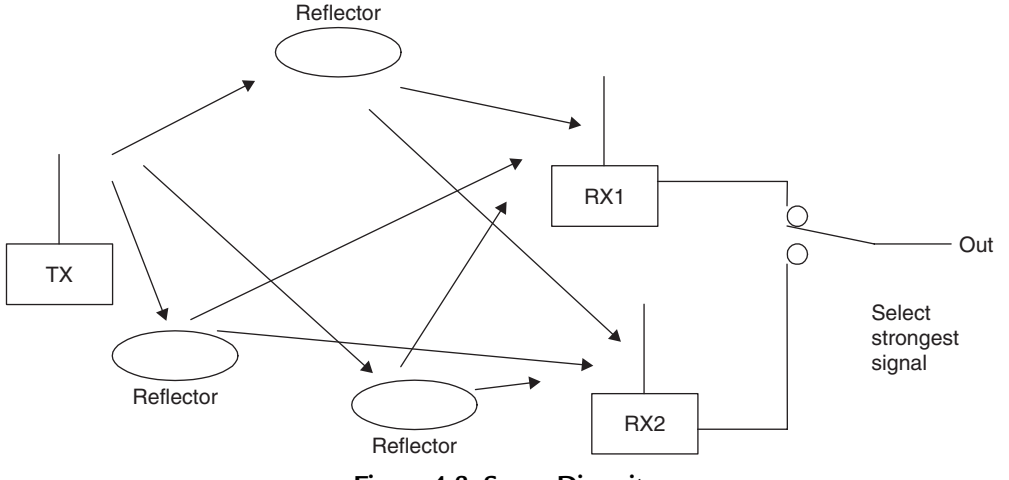


Figure 4.8: Space Diversity

average output after selection will be greater than the average output of one channel alone. To increase even more the probability of getting a higher average output, we could use three or more receivers. From Figure 4.7 you can find the required fading margin using diversity reception having 2, 3, or 4 channels. Note that the plots in Figure 4.7 are based on completely independent channels. When the channels are not completely independent, the results will not be as good as indicated by the plots.

It isn't necessary to use complete receivers at each location, but separate antennas and front ends must be used, at least up to the point where the signal level can be discerned and used to decide on the switch position.

4.8.2 Frequency Diversity

You can get a similar differential in signal strength over two or more signal channels by transmitting on separate frequencies. For the same location of transmitting and receiving antennas, the occurrences of peaks and nulls will differ on the different frequency channels. As in the case of space diversity, choosing the strongest channel will give a higher average signal-to-noise ratio than on either one of the channels. The required frequency difference to get near independent fading on the different channels depends on the diversity of path lengths or signal delays. The larger the difference in path lengths, the smaller the required frequency difference of the channels.

4.8.3 Polarization Diversity

Fading characteristics are dependent on polarization. A signal can be transmitted and received separately on horizontal and vertical antennas to create two diversity channels. Reflections can cause changes in the direction of polarization of a radio wave, so this characteristic of a signal can be used to create two separate signal channels. Thus, cross-polarized antennas can be used at the receiver only. Polarization diversity can be particularly advantageous in a portable handheld transmitter, since the orientation of its antenna will not be rigidly defined.

Polarization diversity doesn't allow the use of more than two channels, and the degree of independence of each channel will usually be less than in the two other cases. However, it may be simpler and less expensive to implement and may give enough improvement to justify its use, although performance will be less than can be achieved with space or frequency diversity.

4.8.4 Diversity Implementation

In the descriptions above, we talked about selecting or switching to the channel having the highest signal level. A more effective method of using diversity is called *maximum ratio combining*. In this technique, the outputs of each independent channel are added together after the channel phases are made equal and channel gains are adjusted for equal signal levels. Maximum ratio combining is known to be optimum as it gives the best statistical reduction of

fading of any linear diversity combiner. In applications where accurate amplitude estimation is difficult, the channel phases only may be equalized and the outputs added without weighting the gains. Performance in this case is almost as good as in maximum ratio combining. [Gibson]

Space diversity has the disadvantage of requiring significant antenna separation, at least in the VHF and lower UHF bands. In the case where multipath signals arrive from all directions, antenna spacing on the order of $.5\lambda$ to $.8\lambda$ is adequate in order to have reasonably independent, or decorrelated, channels. This is at least one-half meter at 300 MHz. When the multipath angle spread is small—for example, when directional antennas are used—much larger separations are required.

Frequency diversity eliminates the need for separate antennas, but the simultaneous use of multiple frequency channels entails increased total power and spectrum utilization. Sometimes data are repeated on different frequencies so that simultaneous transmission doesn't have to be used. Frequency separation must be adequate to create decorrelated channels. The bandwidths allocated for unlicensed short-range use are rarely adequate, particularly in the VHF and UHF ranges (transmitting simultaneously on two separate bands can and has been done). Frequency diversity to reduce the effects of time delay spread is achieved with frequency hopping or direct sequence spread spectrum modulation, but for the spreads encountered in indoor applications, the pulse rate must be relatively high—of the order of several megabits per second—in order to be effective. For long pulse widths, the delay spread will not be a problem anyway, but multipath fading will still occur and the amount of frequency spread normally used in these cases is not likely to solve it.

When polarity diversity is used, the orthogonally oriented antennas can be close together, giving an advantage over space diversity when housing dimensions relative to wavelength are small. Performance may not be quite as good, but may very well be adequate, particularly when used in a system having portable hand-held transmitters, which have essentially random polarization.

Although we have stressed that at least two independent (decorrelated) channels are needed for diversity reception, sometimes shortcuts are taken. In some low-cost security systems, for example, two receiver antennas—space diverse or polarization diverse—are commutated directly, usually by diode switches, before the front end or mixer circuits. Thus, a minimum of circuit duplication is required. In such applications the message frame is repeated many times, so if there happens to be a multipath null when the front end is switched to one antenna and the message frames are lost, at least one or more complete frames will be correctly received when the switch is on the other antenna, which is less affected by the null. This technique works for slow fading, where the fade doesn't change much over the duration of a transmission of message frames. It doesn't appear to give any advantage during fast fading, when used with moving hand-held transmitters, for example. In that case, a receiver with one antenna will have a better chance of decoding at least one of many frames than when

switched antennas are used and only half the total number of frame repetitions are available for each. In a worst-case situation with fast fading, each antenna in turn could experience a signal null.

4.8.5 Statistical Performance Measure

We can estimate the performance advantage due to diversity reception with the help of Figure 4.7. Curves labeled "2 Channels" through "4 Channels" are based on the selection combining technique.

Let's assume, as before, that we require communication reliability of 99 percent, or an error rate of 1 percent. From probability theory, the probability that two independent channels would both have communication errors is the product of the error probabilities of each channel. Thus, if each of two channels has an error probability of 10 percent, the probability that both channels will have signals below the sensitivity threshold level when selection is made is .1 times .1, which equals .01, or 1 percent. This result is reflected in the curve "2 Channels". We see that the signal margin needed for 99 percent reliability (1 percent error) is 10 dB. Using diversity reception with selection from two channels allows a reliability margin of only 10 dB instead of 20 dB, which is required if there is no diversity. Continuing the previous example, we need to transmit only 40 mW for 99 percent reliability instead of 400 mW. Required margins by selection among three channels and four channels is even less—6 dB and 4 dB, respectively.

Remember that the reliability margins using selection combining diversity as shown in Figure 4.7 are ideal cases, based on the Rayleigh fading probability distribution and independently fading channels. However, even if these premises are not realized in practice, the curves still give us approximations of the improvement that diversity reception can bring.

4.9 Noise

The ultimate limitation in radio communication is not the path loss or the fading. Weak signals can be amplified to practically any extent, but it is the noise that bounds the range we can get or the communication reliability that we can expect from our radio system. There are two sources of receiver noise—interfering radiation that the antenna captures along with the desired signal, and the electrical noise that originates in the receiver circuits. In either case, the best signal-to-noise ratio will be obtained by limiting the bandwidth to what is necessary to pass the information contained in the signal. A further improvement can be had by reducing the receiver noise figure, which decreases the internal receiver noise, but this measure is effective only as far as the noise received through the antenna is no more than about the same level as the receiver noise. Finally, if the noise can be reduced no further, performance of digital receivers can be improved by using error correction coding up to a point, which is designated as channel capacity. The capacity is the maximum information rate that the specific

channel can support, and above this rate communication is impossible. The channel capacity is limited by the noise density (noise power per hertz) and the bandwidth.

Figure 4.9 shows various sources of noise over a frequency range of 20 kHz up to 10 GHz. The strength of the noise is shown in microvolts/ meter for a bandwidth of 10 kHz, as received by a half-wave dipole antenna at each frequency. The curve labeled "equivalent receiver noise" translates the noise generated in the receiver circuits into an equivalent field strength so that it can be compared to the external noise sources. Receiver noise figures on which the curve is based vary in a log-linear manner with 2 dB at 50 MHz and 9 dB at 1 GHz. The data in Figure 4.9 are only a representative example of radiation and receiver noise, taken at a particular time and place.

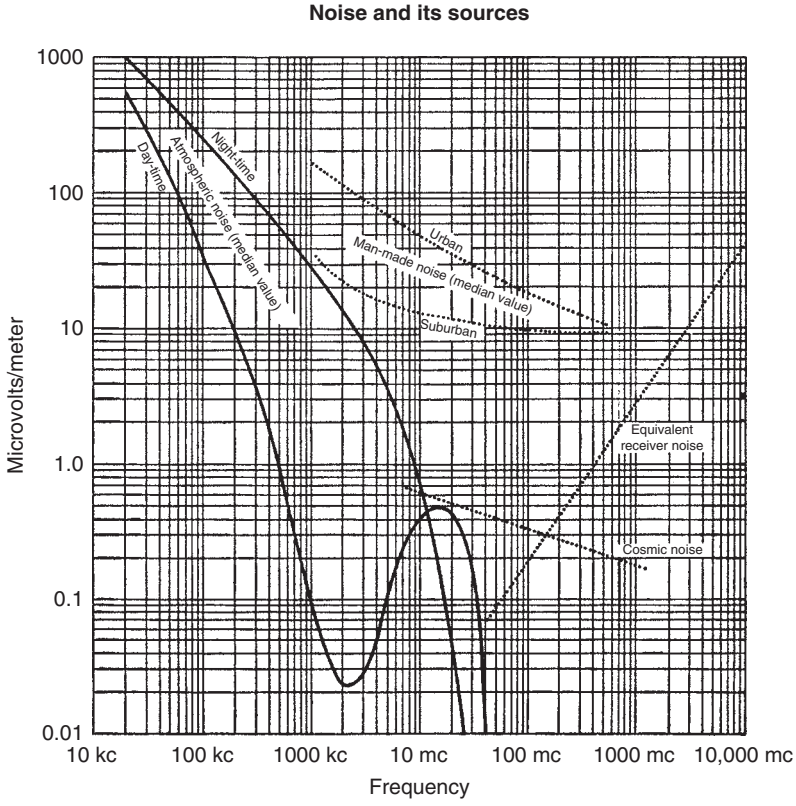


Figure 4.9: External Noise Sources
(Reference Data for Radio Engineers, Fourth Edition)

Note that all of the noise sources shown in Figure 4.9 are dependent on frequency. The relative importance of the various noise sources to receiver sensitivity depends on their strength relative to the receiver noise. Atmospheric noise is dominant on the low radio frequencies but is not significant on the bands used for short-range communication—above around 40 MHz. Cosmic noise comes principally from the sun and from the center of our galaxy. In the figure, it is masked

out by man-made noise, but in locations where man-made noise is a less significant factor, cosmic noise affects sensitivity up to 1 GHz.

Man-made noise is dominant in the range of frequencies widely used for short-range radio systems—VHF and low to middle UHF bands. It is caused by a wide range of ubiquitous electrical and electronic equipment, including automobile ignition systems, electrical machinery, computing devices and monitors. While we tend to place much importance on the receiver sensitivity data presented in equipment specifications, high ambient noise levels can make the sensitivity irrelevant in comparing different devices. For example, a receiver may have a laboratory measured sensitivity of −105 dBm for a signal-to-noise ratio of 10 dB. However, when measured with its antenna in a known electric field and accounting for the antenna gain, −95 dBm may be required to give the same signal-to-noise ratio.

From around 800 MHz and above, receiver sensitivity is essentially determined by the noise figure. Improved low-noise amplifier discrete components and integrated circuit blocks produce much better noise figures than those shown in Figure 4.9 for high UHF and microwave frequencies. Improving the noise figure must not be at the expense of other characteristics—intermodulation distortion, for example, which can be degraded by using a very high-gain amplifier in front of a mixer to improve the noise figure. Intermodulation distortion causes the production of inband interfering signals from strong signals on frequencies outside of the receiving bandwidth.

External noise will be reduced when a directional antenna is used. Regulations on unlicensed transmitters limit the peak radiated power. When possible, it is better to use a high-gain antenna and lower transmitter power to achieve the same peak radiated power as with a lower gain antenna. The result is higher sensitivity through reduction of external noise. Manmade noise is usually less with a horizontal antenna than with a vertical antenna.

4.10 Summary

In this chapter we have looked at various factors that affect the range of reliable radio communication. Propagation of electromagnetic waves is influenced by physical objects in and near the path of line-of-sight between transmitter and receiver. We can get a first rough approximation of communication distance by considering only the reflection of the transmitted signal from the earth. If the communication system site can be classified, an empirically determined exponent can be used to estimate the path loss, and thus the range. When the transmitter or receiver is in motion, or surrounding objects are not static, the path loss varies and must be estimated statistically. We described several techniques of diversity reception that can reduce the required power for a given reliability when the communication link is subject to fading.

Noise was presented as the ultimate limitation on communication range. We saw that the noise sources to be contended with depend on the operating frequency. The importance of low-noise

receiver design depends on the relative intensity of noise received by the antenna to the noise generated in the receiver.

By having some degree of understanding of electromagnetic wave propagation and noise, all those involved in the deployment of a wireless communication system—designers, installers and users—will know what performance they can expect from the system and what concrete measures can be taken to improve it.

References

[4.1] Gibson, Jerry, D., Editor-in-Chief, The Mobile Communications Handbook, CRC Press, Inc., 1996.

[4.2] Rappaport, Theodore S., *Wireless Communications, Principles and Practice*, Prentice Hall, Upper Saddle River, NJ, 1996.

[4.3] Spix, George J., "Maxwell's Electromagnetic Field Equations," unpublished tutorial, copyright 1995 (http://www.connectos.com/spix/rd/gj/nme/maxwell.htm).

This page intentionally left blank

CHAPTER 5

Antennas and Transmission Lines

Alan Bensky

5.1 Introduction

The antenna is the interface between the transmitter or the receiver and the propagation medium, and it therefore is a deciding factor in the performance of a radio communication system. The principal properties of antennas—directivity, gain, and radiation resistance—are the same whether referred to as transmitters or receivers. The principle of reciprocity states that the power transferred between two antennas is the same, regardless of which is used for transmission or reception, if the generator and load impedances are conjugates of the transmitting and receiving antenna impedances in each case.

First we define the various terms used to characterize antennas. Then we discuss several types of antennas that are commonly used in short-range radio systems. Finally, we review methods of matching the impedances of the antenna to the transmitter or receiver RF circuits.

5.2 Antenna Characteristics

Understanding the various characteristics of antennas is a first and most important step before deciding what type of antenna is most appropriate for a particular application. While antennas have several electrical characteristics, often a primary concern in choosing an antenna type is its physical size. Before dealing with the various antenna types and the shapes and sizes they come in, we first must know what the antenna has to do.

5.2.1 Antenna Impedance

As stated in the introduction, the antenna is an interface between circuits and space. It facilitates the transfer of power between space and the transmitter or receiver. The antenna impedance is the load for the transmitter or the input impedance to the receiver. It is composed of two parts—radiation resistance and ohmic resistance. The radiation resistance is a virtual resistance that, when multiplied by the square of the RMS current in the antenna at its feed point, equals the power radiated by the antenna in the case of a transmitter or extracted from space in the case of a receiver. It is customary to refer the radiation resistance to a current maximum in the case of an ungrounded antenna, and to the current at the base of the antenna when the antenna is grounded. Transmitter power delivered to an antenna will always be

greater than the power radiated. The difference between the transmitter power and the radiated power is power dissipated in the ohmic resistance of the antenna conductor and in other losses. The efficiency of an antenna is the ratio of the radiated power to the total power absorbed by the antenna. It can be expressed in terms of the radiation resistance R_r and loss resistance R_l as

$$\text{Antenna Efficiency}\,(\%) = 100 \times (R_r/(R_r + R_l)) \tag{5.1}$$

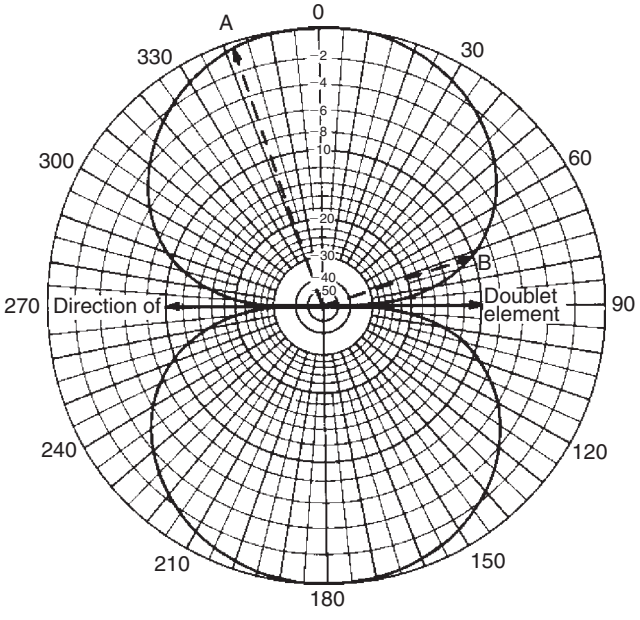
The resistance seen by the transmitter or receiver at the antenna terminals will be equal to the radiation resistance plus the loss resistance only if these terminals are located at the point of maximum current flow in the antenna. The impedance at this point may have a reactive component, too. When there is no reactive component, the antenna is said to be resonant. Maximum power transfer between the antenna and transmitter or receiver will occur only when the impedance seen from the antenna terminals is the complex conjugate of the antenna impedance.

It is important to match the transmitter to the antenna not only to get maximum power transfer. Attenuation of harmonics relative to the fundamental frequency is maximized when the transmitter is matched to the antenna—an important point in meeting the spurious radiation requirements for license-free transmitters. The radiation resistance depends on the proximity of the antenna to conducting and insulating objects. In particular, it depends on the height of the antenna from the ground. Thus, the antenna-matching circuit of a transmitter with integral antenna that is intended to be hand-held should be optimized for the antenna impedance in a typical operating situation.

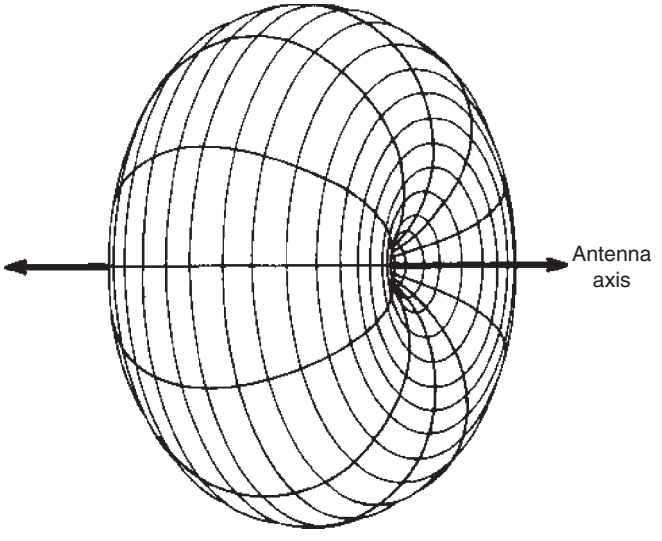
5.2.2 Directivity and Gain

The directivity of an antenna relates to its radiation pattern. An antenna that radiates uniformly in all directions in three-dimensional space is called an isotropic antenna. Such an antenna doesn't exist, but it is convenient to refer to it when discussing the directional properties of an antenna. All real antennas radiate stronger in some directions than in others. The directivity of an antenna is defined as the power density of the antenna in its direction of maximum radiation in three-dimensional space divided by its average power density. The directivity of the hypothetical isotropic radiator is 1, or 0 dB. The directivity of a half-wave dipole antenna is 1.64, or 2.15 dB.

The radiation pattern of a wire antenna of short length compared to a half wavelength is shown in Figure 5.1a. The antenna is high enough so as not to be affected by the ground. If the antenna wire direction is parallel to the earth, then the pattern represents the intersection of a horizontal plane with the solid pattern of the antenna shown in Figure 5.1b. A vertical wire antenna is omnidirectional; that is, it has a circular horizontal radiation pattern and directivity in the vertical plane.



(a) Directive Pattern in Plane Containing Antenna



(b) Solid Diagram

Figure 5.1: Short Dipole Antenna (ARRL Antenna Book)

Courtesy *Antenna Book, 16th Edition*, Published by ARRL

The gain of an antenna is the directivity times the antenna efficiency. When antenna losses are low, the two terms are almost the same. In general, when you are interested in the directional discrimination of an antenna, you will be interested in its directivity. Gain is used to find the maximum radiated power when the power into the antenna is known.

5.2.3 Effective Area

Another term often encountered is the effective area of an antenna. Wave propagation can be described as if all of the radiated power is spread over the surface of a sphere whose area expands according to the square of the distance (in free space). The power captured by the receiving antenna is then the capture area, or effective area, of the antenna times the power density at that location. The power density is the radiated power divided by the surface area of the sphere.

The effective area of an antenna related to gain and wavelength is shown in the following expression:

$$A_e = \frac{\lambda^2 \cdot G}{4\pi} \tag{5.2}$$

This expression shows us that the capture ability of an antenna of given gain G grows proportionally as the square of the wavelength. The antenna of a particular configuration captures less power at higher frequencies. (Remember that frequency is inversely proportional to wavelength λ.)

When the electric field strength E is known, the power density is

$$P_d = \frac{E^2}{120\pi} \tag{5.3}$$

Thus, received power can be found when field strength is known by multiplying (5.2) times (5.3):

$$P_R = \frac{E^2\lambda^2 G}{480\pi^2} \tag{5.4}$$

It's intuitive to note that the effective antenna area has some connection with the physical size of the antenna. This is most obvious in the case of microwave antennas where the effective area approaches the physical aperture. From (5.4) it appears that for a given radiated power and thus field strength, the lower frequency (longer wavelength) systems will give stronger receiver signals than high-frequency equipment. However, short-range devices are often portable or are otherwise limited in size and their antennas may have roughly the same dimensions, regardless of frequency. The lower frequency antennas whose sizes are small fractions of a wavelength have poor efficiency and low gain and therefore may have effective areas similar to their high-frequency counterparts. Thus, using a low frequency doesn't necessarily mean higher power at the receiver, which equation [5.4] may lead us to believe.

5.2.4 Polarization

Electromagnetic radiation is composed of a magnetic field and an electric field. These fields are at right angles to each other, and both are in a plane normal to the direction of propagation.

The direction of polarization refers to the direction of the electric field in relation to the earth. Linear polarization is created by a straight wire antenna. A wire antenna parallel to the earth is horizontally polarized and a wire antenna normal to the earth is vertically polarized.

The electric and magnetic fields may rotate in their plane around the direction of propagation, and this is called elliptical polarization. It may be created by perpendicular antenna elements being fed by coherent RF signals that are not in the same time phase with each other. Circular polarization results when these elements are fed by equal power RF signals which differ in phase by 90°, which causes the electric (and magnetic) field to make a complete 360° rotation every period of the wave (a time of 1/freq. seconds). Some antenna types, among them the helical antenna, produce elliptic or circular polarization inherently, without having two feed points. There are two types of elliptical polarization, right hand and left hand, which are distinguished by the direction of rotation of the electric field.

The polarization of a wave, or an antenna, is important for several reasons. A horizontally polarized receiving antenna cannot receive vertically polarized radiation from a vertical transmitting antenna, and vice versa. Similarly, right-hand and left-hand circular antenna systems are not compatible. Sometimes this quality is used to good advantage. For example, the capacity of a microwave link can be doubled by transmitting two different information channels between two points on the same frequency using oppositely polarized antenna systems.

The degree of reflection of radio signals from the ground is affected by polarization. The phase and amount of reflection of vertically polarized waves from the ground are much more dependent on the angle of incidence than horizontally polarized waves.

Except for directional, line-of-sight microwave systems, the polarity of a signal may change during propagation between transmitter and receiver. Thus, in most short-range radio applications, a horizontal antenna will receive transmissions from a vertical antenna, for example, albeit with some attenuation. The term *cross polarization* defines the degree to which a transmission from an antenna of one polarization can be received by an antenna of the opposite polarization. Often, the polarization of a transmitter or receiver antenna is not well defined, such as in the case of a handheld device. A circular polarized antenna can be used when the opposite antenna polarization is not defined, since it does not distinguish between the orientation of the linear antenna.

5.2.5 Bandwidth

Antenna bandwidth is the range of frequencies over which the antenna can operate while some other characteristic remains within a given range. Very frequently, the bandwidth is related to the antenna impedance expressed as standing wave ratio. Obviously, a device that must operate over a number of frequency channels in a band must have a comparatively wide bandwidth antenna. Less obvious are the bandwidth demands for a single frequency device.

A narrow bandwidth or high Q antenna will discriminate against harmonics and other spurious radiation and thereby will reduce the requirements for a supplementary filter, which may be necessary to allow meeting the radio approval specifications. On the other hand, drifting of antenna physical dimensions or matching components could cause the power output (or sensitivity) to fall with time. Changing proximity of nearby objects or the "hand effect" of portable transmitters can also cause a reduction of power or even a pulling of frequency, particularly in low-power transmitters with a single oscillator stage and no buffer or amplifier stage.

5.2.6 Antenna Factor

The antenna factor is commonly used with calibrated test antennas to make field strength measurements on a test receiver or spectrum analyzer. It relates the field strength to the voltage across the antenna terminals when the antenna is terminated in its specified impedance (usually 50 or 75 ohms):

$$AF = \frac{E}{V} \tag{5.5}$$

where

AF = antenna factor in $(\text{meters})^{-1}$

E = field strength in V/m

V = load voltage in V

Usually the antenna factor is stated in dB:

$AF_{\text{dB}}(\text{m}^{-1}) = 20 \log(E/V)$

The relationship between numerical gain G and antenna factor AF is:

$$AF = \frac{4\pi}{\lambda} \cdot \sqrt{\frac{30}{R_{\text{L}} \cdot G}} \tag{5.6}$$

where R_{L} is the load resistance, usually 50 ohms.

5.3 Types of Antennas

In this section we review the characteristics of several types of antennas that are used in short-range radio devices. The size of the antenna is related to the wavelength, which in turn is found when frequency is known from

wavelength = (velocity of propagation)/(frequency)

The maximum velocity of propagation occurs in a vacuum. It is approximately 300,000,000 meters/second, with little difference in air. This figure is less in solid materials, so the

wavelength will be shorter for antennas printed on circuit board materials or protected with a plastic coating.

5.3.1 Dipole

The dipole is a wire antenna fed at its center. The term usually refers to an antenna whose overall length is one-half wavelength. In free space its radiation resistance is 73 ohms, but that value will vary somewhat when the ground or other large conducting objects are within a wavelength distance from it. The dipole is usually mounted horizontally, but if mounted vertically, its transmission line feeder cable should extend from it at a right angle for a distance of at least a quarter wavelength.

In free space, the radiation pattern of a horizontal half-wave dipole at zero degrees elevation is very similar to that of the dipole shown in Figure 5.1a. It has a directivity of 1.64, or 2.15 dB. It would seem that this antenna would not be usable for short-range devices that need an omnidirectional radiation, but this isn't necessarily the case. The radiation pattern at an elevation of 30 degrees when the antenna is a half wavelength above a conductive plate is shown in Figure 5.2. The radiation is 8 dB down from maximum along the direction of the wire. When used indoors where there are a multitude of reflections from walls, floor and ceiling, the horizontal dipole can give good results in all directions.

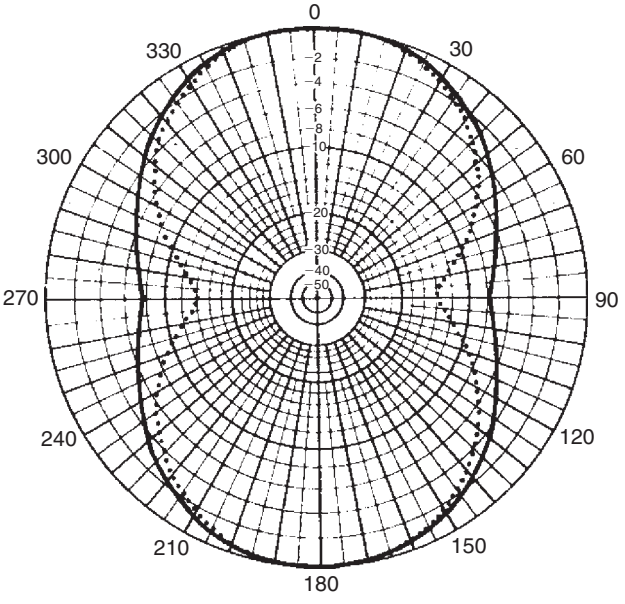


Figure 5.2: Dipole Pattern at 30° Elevation (Solid Line)
Courtesy 1990 ARRL Handbook

The half-wave dipole antenna is convenient to use because it is easy to match a transmitter or receiver to its radiation resistance. It has high efficiency, since wire ohmic losses are only

a small fraction of the radiation resistance. Also, the antenna characteristics are not much affected by the size or shape of the device it is used with, and it doesn't use a ground plane. Devices whose dimensions are small relative to the antenna size can directly feed the dipole, with little or no transmission line. For increased compactness, the two antenna elements can be extended at an angle instead of being in a straight line.

In spite of its many attractive features, the half-wave dipole is not commonly used with short-range radio equipment. On the common unlicensed frequency bands, it is too large for many applications, particularly portable devices. The antenna types below are smaller than the half-wave dipole, and generally give reduced performance.

5.3.2 Groundplane

We mentioned that the dipole can be mounted vertically. If we take one dipole element and mount it perpendicular to a large metal plate, then we don't need the bottom element—a virtual element will be electrically reflected from the plate. When the metal plate is approximately one-half wavelength square or larger, the radiation resistance of the antenna is 36 ohms, and a good match to the receiver or transmitter can be obtained.

The quarter-wave groundplane antenna is ideal if the receiver or transmitter is encased in a metal enclosure that has the required horizontal area for an efficient vertical antenna. However, in many short-range devices, a quarter-wave vertical element is used without a suitable groundplane. In this case, the radiation resistance is much lower than 36 ohms and there is considerable capacitance reactance. An inductor is needed to cancel the reactance as well as a matching circuit to assure maximum power transfer between the antenna and the device. The ohmic losses in the inductor and other matching components, together with the low radiation resistance, result in low antenna efficiency. When possible, the antenna length should be increased to a point where the antenna is resonant, that is, has no reactance. The electrical length can be increased and capacitive reactance reduced by winding the bottom part of the antenna element into a coil having several turns. In this way, the loss resistance may be reduced and efficiency increased.

5.3.3 Loop

The loop antenna is popular for hand-held transmitters particularly because it can be printed on a small circuit board and is less affected by nearby conducting objects than other small resonant antennas. Its biggest drawback is that it is very inefficient.

A loop antenna whose dimensions are small compared to a wavelength—less than 0.1λ—has essentially constant current throughout. Its radiation field is expressed as

$$E(\theta) = \frac{120\pi^2 \cdot I \cdot N \cdot A}{r \cdot \lambda^2} \cdot \cos\theta \tag{5.7}$$

where

I is its current

A is the loop area

r is the distance

θ is the angle from the plane of the loop

N is the number of turns

From this expression we can derive the radiation resistance which is

$$R_r = 320 \cdot \pi^4 \cdot \frac{(A \cdot N)^2}{\lambda^4} \tag{5.8}$$

Loop antennas are frequently used in small hand-held remote control transmitters on the low UHF frequencies. The radiation resistance is generally below a tenth of an ohm and the efficiency under 10%. In order to match the transmitter output stage to the low antenna resistance, parallel resonance is created using a capacitor across the loop terminals. While it may appear that the radiation resistance and hence the efficiency could be raised by increasing the number of turns or the area of the loop, the possibilities with this approach are very limited. Increasing the area increases the loop inductance, which requires a smaller value of resonating capacitance. The limit on the area is reached when this capacitance is several picofarads, and then we get the radiation resistance and approximate efficiency as mentioned above.

Because of the low efficiency of the loop antenna, it is rarely used in UHF short-range receivers. An exception is pager receivers, which use low data rates and high sensitivity to help compensate for the low antenna efficiency. One advantage of the loop antenna is that it doesn't require a ground plane.

In low-power unlicensed transmitters, the low efficiency of the loop is not of much concern, since it is the radiated power that is regulated, and at the low powers in question, the power can be boosted enough to make up for the low efficiency. A reasonably high Q is required in the loop circuit, however, in order to keep harmonic radiation low in respect to the fundamental. In many short-range transmitters, it is the harmonic radiation specification that limits the fundamental power output to well below the allowed level.

For detailed information on loop antenna performance and matching, see References 5.1 and 5.7.

Example

We will design a loop antenna for a transmitter operating on 315 MHz. The task is easy using the Mathcad worksheet "Loop Antenna."

Given data: $f = 315$ MHz; G10 circuit board 1/16″ thick, 1 oz. copper plating, and dielectric constant = 4.7; loop sides 25 mm and 40 mm, conductor width 2 mm.

Enter the relevant data in the worksheet.

The results of this example are:

Radiation resistance = .038 ohm

Loss resistance = 0.15 ohm

Efficiency = 20.1% = −7 dB

Resonating capacitance = 3.65 pF

The results from using the loop antenna worksheet are not particularly accurate, but they do give a starting point for design. Efficiency can be expected to be worse than that calculated because circuit board losses were not accounted for, nor were the effects of surrounding components. There will also be significant losses in the matching circuit because of the difficulty of matching the high output impedance of the low-power transmitter to the very low impedance of the loop. Transmitters designed to operate from low battery voltages can be expected to be better in this respect (see "Transmitter Output Impedance" later in this section).

5.3.4 Helical

The helical antenna can give much better results than the loop antenna, when radiation efficiency is important, while still maintaining a relatively small size compared to a dipole or quarter-wave ground plane.

The helical antenna is made by winding stiff wire in the form of a spring, whose diameter and pitch are very much smaller than a wavelength, or by winding wire on a cylindrical form. See Figure 5.3. This helical winding creates an apparent axial velocity along the "spring" which is much less than the velocity of propagation along a straight wire approximately the

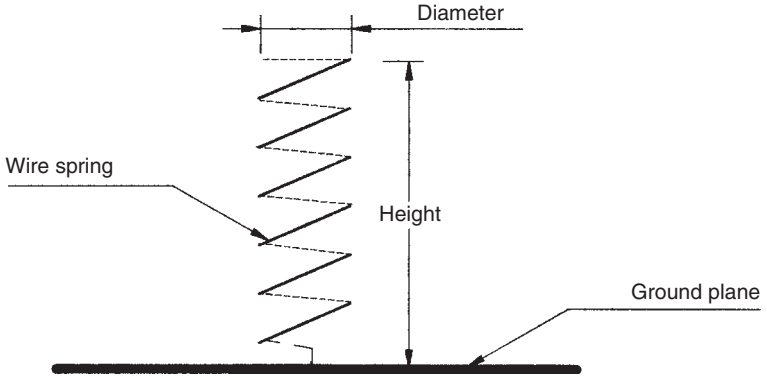


Figure 5.3: Helical Antenna

speed of light in space. Thus, a quarter wave on the helical spring will be much shorter than on a straight wire. The antenna is resonant for this length, but the radiation resistance will be lower and consequently the efficiency is less than that obtained from a standard quarter-wave antenna. The helical antenna resonates when the wire length is in the neighborhood of a half wavelength. Impedance matching to a transmitter or receiver is relatively easy.

The radiating surface of the helical antenna has both vertical and horizontal components, so its polarization is elliptic. However, for the form factors most commonly used, where the antenna length is several times larger than its diameter, polarization is essentially vertical.

The helical antenna should have a good ground plane for best and predictable performance. In hand-held devices, the user's arm and body serve as a counterpoise, and the antenna should be designed for this configuration.

The Mathcad worksheet "Helical Antenna" helps design a helical antenna. We'll demonstrate by an example.

Example

Our antenna will be designed for 173 MHz. We will wind it on a 10 mm form with AWG 20 wire. We want to find the number of turns to get a resonant antenna 16 cm high. We also want an approximation of the radiation resistance and the antenna efficiency.

Given: The mean diameter of the antenna $D = 10.8$ mm (includes the wire diameter). Wire diameter of AWG 20 is $d = .8$ mm. Antenna height $h = 160$ mm. Frequency $f = 173$ MHz.

We insert these values into the helical antenna worksheet and get the following results:

Number of turns = 26

Wire length = 89 cm = .514 λ

Radiation efficiency = 90 percent

Total input resistance = 6.1 ohm

The prototype antenna should have a few more turns than the design value so that the length can be gradually reduced while return loss is monitored, until a resonant condition or good match is obtained for the ground plane that results from the physical characteristics of the product. The input resistance of the antenna can be raised by grounding the bottom end of the antenna wire and tapping the wire up at a point where the desired impedance is found.

5.3.5 Patch

The patch antenna is convenient for microwave frequencies, specifically on the 2.4-GHz band and higher. It consists of a plated geometric form (the patch) on one side of a printed circuit board, backed up on the opposite board side by a groundplane plating which extends beyond

the dimensions of the radiating patch. Rectangular and circular forms are the most common, but other shapes—for example, a trapezoid—are sometimes used. Maximum radiation is generally perpendicular to the board. A square half-wave patch antenna has a directivity of 7 to 8 dB.

A rectangular patch antenna is shown in Figure 5.4. The dimension L is approximately a half wavelength, calculated as half the free space wavelength (λ) divided by the square root of the effective dielectric constant (ε) of the board material. It must actually be slightly less than a half wavelength because of the fringing effect of the radiation from the two opposite edges that are L apart. As long as the feed is on the centerline, the two other edges don't radiate. The figure shows a microstrip feeder, which is convenient because it is etched on the board together with the patch and other component traces on the same side.

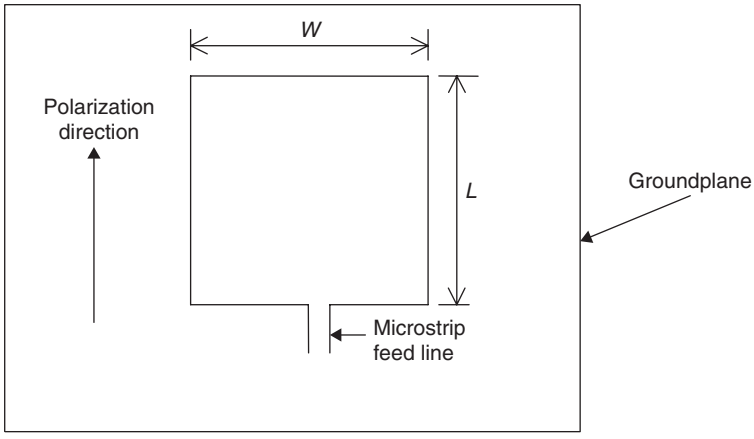


Figure 5.4: Patch Antenna

The impedance at the feed point depends on the width W of the patch. A microstrip transforms it to the required load (for transmitter) or source (for receiver) impedance. The feed point impedance can be made to match a transmission line directly by moving the feed point from the edge on the centerline toward the center of the board. In this way a 50-ohm coax transmission line can be connected directly to the underside of the patch antenna, with the center conductor going to the feed point through a via and the shield soldered to the groundplane.

The Mathcad "Patch Antenna" worksheet on the enclosed CDROM helps design a rectangular patch antenna. It includes calculations for finding the coax cable feed point location.

5.4 Impedance Matching

Impedance matching is important in transmitters and receivers for getting the best transfer of power between the antenna and the device. In a receiver, matching is often done in two stages—matching the receiver input to 50 ohms to suit a bandpass filter and to facilitate laboratory sensitivity measurements, and then matching from 50 ohms to the antenna

impedance. Receiver modules most often have 50 ohms input impedance. Receiver integrated circuits or low-noise RF amplifiers may have 50 ohms input, or the input impedance may be specified for various frequencies of operation. Sometimes a particular source impedance is specified that the input RF stage must "see" in order to obtain minimum noise figure.

Impedances to be matched are specified in different ways in component or module data sheets. A complex impedance may be specified directly for the operating frequency or for several possible operating frequencies. Another type of specification is by a resistance with capacitor or inductor in parallel or in series. The degree of matching to a specified impedance, usually 50 ohms, can be expressed by the *reflection coefficient*. This will be discussed later.

There are various circuit configurations that can be used for impedance matching and we present some simple examples here to match a pure resistance to a complex impedance. First, you should be able to express an impedance or resistance-reactance combination in parallel or serial form, whichever is convenient for the matching topography you wish to use. You can do this using the Mathcad worksheet "Impedance Transformations." Then use the worksheet "Impedance Matching" to find component values to match a wide range of impedances. The parallel or series source reactance must be separated from the total adjacent derived reactance value to get the value of the component to use in the matching circuit. Example 1 demonstrates this for a parallel source capacitor.

Remember that coils and capacitors are never purely reactive. The losses in coils, specified by the quality factor Q, are often significant whereas those in capacitors are usually ignored. In a parallel equivalent circuit (loss resistance in parallel to the reactance), $Q = R/X$ (X is the reactance). In a series equivalent circuit, $Q = X/R$. If the loss resistance is within a factor of up to around 5 of a resistance to be matched, it should be combined with that resistance before using the impedance matching formula. Example 2 shows how to do it.

Use nearest standard component values for the calculated values. Variable components may be needed, particularly in a high Q circuit. Remember also that stray capacitance and inductance will affect the matching and should be considered in selecting the matching circuit components.

Example 1

Figure 5.5 shows a circuit that can be used for matching a high impedance, such as may be found in a low-power transmitter, to 50 ohms.

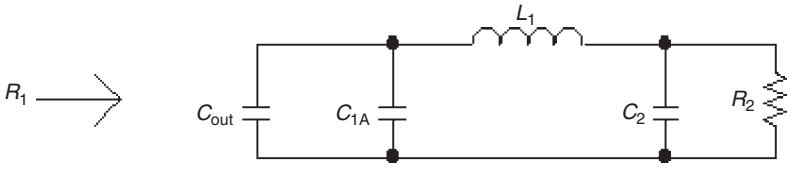


Figure 5.5: Impedance Transformation, Example 1

Let's use it to match a low-power transmitter output impedance of 1000 ohms (see the section "Transmitter Output Impedance") and 1.5-pF parallel capacitance to a 50-ohm bandpass filter or antenna. The frequency is 315 MHz. We use "Impedance Matching" worksheet, circuit 3.

Given values are:

$R_1 = 1000$ ohms $\quad C_{out} = 1.5$ pF $\quad R_2 = 50$ ohms $\quad f = 315$ MHz

(a) At the top of the worksheet, set $f = 315$ MHz. Under circuit (3) set R_1 and R_2 to 1000 and 50 ohms, respectively. Select a value for Q.

$Q = 10$

(b) Rounded off results are:

$C_1 = 5.1$ pF

$C_2 = 20.3$ pF

$L_1 = 101$ nH

(c) C_1 of the worksheet is made up of the parallel combination of C_{out} and C_{1A} of Figure 5.5:

$C_{1A} = C_1 - C_{out} = 5.1\,\text{pF} - 1.5\,\text{pF} = 3.6\,\text{pF}$

Example 2

We want to match the input of an RF mixer and IF amplifier integrated circuit (such as Philips NE605) to a 50-ohm antenna at 45 MHz. The equivalent input circuit is 4500 ohms in parallel with 2.5 pF. We choose to use a parallel coil L_{1A} having a value of 220 nH and a Q_{L1A} of 50. See Figure 5.6.

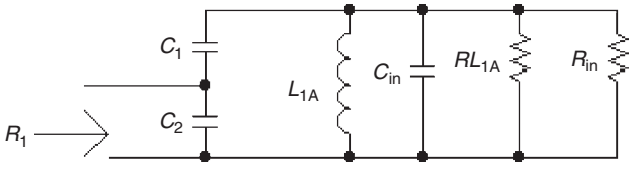


Figure 5.6: Impedance Transformation, Example 2

Given: $f = 45$ MHz, $R_1 = 50$ ohms, $R_{in} = 4.5$ K ohms, $C_{in} = 2.5$ pF, $L_{1A} = 220$ nH, $Q_{1A} = 50$

Find: C_1 and C_2

(a) Calculate RL_{1A}

$XL_{1A} = 62.2$ ohms (you can use "Conversions" worksheet)

$RL_{1A} = Q_{1A} \times XL_{1A} = 3110$ ohms

(b) Find equivalent input resistance to be matched, RL_1:

$RL_1 = RL_{1A} \parallel R_{in} = (3110 \times 4500)/(3110 + 4500) = 1839$ ohms

(c) Find equivalent parallel inductance L_1

$XC_{in} = -1415$ ohms ("Conversions" worksheet)

$XL_1 = XL_{1A} \parallel XC_{in} = (62.2 \times (-1415))/(62.2 - 1415) = 65.06$ ohms

(d) Find Q, which is needed for the calculation of C_1 and C_2 using the "Impedance Matching" worksheet:

$Q = RL_1/XL_1 = 28.27$

(e) Use the worksheet "Impedance Matching" circuit (4) to find C_1 and C_2, after specifying R_1, R_2, and Q:

$R_1 = 50$ ohms $\quad R_2 = RL_1 = 1839$ ohms $\quad Q = 28.27$

Results:

$C_1 = 65$ pF

$C_2 = 322$ pF

It may seem that the choice of the parallel inductor was arbitrary, but that's the designer's prerogative, as long as the resultant Q is greater than the minimum Q given in the worksheet example (in this case approximately 6). The choice of inductance determines the circuit Q, and consequently the bandwidth of the matching circuit. The total Q of the circuit includes the loading effect of the source resistance. Its value is one half the Q used in the design procedure, or 28.27/2 = approximately 14 in this example.

Example 3

We have a helical antenna with 15-ohm impedance that will be used with a receiver module having a 50-ohm input. The operating frequency is 173 MHz. The matching network is shown in Figure 5.7.

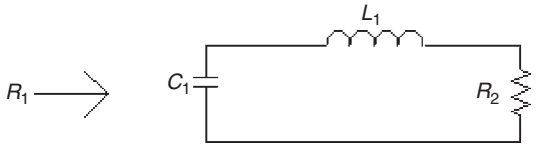


Figure 5.7: Impedance Transformation, Example 3

Given: $f = 173$ MHz, $R_1 = 50$ ohms, $R_2 = 15$ ohms

Use these values in the "Matching Impedance" worksheet, circuit (1), to get the matching network components:

$L_1 = 21.1$ nH

$C_1 = 28.1$ pF

5.4.1 Transmitter Output Impedance

In order to get maximum power transfer from a transmitter to an antenna, the RF amplifier output impedance must be known, as well as the antenna impedance, so that a matching network can be designed as shown in the previous section. For very low-power transmitters with radiated powers of tens of microwatts at the most, close matching is not critical. However, for a radiated power of 10 milliwatts and particularly when low-voltage lithium battery power is used, proper matching can save battery energy due to increased efficiency, and can generally simplify transmitter design. Besides, an output bandpass filter needs reasonably good matching to deliver a predictable frequency response.

A simplified estimate of an RF amplifier's output impedance R_L is given by the following expression:

$$R_{\mathrm{L}} = \frac{(V_{\mathrm{CC}} - V_{\mathrm{CE(sat)}})^2}{2P} \tag{5.9}$$

V_{CC} is the supply voltage to the RF stage, $V_{\mathrm{CE(sat)}}$ is the saturation voltage of the RF transistor at the operating frequency, and P is the power output.

5.4.2 Transmission Lines

In many short-range radio devices the transmitter or receiver antenna is an integral part of the device circuitry and is coupled directly to the transmitter output or receiver input circuit through discrete components. This is particularly the case with portable equipment. Devices with an external antenna located away from the equipment housing need a transmission line to connect the antenna to the input or output circuit. The transmission line is an example of a *distributed circuit* and it affects the coupling or transfer of the RF signal between the device RF circuit and the antenna. At high UHF and microwave frequencies, even the short connection between an internal antenna and the RF circuit is considered a transmission line whose characteristics must be designed to achieve proper impedance matching.

The transmission line can take several forms, among them coaxial cable, balanced two-wire cable, microstrip, and waveguide (a special case, not considered below).

A basic characteristic of a transmission line is its *characteristic impedance*. Its value depends on the capacitance per unit length C and inductance per unit length L, which in turn are functions of the physical characteristics of the line and the dielectric constant of the material surrounding the conductors. In the ideal case when there are no losses in the line the relationship is

$$Z_0 = \sqrt{\frac{L}{C}} \tag{5.10}$$

where L is the inductance per unit length in henrys and C is the capacitance per unit length in farads. Another important characteristic is the velocity factor, which is the ratio of the

propagation velocity, or phase velocity, of the wave in the line to the speed of light. The velocity factor depends on the dielectric constant, ε, of the material enclosing the transmission line conductors as

$$VF = \frac{1}{\sqrt{\varepsilon}} \tag{5.11}$$

Also important in specifying a transmission line, particularly at VHF and higher frequencies and relatively long lines, is the attenuation or line loss.

For example, the characteristics of a commonly used coaxial cable, RG-58C, are:

Characteristic impedance	50 ohms
Inductance per meter	0.25 microhenry
Capacitance per meter	101 picofarad
Velocity factor	0.66
Attenuation at 50 MHz	3.5 dB per 30 meters
Attenuation at 300 MHz	10 dB per 30 meters

In the previous section we talked about matching the antenna impedance to the circuit impedance. When the antenna is connected to the circuit through a transmission line, the impedance to be matched, seen at the circuit end of the transmission line, may be different from the impedance of the antenna itself. It depends on the characteristic impedance of the transmission line and the length of the line.

Several terms which define the degree of impedance matching, usually relating to transmission lines and antennas, are presented below.

Standing Wave Ratio is a term commonly used in connection with matching a transmission line to an antenna. When the load impedance differs from the characteristic impedance of the transmission line, the peak voltage on the line will differ from point to point. On a line whose length is greater than a half wavelength, the distance between voltage peaks or between voltage nulls is one-half wavelength. The ratio of the voltage peak to the voltage null is the standing wave ratio, abbreviated *SWR*, or *VSWR* (voltage standing wave ratio). The ratio of the peak current to the minimum current is the same as the *VSWR*, or *SWR*.

When the load is a pure resistance, R, and is larger than the characteristic impedance of the line (also considered to be a pure resistance), we have

$$SWR = R/Z_0 \tag{5.12}$$

If the load resistance is less than the line characteristic impedance, then

$$SWR = Z_0/R \tag{5.13}$$

In the general case where both the line impedance and the characteristic impedance may have reactance components, thus being complex, the voltage standing wave ratio is

$$SWR = \frac{1 + \left| \frac{Z_{\text{load}} - Z_0}{Z_{\text{load}} + Z_0} \right|}{1 - \left| \frac{Z_{\text{load}} - Z_0}{Z_{\text{load}} + Z_0} \right|} \tag{5.14}$$

Reflection Coefficient is the ratio of the voltage of the reflected wave from a load to the voltage of the forward wave absorbed by the load:

$$\rho = \frac{E_{\text{r}}}{E_{\text{f}}} \tag{5.15}$$

When the load is perfectly matched to the transmission line, the maximum power available from the generator is absorbed by the load, there is no reflected wave, and the reflection coefficient is zero. For any other load impedance, less power is absorbed by the load, and what remains of the available power is reflected back to the generator. When the load is an open or short circuit, or a pure reactance, all of the power is reflected back, the reflected voltage equals the forward voltage, and the reflection coefficient is unity. We can express the reflection coefficient in terms of the load impedance and characteristic impedance as

$$\rho = \frac{Z_{\text{load}} - Z_0}{Z_{\text{load}} + Z_0} \tag{5.16}$$

The relation between the standing wave ratio and the reflection coefficient is

$$SWR = \frac{1 + |\rho|}{1 - |\rho|} \tag{5.17}$$

Return Loss is an expression of the amount of power returned to the source relative to the available power from the generator. It is expressed in decibels as

$$RL = -20 \log (|\rho|) \tag{5.18}$$

Note that the return loss is always equal to or greater than zero.

Of the three terms relating to transmission line matching, the reflection coefficient gives the most information, since it is a complex number. As for the other two terms, *SWR* may be more accurate for large mismatches, whereas return loss presents values with greater resolution than *SWR* when the load impedance is close to the characteristic impedance of the line.

A plot of forward and reflected powers for a range of *SWRs* is given in Figure 5.8. This plot is convenient for seeing the effect of an impedance mismatch on the power actually dissipated in the load or accepted by the antenna, which is the forward power minus the reflected power.

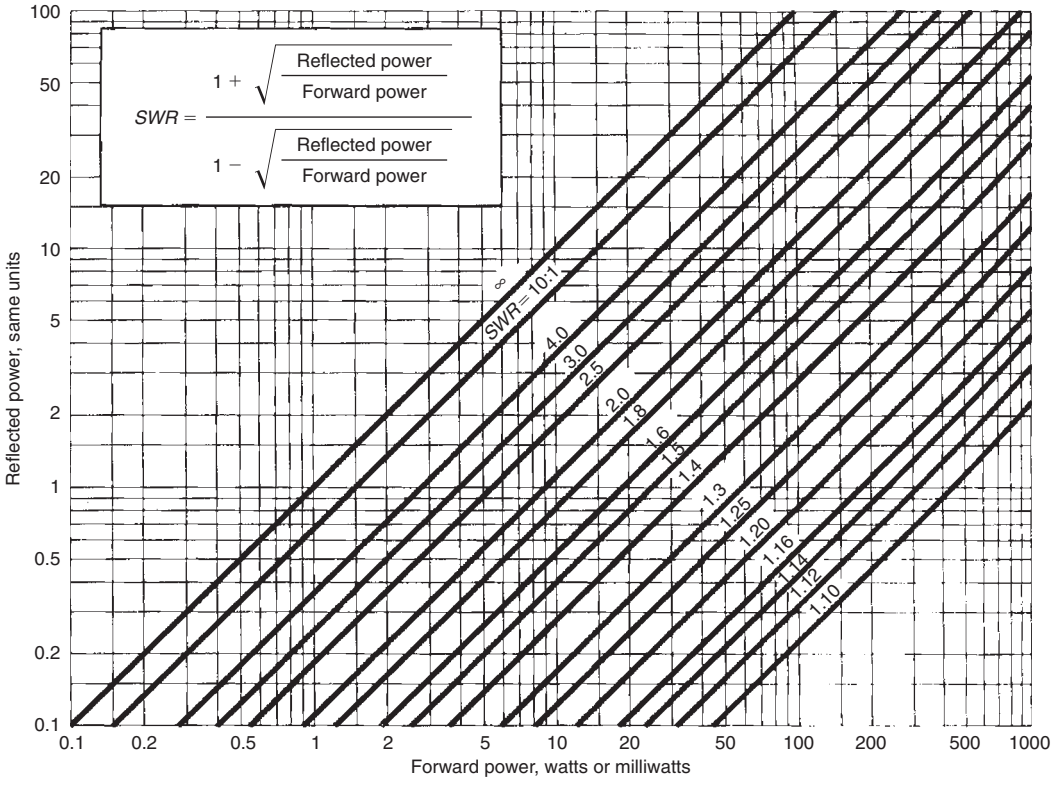


Figure 5.8: SWR, Forward and Reflected Power

Courtesy *Antenna Book, 16th Edition*, ARRL

Transmission line losses are not represented in the above definitions. Their effect is to reduce the *SWR* and increase the return loss, compared to a lossless line with the same load. This may seem to contradict the expressions given, which are in terms of load impedance, but that is not so. For instance, the load impedance in the expression for *SWR* equation [5.10]) is *the impedance at a particular point on the line where the SWR is wanted* and not necessarily the impedance at the end of the line. Thus, a long line with high losses may have a low *SWR* measured at the generator end, but a high *SWR* at the load. Transmission line loss is specified for a perfectly matched line, but when a mismatch exists, the loss is higher because of higher peak current and a resulting increased I^2R power dissipation in the line.

5.4.3 Smith Chart

A convenient tool for finding impedances in transmission lines and designing matching networks is the Smith chart, shown in Figure 5.9.

The Smith chart is a graph on which you can plot complex impedances and admittances (admittance is the inverse of impedance). An impedance value on the chart is the intersection of a resistance circle, labeled on the straight horizontal line in the middle, and a reactance arc,

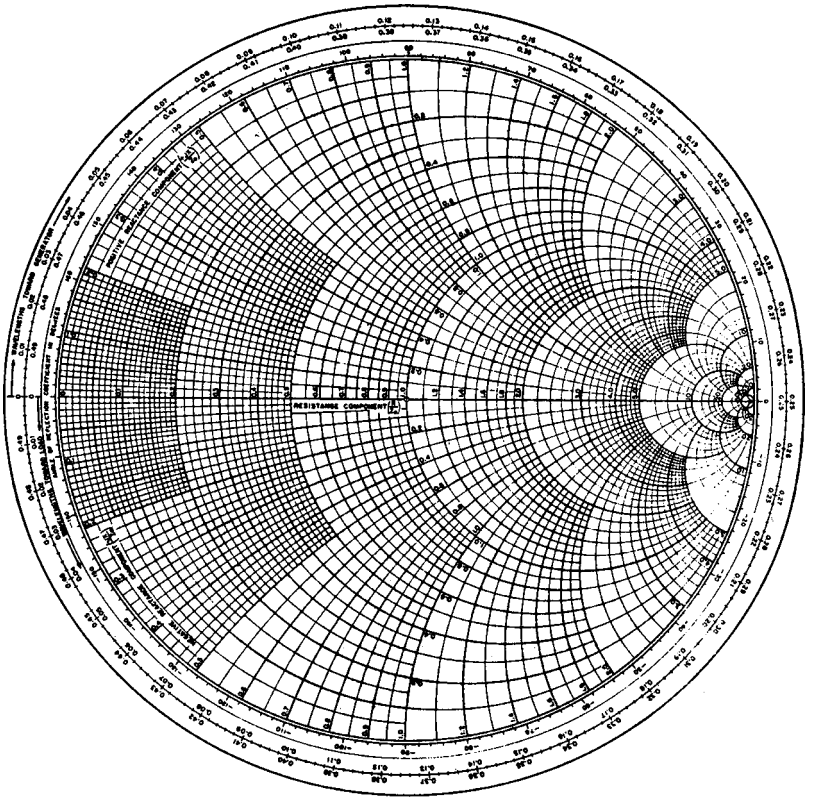


Figure 5.9: Smith Chart

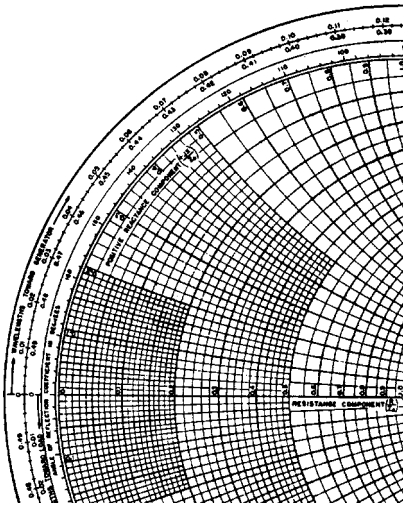


Figure 5.10: Expanded View of Smith Chart

labeled along the circumference of the "0" resistance circle. Figure 5.10 gives an expanded view of the chart with some of the labels. The unique form of the chart was devised for convenient graphical manipulation of impedances and admittances when designing matching networks, particularly when transmission lines are involved. The Smith chart is useful for dealing with distributed parameters which describe the characteristics of circuit board traces at UHF and microwave frequencies.

We'll describe some features of the Smith chart by way of an example. Let's say we need to match an antenna having an impedance of 15 ohms resistance in series with a capacitance reactance of 75 ohms to a transmitter with 50 ohms output impedance. The operating frequency is 173 MHz. The antenna is connected to the transmitter through 73 cm of RG-58C coaxial cable. What is the impedance at the transmitter that the matching network must convert to 50 ohms? The example is sketched in Figure 5.11 and Figure 5.12 shows the use of the Smith chart.

Step 1. First we mark the antenna impedance on the chart. Note that the resistance and reactance coordinates are normalized. The center of the chart, labeled 1.0, is the characteristic impedance of the transmission line, which is 50 ohms. We divide the resistance and capacitive reactance of the antenna by 50 and get, in complex form:

$$Z_{\text{load}} = 0.3 - j1.5 \tag{5.19}$$

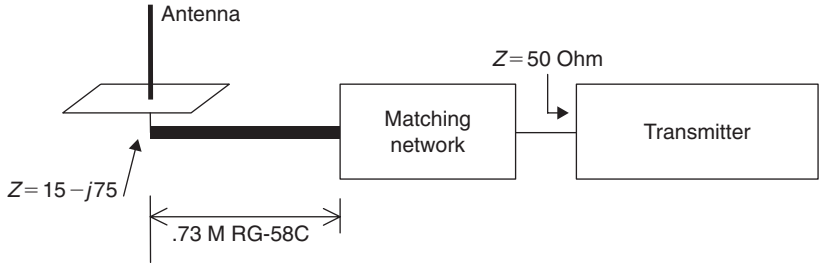


Figure 5.11: Antenna Matching Example

This is marked at the intersection of the 0.3 resistance circle with the 1.5 capacitive reactance coordinate in the bottom half of the chart. This point is marked "A" in Figure 5.12.

Step 2. The impedance at the transmitter end of the transmission line is located on a circle whose radius is the length of a line from the center of the chart to point "A" (assuming no cable losses). In order to find the exact location of the impedance on this circle for the 73-cm coax cable, we must relate the physical cable length, l, to the electrical length, L, in wavelengths.

$$L = \frac{1}{\eta\lambda} \tag{5.20}$$

where η is the velocity factor (.66) and λ is the wavelength in free space (3×10^8/frequency). Inserting the values for this example we find the electrical length of the line is 0.64 wavelengths.

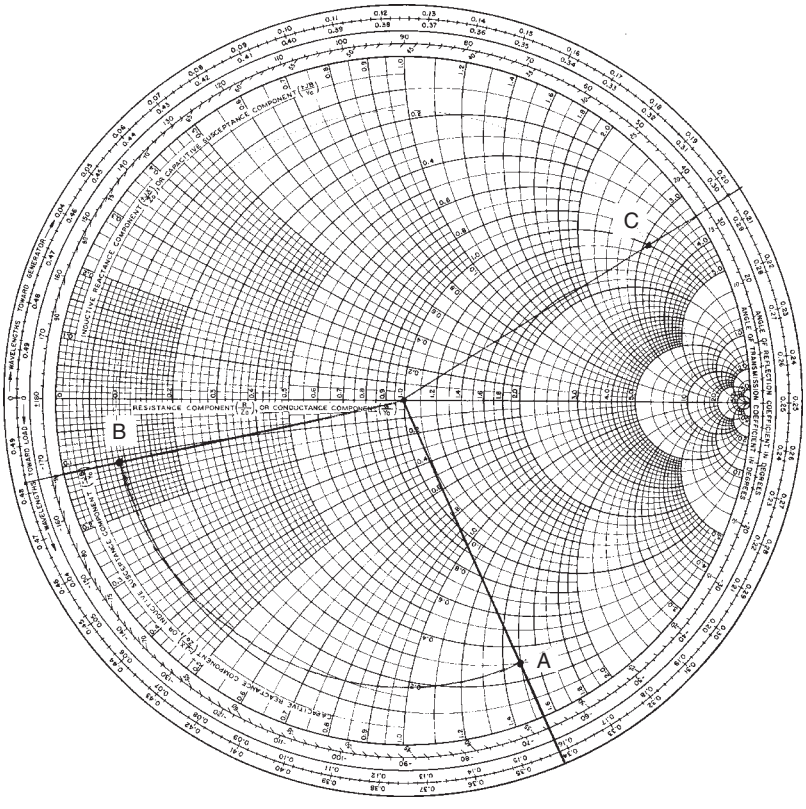


Figure 5.12: Using the Smith Chart

The Smith chart instructs us to move toward the generator in a clockwise direction from the load. The wavelength measure is marked on the outmost marked circle. Every 0.5 wavelengths, the load impedance is reflected to the end of the cable with no change, so we subtract a whole number of half wavelengths, in this case one, from the cable length, giving us

$$.64\lambda_c - .5\lambda_c = 0.14\lambda_c \tag{5.21}$$

where λ_c is the wavelength in the cable.

The line drawn on Figure 5.12 from the center through Z_{load} (point A) intersects the "wavelengths to generator" circle at 0.342. We add 0.14 to get 0.482 and draw a line from the center to this point. Mark the line at point B, which is the same distance from the center as point A. This is done conveniently with a compass.

Step 3. Read off point B. It is $0.1 - j0.113$. Multiplying by the 50-ohm characteristic impedance we get

$$Z_{gen} = 5 - j5.65 \text{ ohms} \tag{5.22}$$

Step 4. Now using the procedure of Example 3 above, we can design a matching network to match the impedance seen at the coax cable to the 50-ohm impedance of the transmitter. You have to add the capacitive reactance, 5.65 ohms, to the reactance you find for L_1. The resulting network components are $L_1 = 19$ nH and $C_1 = 55$ pF.

By examining the Smith chart in Figure 5.12, we note that if we use a longer coaxial cable we can get an impedance at its end that has a real part, or resistance, of 50 ohms, and an inductive reactance of $50 \times 3.1 = 155$ ohms. This is point C in the figure. We attain this impedance by adding $.22\lambda_c = 25$ cm to the original transmission line, for a total coax cable length of 98 cm. Now the only matching component we need is a series capacitor to cancel out the inductive reaction of 155 ohms. Using the "Conversions" worksheet we find its capacitance to be approximately 6 pF.

Other transmission-line matching problems can be solved using the Smith chart. Using the chart, you can easily determine *SWR*, reflection coefficient and return loss. The chart also has provision for accounting for line losses.

The Smith chart is very handy for seeing at a glance the effects on impedance of changing transmission line lengths, and also for using transmission lines as matching networks. Computer programs are available for doing Smith chart plotting. The enclosed Mathcad worksheet "Transmission Lines" solves transmission line problems directly from mathematical formulas.

5.4.4 Microstrip

From around 800 MHz and higher, the lengths of printed circuit board conductors are a significant fraction of a wavelength, so they act as transmission lines. Thus, if the input to a receiver integrated circuit or low-noise amplifier is 50 ohms, and a conductor length of 6 cm connects to the antenna socket, the RF plug from the antenna will *not* see 50 ohms, unless the conductor is designed to have a characteristic impedance of 50 ohms. A printed conductor over a ground plane (copper plating on the opposite side of the board) is called microstrip. The transmission line characteristics of conductors on a board are used in UHF and microwave circuits as matching networks between the various components.

Using the attached Mathcad worksheet "Microstrip," you can find the conductor width required to get a required characteristic line impedance, or you can find the impedance if you know the width. Then you can use the Smith chart to do impedance transformations and design matching networks using microstrip components. The "Microstrip" worksheet also gives you the wavelength on the pc board for a given frequency. In order to use this worksheet, you have to know the dielectric constant of your board material and the board's thickness.

5.5 Measuring Techniques

If you happen to have a vector analyzer, you can measure the impedances you want to match, design a matching network, and check the accuracy of your design. When a matching

network is designed and adjusted correctly, the impedance looking into the network where it is connected to the load or source is the complex conjugate of the impedance of the load or of the source impedance. The complex conjugate of an impedance has the same real part as the impedance and minus the imaginary part of the impedance. For example, if $Z_{source} = 30 - j12$ ohms, then the impedance seen at the input to the matching network should be $30 + j12$ ohms when its output port is connected to the load.

Without a vector analyzer, you need considerable cut-and-try to optimize the antenna and matching components. Other instruments, usually available in RF electronics laboratories, can be a big help. Here are some ideas for adjusting antennas and circuits for resonance at the operating frequency using relatively inexpensive equipment (compared to a vector analyzer).

A grid dip meter (still called that, although for years it has been based on a transistor oscillator, not a vacuum tube) is a simple, inexpensive tool, popular with radio amateurs. It consists of a tunable RF oscillator with external coil, allowing it to be lightly coupled to a resonant circuit, which can be an antenna of almost any type. When the dip meter is tuned across the resonant frequency of the passive circuit under test, its indicating meter shows a current dip due to absorption of energy from the instrument's oscillator. A loop antenna with resonating capacitor is easy to adjust using this method. A dipole, ground plane, or helical antenna can also be checked for resonance by connecting a small one-turn loop to the antenna terminals with matching circuit components disconnected. Set the dip meter coil close to the loop and tune the instrument to find a dip.

The main limitation to the grid dip meter is its frequency range, usually no more than 250 MHz. Higher frequency resonances can be measured with a return loss bridge, also called directional bridge or impedance bridge. This device, which is an integral part of a scalar network analyzer, can be used with a spectrum analyzer and tracking generator or a noise source to give a relative display of return loss versus frequency.

The return loss bridge is a three-port device that indicates power reflected from a mismatched load. Figure 5.13 shows a diagram of the bridge. Power applied at the source port passes to the

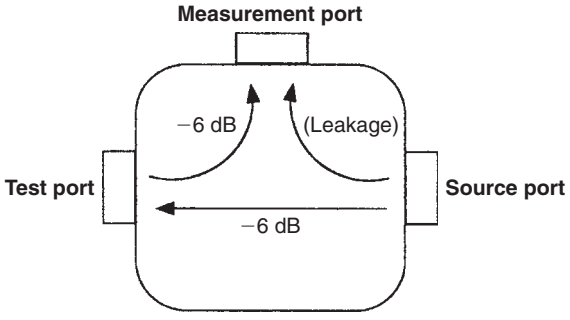


Figure 5.13: Return Loss Bridge

device under test at the test port with a nominal attenuation of 6 dB. Power reflected from the device under test appears at the measurement port, attenuated approximately 6 dB. If the tested circuit presents the same impedance as the characteristic impedance of the bridge, there will be no output at the measurement port, except for a leakage output on the order of 50 dB below the output of the source. If the test port sees an open or short circuit, all power will be reflected and the measurement port output will be around −12 dB. The return loss is the difference between the output measured in dBm at the measurement port when the test port is open or shorted, and the dBm output when the circuit under test is connected to the test port.

A setup to determine the resonant frequency of an antenna is shown in Figure 5.14 (a). The antenna is connected to the test terminal of the bridge through a short length of 50-ohm coaxial cable. A spectrum analyzer is connected to the measurement port and a tracking generator, whose frequency is swept in tandem with the frequency sweep of the spectrum analyzer, drives the source port.

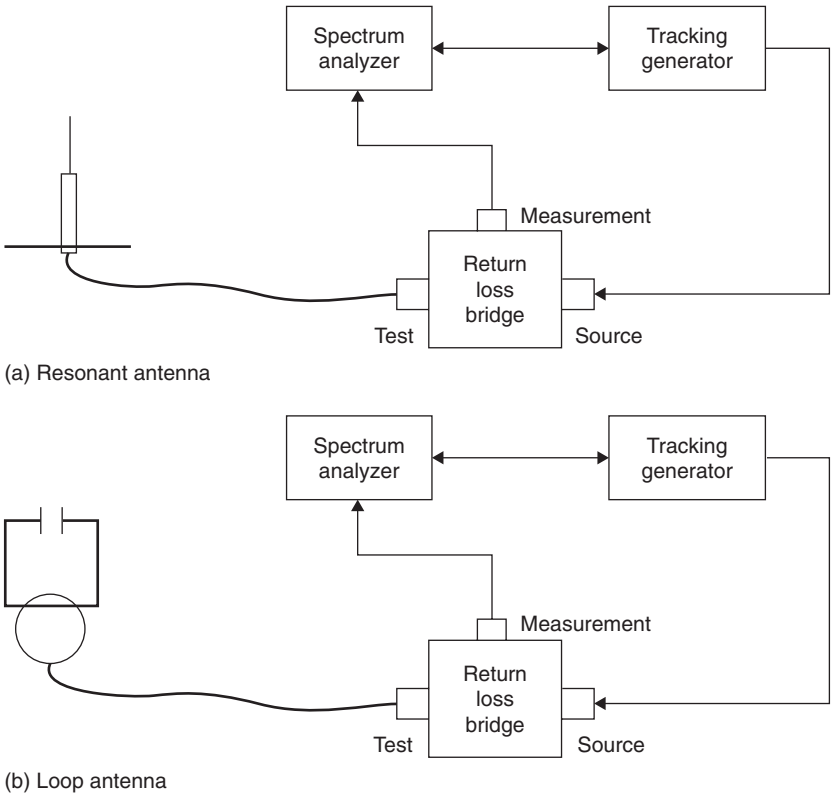


Figure 5.14: Resonant Circuit Test Setup

When the swept frequency passes the resonant frequency of the antenna, the analyzer display dips at that frequency. At the resonant frequency, reactance is cancelled and the antenna presents a pure resistance. The closer the antenna impedance is to 50 ohms, the deeper the

dip. Antenna parameters may be changed—length, loading coil or helical coil dimensions, for example—until the dip occurs at the desired operating frequency. Dips are usually observed at several frequencies because of more than one resonance in the system. By noting the effect of changes in the antenna on the various dips, as well as designing the antenna properly in the first place to give approximately the correct resonant frequency, the right dip can usually be correctly identified.

You can get an approximation of the resonant antenna resistance R_{ant} by measuring the return loss *RL* and then converting it to resistance using the following equations, derived from equations [5.16] and [5.18], or by using the Mathcad "Transmission Lines" worksheet.

$$\rho = \pm 10^{\frac{-RL}{20}} \tag{5.23}$$

$$R_{ant} = Z_0 \cdot \frac{1+\rho}{1-\rho} \tag{5.24}$$

The return loss is a positive value, so when solving for the reflection coefficient in equation [5.23], ρ can be either plus or minus, and R_{ant} found in equation [5.24] has two possible values. For example, if the return loss is 5 dB, the antenna resistance is either 14 ohms or 178 ohms. You decide between the two values using an educated guess. A monopole antenna over a ground plane, helically wound or having a loading coil, whose length is less than a quarter wave will have an impedance less than 50 ohms. Once the resistance is known, you can design a matching network as described above.

The arrangement shown in Figure 5.14(b) is convenient for checking the resonant frequency of a loop antenna, up to around 500 MHz. Use a short piece of coax cable and a loop of stiff magnet wire with a diameter of 2 cm. Use two or three turns in the loop for VHF and lower frequencies. At loop resonance, the spectrum analyzer display shows a sharp dip. Keep the test coil as far as possible from the loop, while still seeing the dip, to avoid influencing the circuit. You can easily tune the loop circuit, if it has a trimmer capacitor, by observing the location of the dip. The same setup can be used for checking resonance of tuning coils in the transmitter or receiver. There must be no radiation from the circuit when this test is made. If possible, disable the oscillator and apply power to the device being tested. The resonant frequency of a tuned circuit that is coupled to a transistor stage will be different when voltage is applied and when it is not.

5.6 Summary

We have covered in this chapter the most important properties of antennas and transmission lines that one needs to know to get the most from a short-range radio system. Antenna characteristics were defined. Then we discussed some of the types of antennas commonly used in short-range systems and gave examples of design. Impedance matching is imperative to get the most into,

and out of, an antenna, and we presented several matching circuits and gave examples of how to use them. We introduced the Smith chart, which may not be as widely used now as it once was as a design tool, but understanding it helps us visualize the concepts of circuit matching, particularly with distributed components.

Finally, we showed some simple measurements which help in realizing a design and which considerably shorten the cut-and-try routine that is almost inevitable when perfecting a product.

References

[5.1] Dacus, Farron L., Van Niekerk, Jan, and Bible, Steven, "Introducing Loop Antennas for Short-range Radios," Microwaves & RF, July 2002, p. 80.

[5.2] Drabowitch, S., Papiernik, A., Griffiths, H., Encinas, J. *Modern Antennas*, Chapman & Hall, London, 1998.

[5.3] Fujimoto, K., Henderson, A., Hirasawa, K., and James, J.R., *Small Antennas*, Research Studies Press, Ltd, England, 1987.

[5.4] Jaskik, Henry, Editor, *Antenna Engineering Handbook*, McGraw-Hill, NY, 1961.

[5.5] Milligan, Thomas A., *Modern Antenna Design*, McGraw-Hill, NY, 1985.

[5.6] Stutzman & Thiele, *Antenna Theory and Design*, John Wiley & Sons, Inc., 1998.

[5.7] Van Niekerk, Jan, Dacus, Farron L., and Bible, Steven, "Matching Loop Antennas to Short-range Radios," Microwaves & RF, August 2002, p. 72.

[5.8] Weeks, W. L., *Antenna Engineering*, McGraw Hill, NY, 1968.

[5.9] Yestrebsky, Tom, "MICRF001 Antenna Design Tutorial," Application Note 23, Micrel, Inc. San Jose, California.

This page intentionally left blank

CHAPTER 6

Communication Protocols and Modulation

Alan Bensky

In this chapter we take an overall view of the characteristics of the communication system. While these characteristics are common to any wireless communication link, for detail we'll address the peculiarities of short-range systems.

A simple block diagram of a digital wireless link is shown in Figure 6.1. The link transfers information originating at one location, referred to as source data, to another location where it is referred to as reconstructed data. A more concrete implementation of a wireless system, a security system, is shown in Figure 6.2.

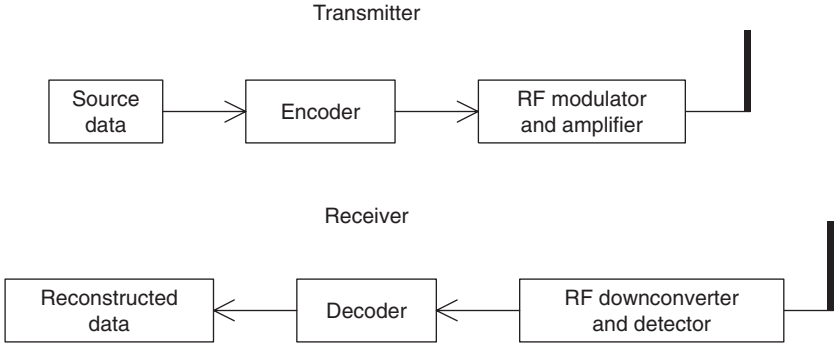


Figure 6.1: Radio Communication Link Diagram

6.1 Baseband Data Format and Protocol

Let's first take a look at what information we may want to transfer to the other side. This is important in determining what bandwidth the system needs.

6.1.1 Change-of-State Source Data

Many short-range systems only have to relay information about the state of a contact. This is true of the security system of Figure 6.2 where an infrared motion detector notifies the control panel when motion is detected. Another example is the push-button transmitter, which may be used as a panic button or as a way to activate and deactivate the control system, or

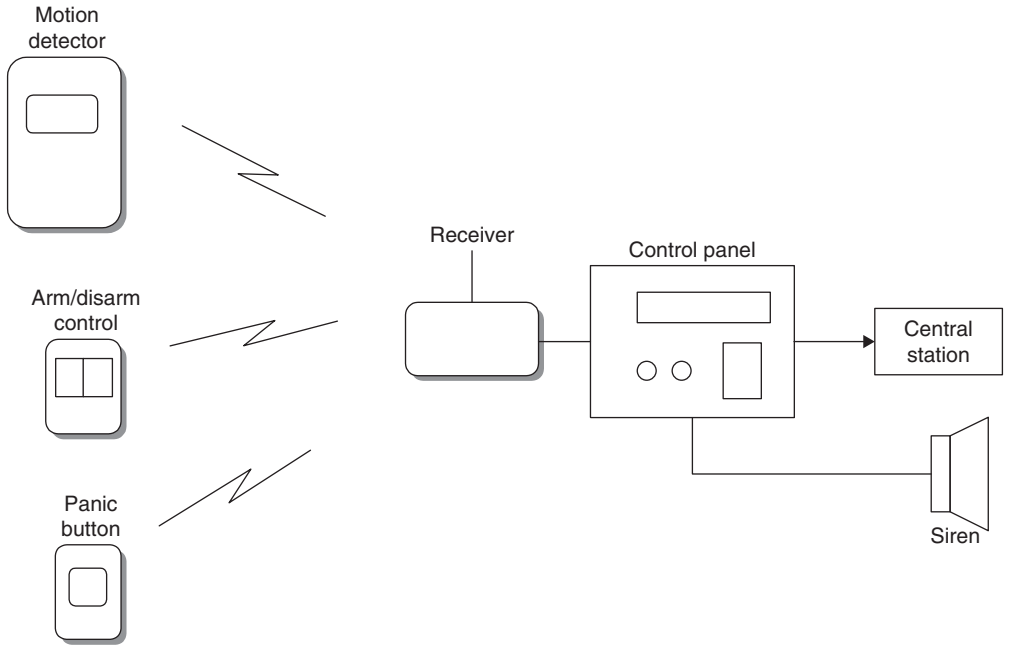


Figure 6.2: Security System

a wireless smoke detector, which gives advance warning of an impending fire. There are also what are often referred to as "technical" alarms—gas detectors, water level detectors, and low and high temperature detectors—whose function is to give notice of an abnormal situation.

All these examples are characterized as very low-bandwidth information sources. Change of state occurs relatively rarely, and when it does, we usually don't care if knowledge of the event is signaled tens or even hundreds of milliseconds after it occurs. Thus, required information bandwidth is very low—several hertz.

It would be possible to maintain this very low bandwidth by using the source data to turn on and off the transmitter at the same rate the information occurs, making a very simple communication link. This is not a practical approach, however, since the receiver could easily mistake random noise on the radio channel for a legitimate signal and thereby announce an intrusion, or a fire, when none occurred. Such false alarms are highly undesirable, so the simple on/off information of the transmitter must be coded to be sure it can't be misinterpreted at the receiver.

This is the purpose of the encoder shown in Figure 6.1. This block creates a group of bits, assembled into a frame, to make sure the receiver will not mistake a false occurrence for a real one. Figure 6.3 is an example of a message frame. The example has four fields. The first

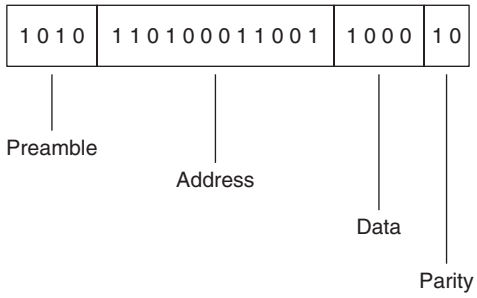


Figure 6.3: Message Frame

field is a preamble with start bit, which conditions the receiver for the transfer of information and tells it when the message begins. The next field is an identifying address. This address is unique to the transmitter and its purpose is to notify the receiver from where, or from what unit, the message is coming. The data field follows, which may indicate what type of event is being signaled, followed, in some protocols, by a parity bit or bits to allow the receiver to determine whether the message was received correctly.

6.1.1.1 Address Field

The number of bits in the address field depends on the number of different transmitters there may be in the system. Often the number of possibilities is far greater than this, to prevent confusion with neighboring, independent systems and to prevent the statistically possible chance that random noise will duplicate the address. The number of possible addresses in the code is 2^{L1}, where $L1$ is the length of the message field. In many simple security systems the address field is determined by dip switches set by the user. Commonly, eight to ten dip switch positions are available, giving 256 to 1024 address possibilities. In other systems, the address field, or device identity number, is a code number set in the unit microcontroller during manufacture. This code number is longer than that produced by dip switches, and may be 16 to 24 bits long, having 65,536 to 16,777,216 different codes. The longer codes greatly reduce the chances that a neighboring system or random event will cause a false alarm. On the other hand, the probability of detection is lower with the longer code because of the higher probability of error. This means that a larger signal-to-noise ratio is required for a given probability of detection.

In all cases, the receiver must be set up to recognize transmitters in its own system. In the case of dip-switch addressing, a dip switch in the receiver is set to the same address as in the transmitter. When several transmitters are used with the same receiver, all transmitters must have the same identification address as that set in the receiver. In order for each individual transmitter to be recognized, a subfield of two to four extra dip switch positions can be used for this differentiation. When a built-in individual fixed identity is used instead of

dip switches, the receiver must be taught to recognize the identification numbers of all the transmitters used in the system; this is done at the time of installation. Several common ways of accomplishing this are:

(a) *Wireless* "*learn*" *mode.* During a special installation procedure, the receiver stores the addresses of each of the transmitters which are caused to transmit during this mode;

(b) *Infrared transmission*. Infrared emitters and detectors on the transmitter and receiver, respectively, transfer the address information;

(c) *Direct key-in*. Each transmitter is labeled with its individual address, which is then keyed into the receiver or control panel by the system installer;

(d) *Wired learn mode*. A short cable temporarily connected between the receiver and transmitter is used when performing the initial address recognition procedure during installation.

Table 6.1: Advantages and Disadvantages of the Two Addressing Systems

Dip Switch	
Advantages	**Disadvantages**
Unlimited number of transmitters can be used with a receiver.	Limited number of bits increases false alarms and interference from adjacent systems.
Can be used with commercially available data encoders and decoders.	Device must be opened for coding during installation.
Transmitter or receiver can be easily replaced without recoding the opposite terminal.	Multiple devices in a system are not distinguishable in most simple systems.
	Control systems are vulnerable to unauthorized operation since the address code can be duplicated by trial and error.
Internal Fixed Code Identity	
Advantages	**Disadvantages**
Large number of code bits reduces possibility of false alarms.	Longer code reduces probability of detection.
System can be set up without opening transmitter.	Replacing transmitter or receiver involves redoing the code learning procedure.
Each transmitter is individually recognized by receiver.	Limited number of tranmitters can be used with each receiver.
	Must be used with a dedicated microcontroller. Cannot be used with standard encoders and decoders.

6.1.2 Code-Hopping Addressing

While using a large number of bits in the address field reduces the possibility of false identification of a signal, there is still a chance of purposeful duplication of a transmitter code to gain access to a controlled entry. Wireless push buttons are used widely for access control to vehicles and buildings. Radio receivers exist, popularly called "code grabbers," which receive the transmitted entry signals and allow retransmitting them for fraudulent access to a protected vehicle or other site. To counter this possibility, addressing techniques were developed that cause the code to change every time the push button is pressed, so that even if the transmission is intercepted and recorded, its repetition by a would-be intruder will not activate the receiver, which is now expecting a different code. This method is variously called code rotation, code hopping, or rolling code addressing. In order to make it virtually impossible for a would-be intruder to guess or try various combinations to arrive at the correct code, a relatively large number of address bits are used. In some devices, 36-bit addresses are employed, giving a total of over 68 billion possible codes.

In order for the system to work, the transmitter and receiver must be synchronized. That is, once the receiver has accepted a particular transmission, it must know what the next transmitted address will be. The addresses cannot be sequential, since that would make it too easy for the intruder to break the system. Also, it is possible that the user might press the push button to make a transmission but the receiver may not receive it, due to interference or the fact that the transmitter is too far away. This could even happen several times, further unsynchronizing the transmitter and the receiver. All of the code-hopping systems are designed to prevent such unsynchronization.

Following is a simplified description of how code hopping works, aided by Figure 6.4.

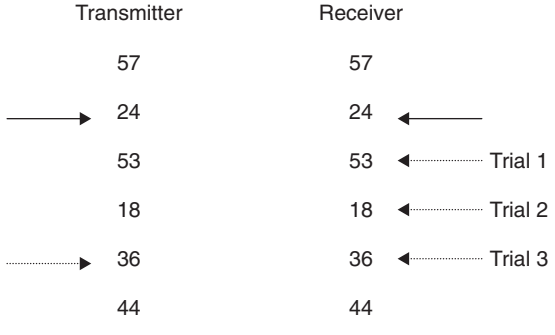


Figure 6.4: Code Hopping

Both the receiver and the transmitter use a common algorithm to generate a pseudorandom sequence of addresses. This algorithm works by manipulating the address bits in a certain fashion. Thus, starting at a known address, both sides of the link will create the same next

address. For demonstration purposes, Figure 6.4 shows the same sequence of two-digit decimal numbers at the transmitting side and the receiving side. The solid transmitter arrow points to the present transmitter address and the solid receiver arrow points to the expected receiver address. After transmission and reception, both transmitter and receiver calculate their next addresses, which will be the same. The arrows are synchronized to point to the same address during a system set-up procedure. As long as the receiver doesn't miss a transmission, there is no problem, since each side will calculate an identical next address. However, if one or more transmissions are missed by the receiver, when it finally does receive a message, its expected address will not match the received address. In this case it will perform its algorithm again to create a new address and will try to match it. If the addresses still don't match, a new address is calculated until either the addresses match or a given number of trials have been made with no success. At this point, the transmitter and receiver are unsynchronized and the original setup procedure has to be repeated to realign the transmitter and receiver addresses.

The number of trials permitted by the receiver may typically be between 64 and 256. If this number is too high, the possibility of compromising the system is greater (although with a 36-bit address a very large number of trials would be needed for this) and with too few trials, the frequency of inconvenient resynchronization would be greater. Note that a large number of trials takes a lot of time for computations and may cause a significant delay in response.

Several companies make rolling code components, among them Microchip, Texas Instruments, and National Semiconductor.

6.1.3 Data Field

The next part of the message frame is the data field. Its number of bits depends on how many pieces of information the transmitter may send to the receiver. For example, the motion detector may transmit three types of information: motion detection, tamper detection, or low battery.

6.1.3.1 Parity Bit Field

The last field is for error detection bits, or parity bits. As discussed later, some protocols have inherent error detection features so the last field is not needed.

6.1.3.2 Baseband Data Rate

Once we have determined the data frame, we can decide on the appropriate baseband data rate. For the security system example, this rate will usually be several hundred hertz up to a maximum of a couple of kilohertz. Since a rapid response is not needed, a frame can be repeated several times to be more certain it will get through. Frame repetition is needed in systems where space diversity is used in the receiver. In these systems, two separate antennas are periodically switched to improve the probability of reception. If signal nulling occurs at

one antenna because of the multipath phenomena, the other antenna will produce a stronger signal, which can be correctly decoded. Thus, a message frame must be sent more often to give it a chance to be received after unsuccessful reception by one of the antennas.

6.1.3.3 Supervision

Another characteristic of digital event systems is the need for link supervision. Security systems and other event systems, including medical emergency systems, are one-way links. They consist of several transmitters and one receiver. As mentioned above, these systems transmit relatively rarely, only when there is an alarm or possibly a low-battery condition. If a transmitter ceases to operate, due to a component failure, for example, or if there is an abnormal continuing interference on the radio channel, the fact that the link has been broken will go undetected. In the case of a security system, the installation will be unprotected, possibly, until a routine system inspection is carried out. In a wired system, such a possibility is usually covered by a normally energized relay connected through closed contacts to a control panel. If a fault occurs in the device, the relay becomes unenergized and the panel detects the opening of the contacts. Similarly, cutting the connecting wires will also be detected by the panel. Thus, the advantages of a wireless system are compromised by the lower confidence level accompanying its operation.

Many security systems minimize the risk of undetected transmitter failure by sending a supervisory signal to the receiver at a regular interval. The receiver expects to receive a signal during this interval and can emit a supervisory alarm if the signal is not received. The supervisory signal must be identified as such by the receiver so as not to be mistaken for an alarm.

The duration of the supervisory interval is determined by several factors:

- Devices certified under FCC Part 15 paragraph 15.231, which applies to most wireless security devices in North America, may not send regular transmissions more frequently than one per hour.
- The more frequently regular supervision transmissions are made, the shorter the battery life of the device.
- Frequent supervisory transmissions when there are many transmitters in the system raise the probability of a collision with an alarm signal, which may cause the alarm not to get through to the receiver.
- The more frequent the supervisory transmissions, the higher the confidence level of the system.

While it is advantageous to notify the system operator at the earliest sign of transmitter malfunction, frequent supervision raises the possibility that a fault might be reported when

it doesn't exist. Thus, most security systems determine that a number of consecutive missing supervisory transmissions must be detected before an alarm is given. A system which specifies security emissions once every hour, for example, may wait for eight missing supervisory transmissions, or eight hours, before a supervisory alarm is announced. Clearly, the greater the consequences of lack of alarm detection due to a transmitter failure, the shorter the supervision interval must be.

6.1.4 Continuous Digital Data

In other systems flowing digital data must be transmitted in real time and the original source data rate will determine the baseband data rate. This is the case in wireless LANs and wireless peripheral-connecting devices. The data is arranged in message frames, which contain fields needed for correct transportation of the data from one side to the other, in addition to the data itself.

An example of a frame used in *synchronous data link control* (SDLC) is shown in Figure 6.5. It consists of beginning and ending bytes that delimit the frame in the message, address and control fields, a data field of undefined length, and check bits or parity bits for letting the receiver check whether the frame was correctly received. If it is, the receiver sends a short acknowledgment and the transmitter can continue with the next frame. If no acknowledgment is received, the transmitter repeats the message again and again until it is received. This is called an ARQ (automatic repeat query) protocol. In high-noise environments, such as encountered on radio channels, the repeated transmissions can significantly slow down the message throughput.

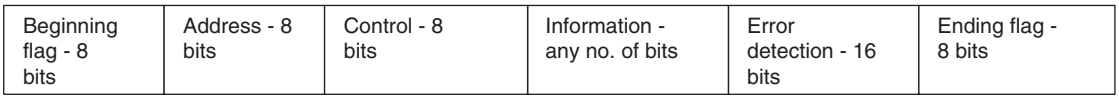


Figure 6.5: Synchronous Data Link Control Frame

More common today is to use a *forward error control* (FEC) protocol. In this case, there is enough information in the parity bits to allow the receiver to correct a small number of errors in the message so that it will not have to request retransmission. Although more parity bits are needed for error correction than for error detection alone, the throughput is greatly increased when using FEC on noisy channels.

In all cases, we see that extra bits must be included in a message to insure proper transmission, and the consequently longer frames require a higher transmission rate than what would be needed for the source data alone. This message overhead must be considered in determining the required bit rate on the channel, the type of digital modulation, and consequently the bandwidth.

6.1.5 Analog Transmission

Analog transmission devices, such as wireless microphones, also have a baseband bandwidth determined by the data source. A high-quality wireless microphone may be required to pass 50 to 15,000 Hz, whereas an analog wireless telephone needs only 100 to 3000 Hz. In this case determining the channel bandwidth is more straightforward than in the digital case, although the bandwidth depends on whether AM or FM modulation is used. In most short-range radio applications, FM is preferred—narrowband FM for voice communications and wide-band FM for quality voice and music transmission.

6.2 Baseband Coding

The form of the information signal that is modulated onto the RF carrier we call here baseband coding. We refer below to both digital and analog systems, although strictly speaking the analog signal is not coded but is modified to obtain desired system characteristics.

6.2.1 Digital Systems

Once we have a message frame, composed as we have shown by address and data fields, we must form the information into signal levels that can be effectively transmitted, received, and decoded. Since we're concerned here with binary data transmission, the baseband coding selected has to give the signal the best chance to be decoded after it has been modified by noise and the response of the circuit and channel elements. This coding consists essentially of the way that zeros and ones are represented in the signal sent to the modulator of the transmitter.

There are many different recognized systems of baseband coding. We will examine only a few common examples.

These are the dominant criteria for choosing or judging a baseband code:

(a) *Timing*. The receiver must be able to take a data stream polluted by noise and recognize transitions between each bit. The bit transitions must be independent of the message content—that is, they must be identifiable even for long strings of zeros or ones.

(b) *DC content*. It is desirable that the average level of the message—that is, its DC level—remains constant throughout the message frame, regardless of the content of the message. If this is not the case, the receiver detection circuits must have a frequency response down to DC so that the levels of the message bits won't tend to wander throughout the frame. In circuits where coupling capacitors are used, such a response is impossible.

(c) *Power spectrum*. Baseband coding systems have different frequency responses. A system with a narrow frequency response can be filtered more effectively to reduce noise before detection.

(d) *Inherent error detection*. Codes that allow the receiver to recognize an error on a bit-by-bit basis have a lower possibility of reporting a false alarm when error-detecting bits are not used.

(e) *Probability of error*. Codes differ in their ability to properly decode a signal, given a fixed transmitter power. This quality can also be stated as having a lower probability of error for a given signal-to-noise ratio.

(f) *Polarity independence*. There is sometimes an advantage in using a code that retains its characteristics and decoding capabilities when inverted. Certain types of modulation and demodulation do not retain polarity information. Phase modulation is an example.

Now let's look at some common codes (Figure 6.6) and rate them according to the criteria above. It should be noted that coding considerations for event-type reporting are far different from those for flowing real-time data, since bit or symbol times are not so critical, and message frames can be repeated for redundancy to improve the probability of detection and reduce false alarms. In data flow messages, the data rate is important and sophisticated error detection and correction techniques are used to improve system sensitivity and reliability.

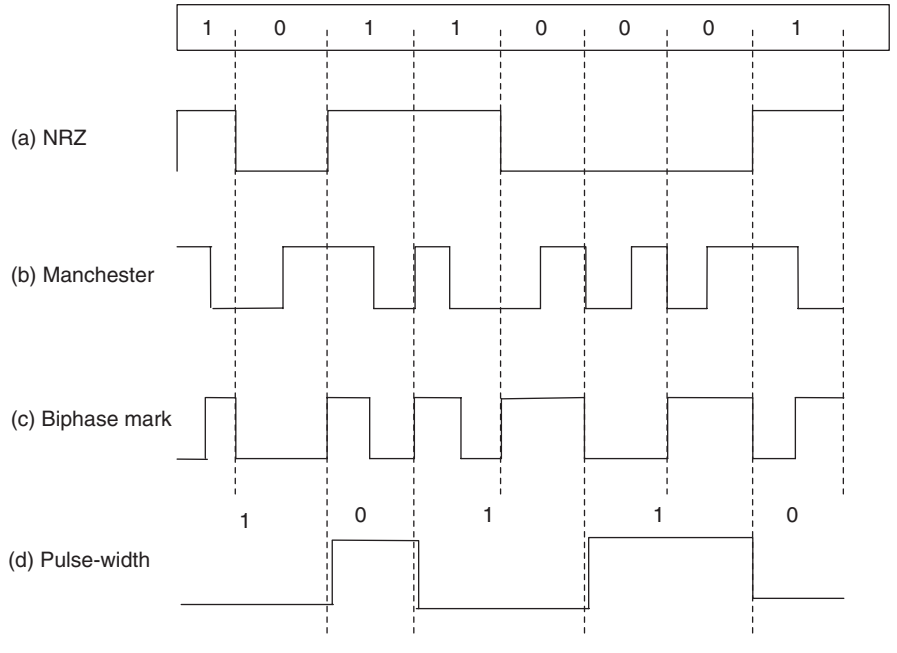


Figure 6.6: Baseband Bit Formats

(a) *Non-return to Zero (NRZ)—Figure 6.6a.*

This is the most familiar code, since it is used in digital circuitry and serial wired short-distance communication links, like RS-232. However, it is rarely used directly for wireless communication. Strings of ones or zeros leave it without defined bit boundaries, and its DC level is very dependent on the message content. There is no inherent error detection. If NRZ coding is used, an error detection or correction field is imperative. If amplitude shift keying (ASK) modulation is used, a string of zeros means an extended period of no transmission at all. In any case, if NRZ signaling is used, it should only be for very short frames of no more than eight bits.

(b) *Manchester code—Figure 6.6b.*

A primary advantage of this code is its relatively low probability of error compared to other codes. It is the code used in Ethernet local area networks. It gives good timing information since there is always a transition in the middle of a bit, which is decoded as zero if this is a positive transition and a one otherwise. The Manchester code has a constant DC component and its waveform doesn't change if it passes through a capacitor or transformer. However, a "training" code sequence should be inserted before the message information as a preamble to allow capacitors in the receiver detection circuit to reach charge equilibrium before the actual message bits appear. Inverting the Manchester code turns zeros to ones and ones to zeros. The frequency response of Manchester code has components twice as high as NRZ code, so a low-pass filter in the receiver must have a cut-off frequency twice as high as for NRZ code with the same bit rate.

(c) *Biphase Mark—Figure 6.6c.*

This code is somewhat similar to the Manchester code, but bit identity is determined by whether or not there is a transition in the middle of a bit. For biphase mark, a level transition in the middle of a bit (going in either direction) signifies a one, and a lack of transition indicates zero. Biphase space is also used, where the space character has a level transition. There is always a transition at the bit boundaries, so timing content is good. A lack of this transition gives immediate notice of a bit error and the frame should then be aborted. The biphase code has constant DC level, no matter what the message content, and a preamble should be sent to allow capacitor charge equalization before the message bits arrive. As with the Manchester code, frequency content is twice as much as for the NRZ code. The biphase mark or space code has the added advantage of being polarity independent.

(d) *Pulse width modulation—Figure 6.6d.*

As shown in the figure, a one has two timing durations and a zero has a pulse width of one duration. The signal level inverts with each bit so timing information for synchronization is good. There is a constant average DC level. Since the average pulse width varies with the

message content, in contrast with the other examples, the bit rate is not constant. This code has inherent error detection capability.

(e) *Motorola MC145026-145028 coding—Figure 6.7.*

Knowledge of the various baseband codes is particularly important if the designer creates his own protocol and implements it on a microcontroller. However, there are several off-the-shelf integrated circuits that are popular for simple event transmission transmitters where a microcontroller is not needed for other circuit functions, such as in panic buttons or door-opening controllers.

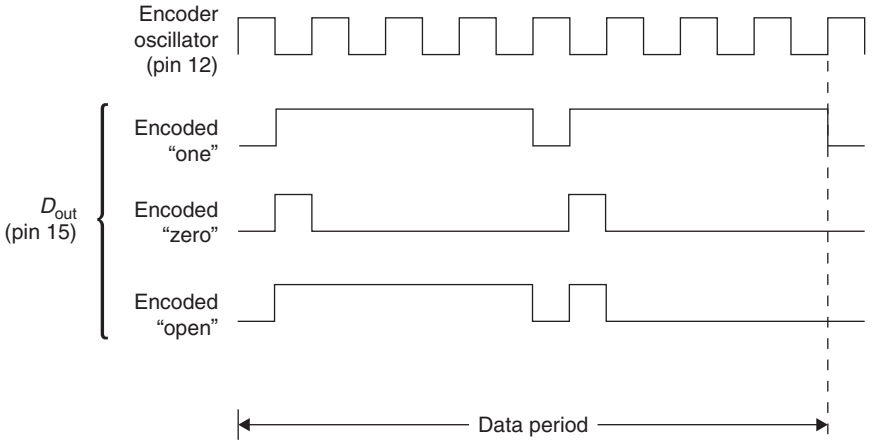


Figure 6.7: Motorola MC145026

The Motorola chips are an example of nonstandard coding developed especially for event transmission where three-state addressing is determined as high, low, or open connections to device pins. Thus, a bit symbol can be one of three different types. The receiver, MC145028, must recognize two consecutive identical frames to signal a valid message. The Motorola protocol gives very high reliability and freedom from false alarms. Its signal does have a broad frequency spectrum relative to the data rate and the receiver filter passband must be designed accordingly. The DC level is dependent on the message content.

6.2.2 Analog Baseband Conditioning

Wireless microphones and headsets are examples of short-range systems that must maintain high audio quality over the vagaries of changing path lengths and indoor environments, while having small size and low cost. To help them achieve this, they have a signal conditioning element in their baseband path before modulation. Two features used to achieve high signal-to-noise ratio over a wide dynamic range are pre-emphasis/deemphasis and compression/ expansion. Their positions in the transmitter/ receiver chain are shown in Figure 6.8.

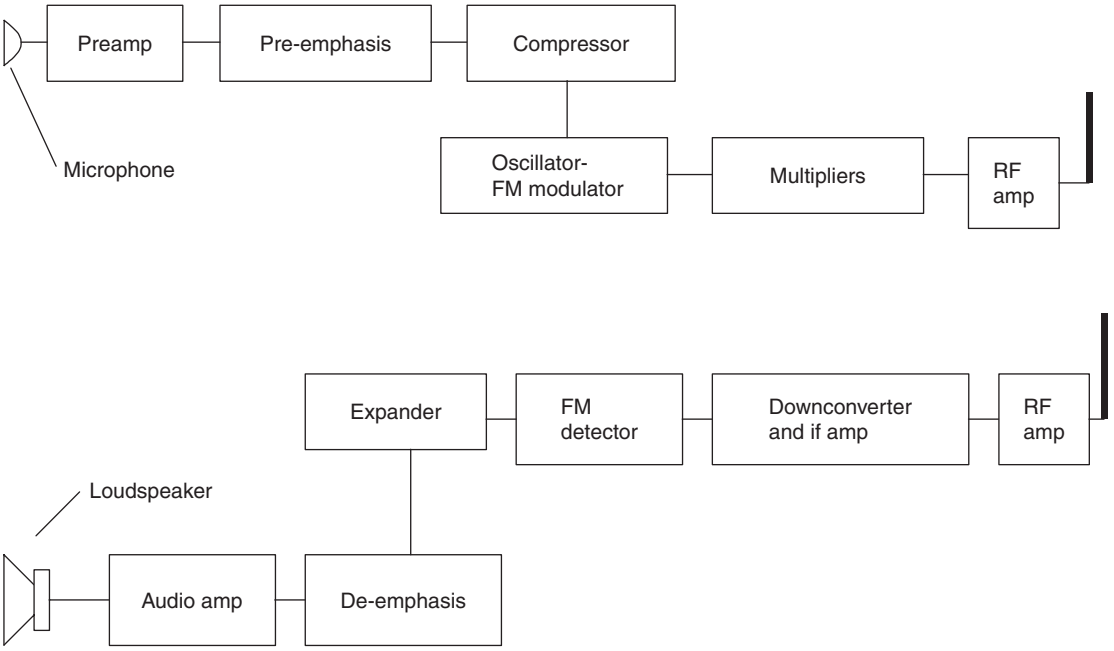


Figure 6.8: Wireless Microphone System

The transmitter audio signal is applied to a high-pass filter (preemphasis) which increases the high frequency content of the signal. In the receiver, the detected audio goes through a complementary low-pass filter (de-emphasis), restoring the signal to its original spectrum composition. However, in so doing, high frequency noise that entered the signal path after the modulation process is filtered out by the receiver while the desired signal is returned to its original quality.

Compression of the transmitted signal raises the weak sounds and suppresses strong sounds to make the modulation more efficient. Reversing the process in the receiver weakens annoying background noises while restoring the signal to its original dynamic range.

6.3 RF Frequency and Bandwidth

There are several important factors to consider when determining the radio frequency of a short-range system:

- Telecommunication regulations
- Antenna size
- Cost
- Interference
- Propagation characteristics

When you want to market a device in several countries and regions of the world, you may want to choose a frequency that can be used in the different regions, or at least frequencies that don't differ very much so that the basic design won't be changed by changing frequencies.

The UHF frequency bands are usually the choice for wireless alarm, medical, and control systems. The bands that allow unlicensed operation don't require rigid frequency accuracy. Various components—SAWs and ICs—have been specially designed for these bands and are available at low prices, so choosing these frequencies means simple designs and low cost. Many companies produce complete RF transmitter and receiver modules covering the most common UHF frequencies, among them 315 MHz and 902 to 928 MHz (U.S. and Canada), and 433.92 MHz band and 868 to 870 MHz (European Community).

Antenna size may be important in certain applications, and for a given type of antenna, its size is proportional to wavelength, or inversely proportional to frequency. When spatial diversity is used to counter multipath interference, a short wavelength of the order of the size of the device allows using two antennas with enough spacing to counter the nulling that results from multipath reflections. In general, efficient built-in antennas are easier to achieve in small devices at short wavelengths.

From VHF frequencies and up, cost is directly proportional to increased frequency.

Natural and manmade background noise is higher on the lower frequencies. On the other hand, certain frequency bands available for short-range use may be very congested with other users, such as the ISM bands. Where possible, it is advisable to chose a band set aside for a particular use, such as the 868–870 MHz band available in Europe.

Propagation characteristics also must be considered in choosing the operating frequency. High frequencies reflect easily from surfaces but penetrate insulators less readily than lower frequencies.

The radio frequency bandwidth is a function of the baseband bandwidth and the type of modulation employed. For security event transmitters, the required bandwidth is small, of the order of several kilohertz. If the complete communication system were designed to take advantage of this narrow bandwidth, there would be significant performance advantages over the most commonly used systems having a bandwidth of hundreds of kilohertz. For given radiated transmitter power, the range is inversely dependent on the receiver bandwidth. Also, narrow-band unlicensed frequency allotments can be used where available in the different regions, reducing interference from other users. However, cost and complexity considerations tend to outweigh communication reliability for these systems, and manufacturers decide to make do with the necessary performance compromises. The bandwidth of the mass production security devices is thus determined by the frequency stability of the transmitter and receiver frequency determining elements, and not by the required signaling bandwidth. The commonly used SAW devices dictate a bandwidth of at least 200 kHz, whereas the signaling bandwidth

may be only 2 kHz. Designing the receiver with a passband of 20 kHz instead of 200 kHz would increase sensitivity by 10 dB, roughly doubling the range. This entails using stable crystal oscillators in the transmitter and in the receiver local oscillator.

6.4 Modulation

Amplitude modulation (AM) and frequency modulation (FM), known from commercial broadcasting, have their counterparts in modulation of digital signals, but you must be careful before drawing similar conclusions about the merits of each. The third class of modulation is phase modulation, not used in broadcasting, but its digital counterpart is commonly used in high-end, high-data-rate digital wireless communication.

Digital AM is referred to as ASK—amplitude shift keying—and sometimes as OOK—on/off keying. FSK is frequency shift keying, the parallel to FM. Commercial AM has a bandwidth of 10 kHz whereas an FM broadcasting signal occupies 180 kHz. The high post-detection signal-to-noise ratio of FM is due to this wide bandwidth. However, on the negative side, FM has what is called a threshold effect, also due to wide bandwidth. Weak FM signals are unintelligible at a level that would still be usable for AM signals. When FM is used for two-way analog communication, narrow-band FM, which occupies a similar bandwidth to AM, also has comparable sensitivity for a given S/N.

6.4.1 Modulation for Digital Event Communication

For short-range digital communication we're not interested in high fidelity, but rather high sensitivity. Other factors for consideration are simplicity and cost of modulation and demodulation. Let's now look into the reasons for choosing one form of modulation or the other.

An analysis of error rates versus bit energy to noise density shows that there is no inherent advantage of one system, ASK or FSK, over the other. This conclusion is based on certain theoretical assumptions concerning bandwidth and method of detection. While practical implementation methods may favor one system over the other, we shouldn't jump to conclusions that FSK is necessarily the best, based on a false analogy to FM and AM broadcasting.

In low-cost security systems, ASK is the simplest and cheapest method to use. For this type of modulation we must just turn on and turn off the radio frequency output in accordance with the digital modulating signal. The output of a microcontroller or dedicated coding device biases on and off a single SAW-controlled transistor RF oscillator. Detection in the receiver is also simple. It may be accomplished by a diode detector in several receiver architectures, to be discussed later, or by the RSSI (received signal strength indicator) output of many superheterodyne receiver ICs employed today. Also, ASK must be used in the still widespread superregenerative receivers.

For FSK, on the other hand, it's necessary to shift the transmitting frequency between two different values in response to the digital code.

More elaborate means is needed for this than in the simple ASK transmitter, particularly when crystal or SAW devices are used to keep the frequency stable. In the receiver, also, additional components are required for FSK demodulation as compared to ASK. We have to decide whether the additional cost and complexity is worthwhile for FSK.

In judging two systems of modulation, we must base our results on a common parameter that is a basis for comparison. This may be peak or average power. For FSK, the peak and average powers are the same. For ASK, average power for a given peak power depends on the duty cycle of the modulating signal. Let's assume first that both methods, ASK and FSK, give the same performance—that is, the same sensitivity—if the average power in both cases are equal. It turns out that in this case, our preference depends on whether we are primarily marketing our system in North America or in Europe. This is because of the difference in the definition of the power output limits between the telecommunication regulations in force in the US and Canada as compared to the common European regulations.

The US FCC Part 15 and similar Canadian regulations specify an *average* field strength limit. Thus, if the transmitter is capable of using a peak power proportional to the inverse of its modulation duty cycle, while maintaining the allowed average power, then under our presumption of equal performance for equal average power, there would be no reason to prefer FSK, with its additional complexity and cost, over ASK.

In Western Europe, on the other hand, the low-power radio specification, ETSI 300 220 limits the *peak* power of the transmitter. This means that if we take advantage of the maximum allowed peak power, FSK is the proper choice, since for a given peak power, the average power of the ASK transmitter will always be less, in proportion to the modulating signal duty cycle, than that of the FSK transmitter.

However, is our presumption of equal performance for equal average power correct? Under conditions of added white Gaussian noise (AWGN) it seems that it is. This type of noise is usually used in performance calculations since it represents the noise present in all electrical circuits as well as cosmic background noise on the radio channel. But in real life, other forms of interference are present in the receiver passband that have very different, and usually unknown, statistical characteristics from AWGN. On the UHF frequencies normally used for short-range radio, this interference is primarily from other transmitters using the same or nearby frequencies. To compare performance, we must examine how the ASK and FSK receivers handle this type of interference. This examination is pertinent in the US and Canada where we must choose between ASK and FSK when considering that the average power, or signal-to-noise ratio, remains constant. Some designers believe that a small duty cycle resulting in high peak power per bit is advantageous since the presence of the bit, or high peak

signal, will get through a background of interfering signals better than another signal with the same average power but a lower peak. To check this out, we must assume a fair and equal basis of comparison. For a given data rate, the low-dutycycle ASK signal will have shorter pulses than for the FSK case. Shorter pulses means higher baseband bandwidth and a higher cutoff frequency for the post detection bandpass filter, resulting in more broadband noise for the same data rate. Thus, the decision depends on the assumptions of the type of interference to be encountered and even then the answer is not clear cut.

An analysis of the effect of different types of interference is given by Anthes of RF Monolithics (see references). He concludes that ASK, which does not completely shut off the carrier on a "0" bit, is marginally better than FSK.

6.4.2 Continuous Digital Communication

For efficient transmission of continuous digital data, the modulation choices are much more varied than in the case of event transmission. We can see this in the three leading cellular digital radio systems, all of which have the same use and basic requirements. The system referred to as DAMPS or TDMA (time division multiple access) uses a type of modulation called Pi/4 DPSK (differential phase shift keying). The GSM network is based on GMSK (Gaussian minimum shift keying). The third major system is CDMA (code division multiple access). Each system claims that its choice is best, but it is clear that there is no simple cut-and-dried answer. We aren't going into the details of the cellular systems here, so we'll look at the relevant trade-offs for modulation methods in what we have defined as short-range radio applications. At the end of this chapter, we review the basic principles of digital modulation and spread-spectrum modulation.

For the most part, license-free applications specify ISM bands where signals are not confined to narrow bandwidth channels. However, noise power is directly proportional to bandwidth, so the receiver bandwidth should be no more than is required for the data rate used. Given a data rate and an average or peak power limitation, there are several reasons for preferring one type of modulation over another. They involve error rate, implementation complexity, and cost. A common way to compare performance of the different systems is by curves of bit error rate (BER) versus the signal-to-noise ratio, expressed as energy per bit divided by the noise density (defined below). The three most common types of modulation system are compared in Figure 6.9. Two of the modulation types were mentioned above. The third, phase shift keying, is described below.

6.4.2.1 Phase Shift Keying (PSK)

Whereas in amplitude shift keying and frequency shift keying the amplitude and frequency are varied according to the digital source data, in PSK it is the phase of the RF carrier that is varied. In its simplest form, the waveform looks like Figure 6.10. Note that the phase of the carrier wave shifts 180 degrees according to the data signal bits. Similar to FSK, the carrier

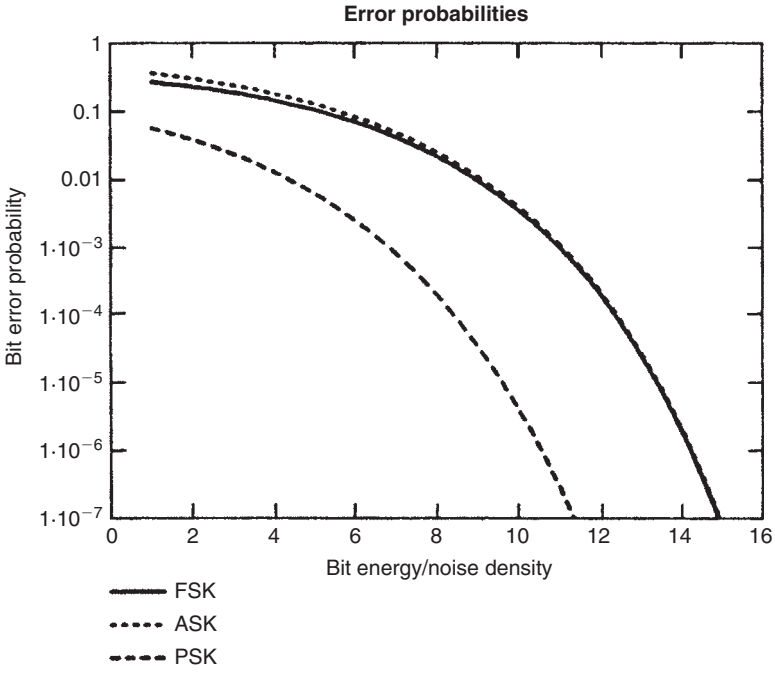


Figure 6.9: Bit Error Rates

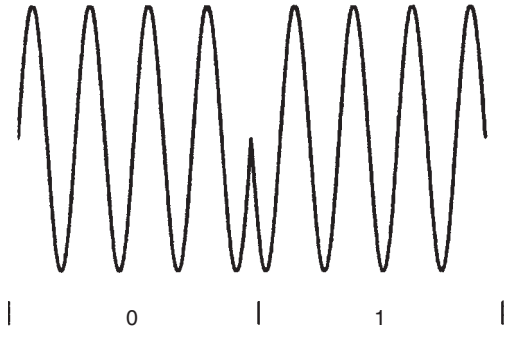


Figure 6.10: Phase Shift Keying

remains constant, thus giving the same advantage that we mentioned for FSK—maximum signal-to-noise ratio when there is a peak power limitation.

6.4.3 Comparing Digital Modulation Methods

In comparing the different digital modulation methods, we need a common reference parameter that reflects the signal power and noise at the input of the receiver, and the bit rate of the data. This common parameter of signal to noise ratio for digital systems is expressed as

the signal energy per bit divided by the noise density, E/N_o. We can relate this parameter to the more familiar signal to noise ratio, S/N, and the data rate R, as follows:

1. S/N = signal power/noise power. The noise power is the noise density N_o in watts/Hz times the transmitted signal bandwidth B_T in Hz:

$$S/N = S/(N_o B_T) \tag{6.1}$$

2. The signal energy in joules is the signal power in watts, S, times the bit time in seconds, which is 1/(data rate) = $1/R$, thus $E = S(1/R)$

3. The minimum transmitted bandwidth (Nyquist bandwidth) to pass a bit stream of R bits per second is $B_T = R$ Hz.

4. In sum:

$$E/N_o = (S/R)/N_o = S/N_o R = S/N$$

The signal power for this expression is the power at the input of the receiver, which is derived from the radiated transmitted power, the receiver antenna gain, and the path loss. The noise density N_o, or more precisely the one-sided noise power spectral density, in units of watts/Hz, can be calculated from the expression:

$$N_o = kT = 1.38 \times 10^{-23} \times T(\text{Kelvin}) \text{ watt/hertz}$$

The factor k is Boltzmann's constant and T is the equivalent noise temperature that relates the receiver input noise to the thermal noise that is present in a resistance at the same temperature T. Thus, at standard room temperature of 290 degrees Kelvin, the noise power density is 4×10^{-21} watts/Hz, or -174 dBm/Hz.

The modulation types making up the curves in Figure 6.9 are:

Phase shift keying (PSK)

Noncoherent frequency shift keying (FSK)

Noncoherent amplitude shift keying (ASK)

We see from the curves that the best type of modulation to use from the point of view of lowest bit error for a given signal-to-noise ratio (E/N_o) is PSK. There is essentially no difference, according to the curves, between ASK and FSK. (This is true only when noncoherent demodulation is used, as in most simple short range systems.) What then must we consider in making our choice?

PSK is not difficult to generate. It can be done by a balanced modulator. The difficulty is in the receiver. A balanced modulator can be used here too but one of its inputs, which switches the polarity of the incoming signal, must be perfectly correlated with the received signal

carrier and without its modulation. The balanced modulator acts as a multiplier and when a perfectly synchronized RF carrier is multiplied by the received signal, the output, after low-pass filtering, is the original bit stream. PSK demodulation is shown in Figure 6.11.

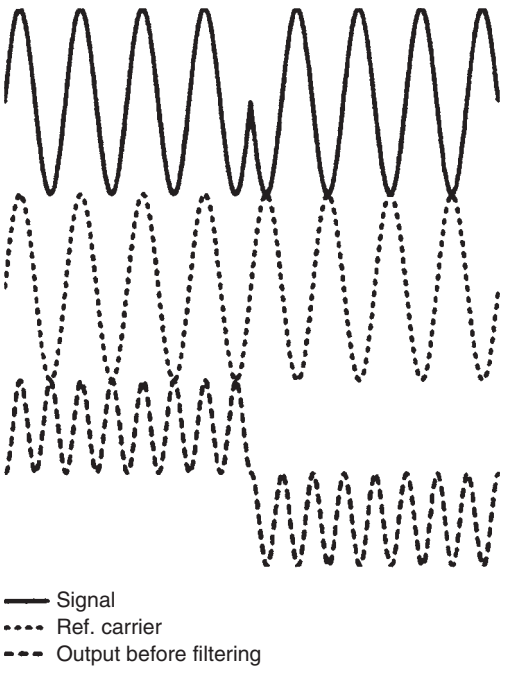


Figure 6.11: PSK Demodulation

There are several ways of generating the required reference carrier signal from the received signal. Two examples are the Costas loop and the squaring loop, which include three multiplier blocks and a variable frequency oscillator (VFO) [Ref. 6.2.]. Because of the complexity and cost, PSK is not commonly used in inexpensive short-range equipment, but it is the most efficient type of modulation for high-performance data communication systems.

Amplitude shift keying is easy to generate and detect, and as we see from Figure 6.9, its bit error rate performance is essentially the same as for FSK. However, FSK is usually the modulation of choice for many systems. The primary reason is that peak power is usually the limitation, which gives FSK a 3 dB advantage, since it has constant power for both bit states, whereas ASK has only half the average power, assuming a 50% duty cycle and equal probability of marks and spaces. FSK has slightly more complexity than ASK, and that's probably why it isn't used in all short-range digital systems.

A modulation system which incorporates the methods discussed above but provides a high degree of interference immunity is spread spectrum, which we'll discuss later on in this chapter.

6.4.4 Analog Communication

For short-range analog communication—wireless microphones, wireless earphones, auditive assistance devices—FM is almost exclusively used. When transmitting high-quality audio, FM gives an enhanced post-detection signal-to-noise ratio, at the expense of greater bandwidth. Even for narrow band FM, which doesn't have post-detection signal-to-noise enhancement, its noise performance is better than that of AM, because a limiting IF amplifier can be used to reduce the noise. AM, being a linear modulation process, requires linear amplifiers after modulation in the transmitter, which are less efficient than the class C amplifiers used for FM. Higher power conversion efficiency gives FM an advantage in battery-operated equipment.

6.4.5 Advanced Digital Modulation

Two leading characteristics of wireless communication in the last few years are the need for increasing data rates and better utilization of the radio spectrum. This translates to higher speeds on narrower bandwidths. At the same time, much of the radio equipment is portable and operated by batteries. So what is needed is:

- high data transmission rates
- narrow bandwidth
- low error rates at low signal-to-noise ratios
- low power consumption

Breakthroughs have occurred with the advancement of digital modulation and coding systems. We deal here with digital modulation principles.

The types of modulation that we discussed previously, ASK, FSK, and PSK, involve modifying a radio frequency carrier one bit at a time. We mentioned the Nyquist bandwidth, which is the narrowest bandwidth of an ideal filter which permits passing a bit stream without intersymbol interference. As the bandwidth of the digital bit stream is further reduced, the bits are lengthened and interfere with the detection of subsequent bits. This minimum, or Nyquist, bandwidth equals one-half of the bit rate at baseband, but twice as much for the modulated signal. We can see this result in Figure 6.12. An alternating series of marks and spaces can be represented by a sine

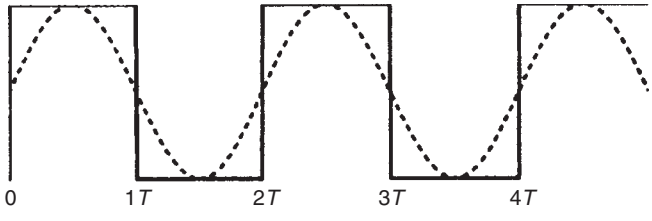


Figure 6.12: Nyquist Bandwidth

wave whose frequency is one-half the bit rate: $f_{sin} = 1/2T$. An ideal filter with a lower cutoff frequency will not pass this fundamental frequency component and the data will not get through.

Any other combination of bits will create other frequencies, all of which are lower than the Nyquist frequency. It turns out then that the maximum number of bits per hertz of filter cutoff frequency that can be passed at baseband is two. Therefore, if the bandwidth of a telephone line is 3.4 kHz, the maximum binary bit rate that it can pass without intersymbol interference is 6.8 k bits per second. We know that telephone line modems pass several times this rate. They do it by incorporating several bits in each *symbol* transmitted, for it is actually the symbol rate that is limited by the Nyquist bandwidth, and the problem that remains is to put several bits on each symbol in such a manner that they can be effectively taken off the symbol at the receiving end with as small as possible chance of error, given a particular *S*/*N*.

In Figure 6.13 we see three ways of combining bits with individual symbols, each of them based on one of the basic types of modulation—ASK, FSK, PSK. Each symbol duration *T* can carry one of four different values, or two bits. Using any one of the modulation types shown, the telephone line, or wireless link, can pass a bit rate twice as high as before over the same bandwidth. Combinations of these types are also employed, particularly of ASK and PSK, to put even more bits on a symbol. Quadrature amplitude modulation, QAM, sends several signal levels on four phases of the carrier frequency to give a high bandwidth efficiency—a high bit-rate relative to the signal bandwidth. It seems then that there is essentially no limit to the number of bits that could be compressed into a given bandwidth. If that were true, the 3.4-kHz

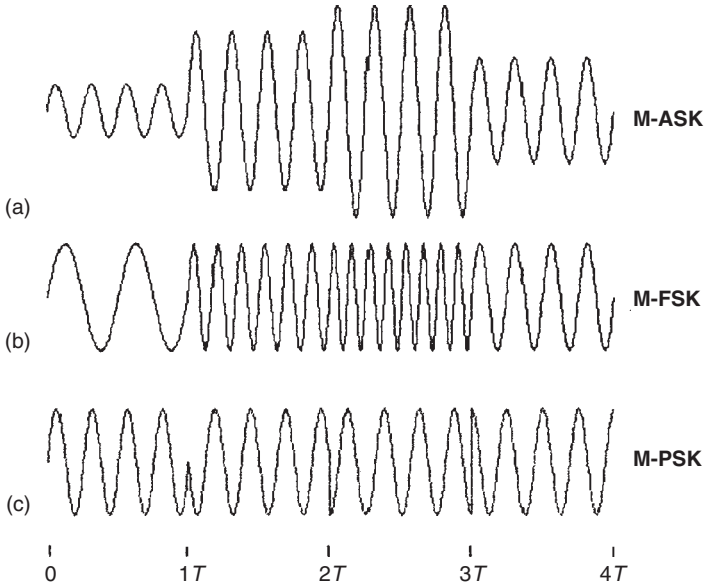


Figure 6.13: M-ary Modulation

telephone line could carry millions of bits per second, and the internet bottleneck to our homes would no longer exist. However, there is a very definite limit to the rate of information transfer over a transmission medium where noise is present, expressed in the Hartley-Shannon law:

$$C = W \log(1 + S/N) \tag{6.2}$$

This expression tells us that the maximum rate of information (the capacity C) that can be sent without errors on a communication link is a function of the bandwidth, W, and the signal-to-noise ratio, S/N.

In investigating the ways of modulating and demodulating multiple bits per symbol, we'll first briefly discuss a method not commonly used in short-range applications (although it could be). This is called M-ary FSK (Figure 6.13b) and in contrast to the aim we mentioned above of increasing the number of bits per hertz, it increases the required bandwidth as the number of bits per symbol is increased. "M" in "M-FSK" is the number of different frequencies that may be transmitted in each symbol period. The benefit of this method is that the required S/N per bit for a given bit error rate decreases as the number of bits per symbol increases. This is analogous to analog FM modulation, which uses a wideband radio channel, well in excess of the bandwidth of the source audio signal, to increase the resultant S/N. M-ary FSK is commonly used in point-to-point microwave transmission links for high-speed data communication where bandwidth limitation is no problem but the power limitation is.

Most of the high-data-rate bandwidth-limited channels use multiphase PSK or QAM. While there are various modulation schemes in use, the essentials of most of them can be described by the block diagram in Figure 6.14. This diagram is the basis of what may be called vector

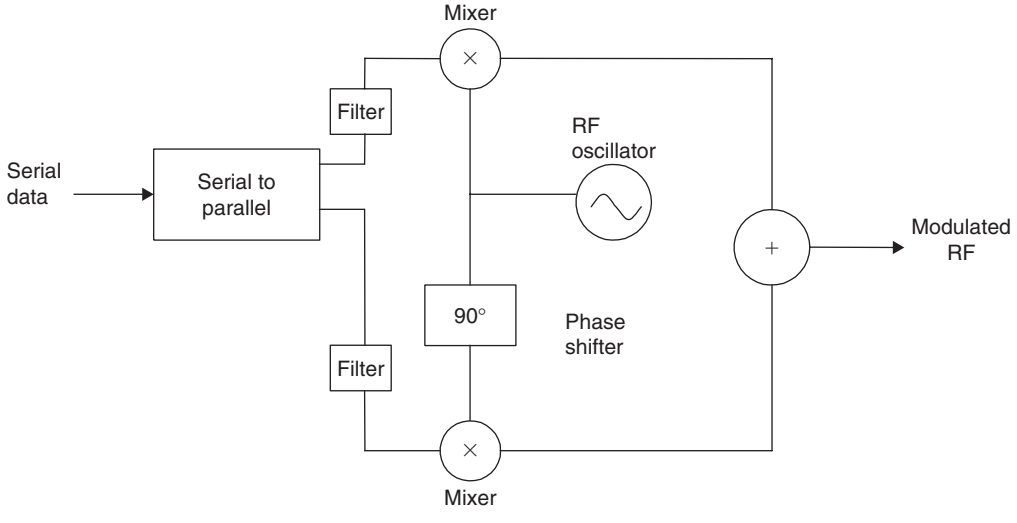


Figure 6.14: Quadrature Modulation

modulation, IQ modulation, or quadrature modulation. "I" stands for "in phase" and "Q" stands for "quadrature."

The basis for quadrature modulation is the fact that two completely independent data streams can be simultaneously modulated on the same frequency and carrier wave. This is possible if each data stream modulates coherent carriers whose phases are 90 degrees apart. We see in the diagram that each of these carriers is created from the same source by passing one of them through a 90-degree phase shifter. Now it doesn't matter which method of modulation is used for each of the phase shifted carriers. Although the two carriers are added together and amplified before transmission, the receiver, by reversing the process used in the transmitter (see Figure 6.15), can completely separate the incoming signal into its two components.

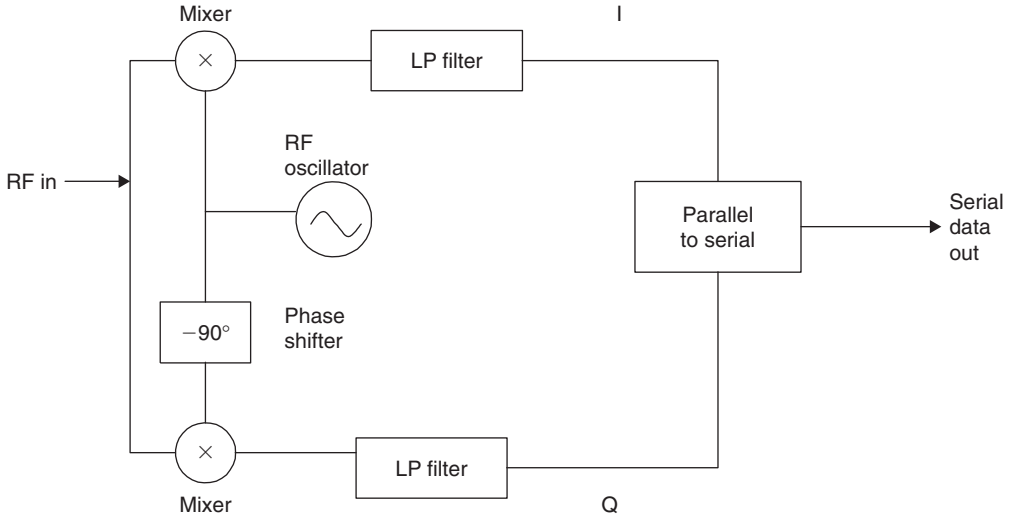


Figure 6.15: Quadrature Demodulation

The diagrams in Figures 6.14 and 6.15 demonstrate quadrature phase shift keying (QPSK) where a data stream is split into two, modulated independently bit by bit, then combined to be transmitted on a single carrier. The receiver separates the two data streams, then demodulates and combines them to get the original serial digital data. The phase changes of the carrier in response to the modulation is commonly shown on a vector or constellation diagram. The constellation diagram for QPSK is shown in Figure 6.16. The Xs on the plot are tips of vectors that represent the magnitude (distance from the origin) and phase of the signal that is the sum of the "I" and the "Q" carriers shown on Figure 6.14, where each is multiplied by -1 or $+1$, corresponding to bit values of 0 and 1. The signal magnitudes of all four possible combinations of the two bits is the same—$\sqrt{2}$ when the I and Q carriers have a magnitude of 1. I and Q bit value combinations corresponding to each vector are shown on the plot.

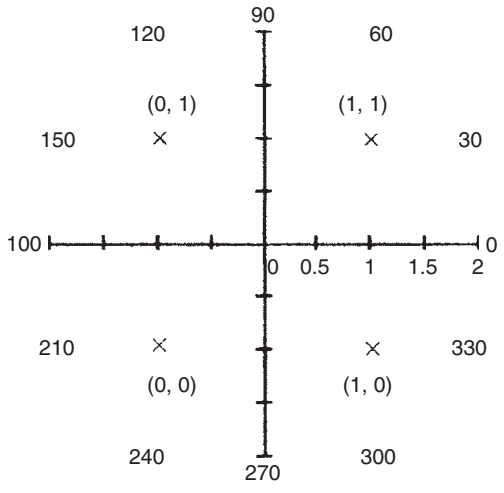


Figure 6.16: QPSK Constellation Diagram

We have shown that two data streams, derived from a data stream of rate R2, can be sent in parallel on the same RF channel and at the same bandwidth as a single data stream having half the bit rate, R1. At the receiver end, the demodulation process of each of the split bit streams is exactly the same as it would be for the binary phase shift modulation shown in Figure 6.11 above, and its error rate performance is the same as is shown for the BPSK curve in Figure 6.9. Thus, we've doubled the data rate on the same bandwidth channel while maintaining the same error rate as before. In short, we've doubled the efficiency of the communication.

However, there are some complications in adopting quadrature modulation. Similar to the basic modulation methods previously discussed, the use of square waves to modulate RF carriers causes unwanted sidebands that may exceed the allowed channel bandwidth. Thus, special low-pass filters, shown in Figure 6.14, are inserted in the signal paths before modulation. Even with these filters, the abrupt change in the phase of the RF signal at the change of data state will also cause increased sidebands. We can realize from Figure 6.13c that changes of data states may cause the RF carrier to pass through zero when changing phase. Variations of carrier amplitude make the signal resemble amplitude modulation, which requires inefficient linear amplifiers, as compared to nonlinear amplifiers that can be used for frequency modulation, for example.

Two variations of quadrature phase shift keying have been devised to reduce phase changes between successive symbols and to prevent the carrier from going through zero amplitude during phase transitions. One of these is offset phase shift keying. In this method, the I and Q data streams in the transmitter are offset in time by one-half of the bit duration, causing the carrier phase to change more often but more gradually. In other words, the new "I" bit will modulate the cosine carrier at half a bit time earlier (or later) than the time that the "Q" bit modulates the sine carrier.

The other variant of QPSK is called Pi/4 DPSK. It is used in the US TDMA (time division multiple access) digital cellular system. In it, the constellation diagram is rotated Pi/4 radians (45 degrees) at every bit time such that the carrier phase angle can change by either 45 or 135 degrees. This system reduces variations of carrier amplitude so that more efficient nonlinear power amplifiers may be used in the transmitter.

Another problem with quadrature modulation as described above is the need for a coherent local oscillator in the receiver in order to separate the in-phase and quadrature data streams. As for bipolar phase shift keying, this problem may be ameliorated by using differential modulation and by multiplying a delayed replica of the received signal by itself to extract the phase differences from symbol to symbol.

The principle of transmitting separate data streams on in phase and quadrature RF carriers may be extended so that each symbol on each carrier contains 2, 3, 4 or more bits. For example, 4 bits per each carrier vector symbol allows up to 16 amplitude levels per carrier and a total of 256 different states of amplitude and phase altogether. This type of modulation is called quadrature amplitude modulation (QAM), and in this example, 256-QAM, 8 data bits can be transmitted in essentially the same time and bandwidth as one bit sent by binary phase shift keying. The constellation of 16-QAM (4 bits per symbol) is shown in Figure 6.17.

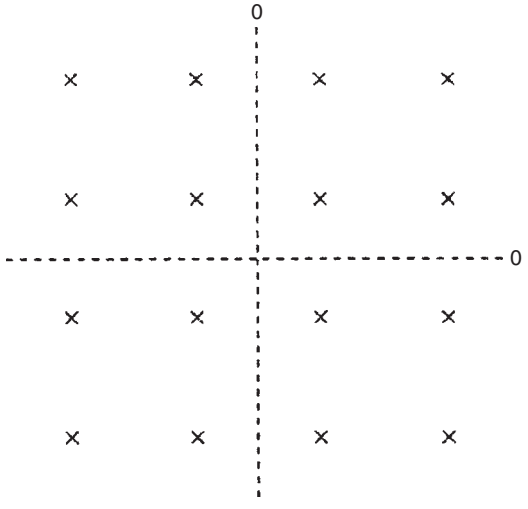


Figure 6.17: 16-QAM Constellation Diagram

Remember that the process of concentrating more and more bits to a carrier symbol cannot go on without limit. While the bit rate per hertz goes up, the required *S/N* of the channel for a given bit error rate (BER) also increases—that is, more power must be transmitted, or as a common alternative in modern digital communication systems, sophisticated coding

algorithms are incorporated in the data protocol. However, the ultimate limit of data rate for a particular communication channel with noise is the Hartley-Shannon limit stated above.

We now turn to examine another form of digital modulation that is becoming very important in a growing number of short-range wireless applications—spread-spectrum modulation.

6.4.6 Spread Spectrum

The regulations for unlicensed communication using unspecified modulation schemes, both in the US and in Europe, determine maximum power outputs ranging from tens of microwatts up to 10 milliwatts in most countries. This limitation greatly reduces the possibilities for wireless devices to replace wires and to obtain equivalent communication reliability. However, the availability of frequency bands where up to one watt may be transmitted in the US and 100 mW in Europe, under the condition of using a specified modulation system, greatly enlarges the possible scope of use and reliability of unlicensed short-range communication.

Spread spectrum has allowed the telecommunication authorities to permit higher transmitter powers because spread-spectrum signals can coexist on the same frequency bands as other types of authorized transmissions without causing undue interference or being unreasonably interfered with. The reason for this is evident from Figure 6.18. Figure 6.18a shows the spread-spectrum signal spread out over a bandwidth much larger than the narrow-band signals.

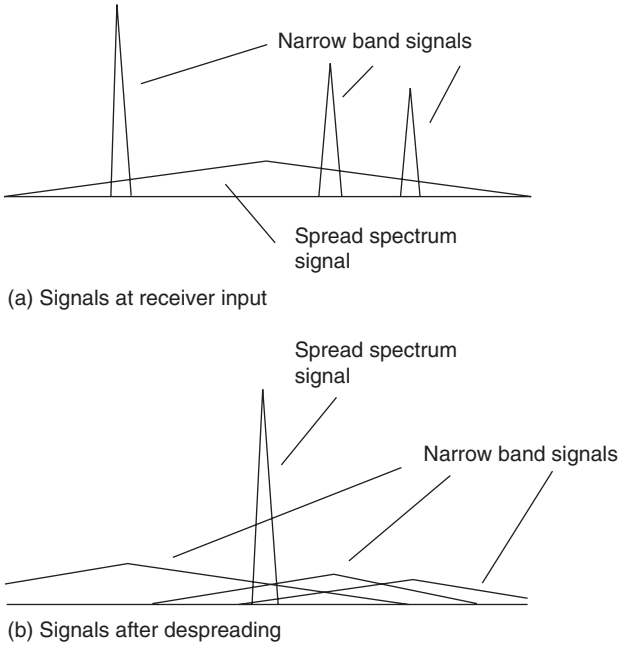


Figure 6.18: Spread Spectrum and Narrow-Band Signals

Although its total power if transmitted as a narrow band signal could completely overwhelm another narrow-band signal on the same or adjacent frequency, the part of it occupying the narrow-band signal bandwidth is small related to the total, so it doesn't interfere with it. In other words, spreading the transmitted power over a wide frequency band greatly reduces the signal power in a narrow bandwidth and thus the potential for interference.

Figure 6.18b shows how spread-spectrum processing reduces interference from adjacent signals. The despreading process concentrates the total power of the spread-spectrum signal into a narrow-band high peak power signal, whereas the potentially interfering narrow-band signals are spread out so that their power in the bandwidth of the desired signal is relatively low.

These are some advantages of spread spectrum modulation:

- FCC rules allow higher power for nonlicensed devices
- Reduces co-channel interference—good for congested ISM bands
- Reduces multipath interference
- Resists intentional and unintentional jamming
- Reduces the potential for eavesdropping
- Permits code division multiplexing of multiple users on a common channel.

There is sometimes a tendency to compare spread spectrum with wide-band frequency modulation, such as used in broadcasting, since both processes achieve performance advantages by occupying a channel bandwidth much larger than the bandwidth of the information being transmitted. However, it's important to understand that there are principal differences in the two systems.

In wide band FM (WBFM), the bandwidth is spread out directly by the amplitude of the modulating signal and at a rate determined by its frequency content. The result achieved is a signal-to-noise ratio that is higher than that obtainable by sending the same signal over baseband (without modulation) and with the same noise density as on the RF channel. In WBFM, the post detection signal-to-noise ratio (S/N) is a multiple of the S/N at the input to the receiver, but that input S/N must be higher than a threshold value, which depends on the deviation factor of the modulation.

In contrast, spread spectrum has no advantage over baseband transmission from the point of view of signal-to-noise ratio. Usual comparisons of modulation methods are based on a channel having only additive wideband Gaussian noise. With such a basis for comparison, there would be no advantage at all in using spread spectrum compared to sending the same data over a narrow-band link. The advantages of spread spectrum are related to its relative immunity to interfering signals, to the difficulty of message interception by a chance

eavesdropper, and to its ability to use code selective signal differentiation. Another often-stated advantage to spread spectrum—reduction of multipath interference—is not particularly relevant to short-range communication because the pulse widths involved are much longer than the delay times encountered indoors. (A spread-spectrum specialist company, Digital Wireless, claims a method of countering multipath interference over short distances.)

The basic difference between WBFM and spread spectrum is that the spreading process of the latter is completely independent of the baseband signal itself. The transmitted signal is spread by one (or more) of several different spreading methods, and then unspread in a receiver that knows the spreading code of the transmitter.

The methods for spreading the bandwidth of the spread-spectrum transmission are frequency-hopping spread spectrum (FHSS), direct-sequence spread spectrum (DSSS), pulsed-frequency modulation or chirp modulation, and time-hopping spread spectrum. The last two types are not allowed in the FCC rules for unlicensed operation and after giving a brief definition of them we will not consider them further.

1. In frequency-hopping spread spectrum, the RF carrier frequency is changed relatively rapidly at a rate of the same order of magnitude as the bandwidth of the source information (analog or digital), but not dependent on it in any way. At least several tens of different frequencies are used, and they are changed according to a pseudo-random pattern known also at the receiver. The spectrum bandwidth is roughly the number of the different carrier frequencies times the bandwidth occupied by the modulation information on one hopping frequency.

2. The direct-sequence spread-spectrum signal is modulated by a pseudo-random digital code sequence known to the receiver. The bit rate of this code is much higher than the bit rate of the information data, so the bandwidth of the RF signal is consequently higher than the bandwidth of the data.

3. In chirp modulation, the transmitted frequency is swept for a given duration from one value to another. The receiver knows the starting frequency and duration so it can unspread the signal.

4. A time-hopping spread-spectrum transmitter sends low duty cycle pulses with pseudo-random intervals between them. The receiver unspreads the signal by gating its reception path according to the same random code as used in the transmitter.

Actually, all of the above methods, and their combinations that are sometimes employed, are similar in that a pseudo-random or arbitrary (in the case of chirp) modulation process used in the transmitter is duplicated in reverse in the receiver to unravel the wide-band transmission and bring it to a form where the desired signal can be demodulated like any narrowband transmission.

The performance of all types of spread-spectrum signals is strongly related to a property called process gain. It is this process gain that quantifies the degree of selection of the desired signal over interfering narrow-band and other wide-band signals in the same passband. Process gain is the difference in dB between the output *S*/*N* after unspreading and the input *S*/*N* to the receiver:

$$PG_{\text{dB}} = (S/N)_{\text{out}} - (S/N)_{\text{in}} \tag{6.3}$$

The process gain factor may be approximated by the ratio:

PG_f = (RF bandwidth)/(rate of information)

A possibly more useful indication of the effectiveness of a spread-spectrum system is the jamming margin:

$$\text{Jamming Margin} = PG - (L_{\text{sys}} + (S/N)_{\text{out}}) \tag{6.4}$$

where L_{sys} is system implementation losses, which may be of the order of 2 dB.

The jamming margin is the amount by which a potentially interfering signal in the receiver's passband may be stronger than the desired signal without impairing the desired signal's ability to get through.

Let's now look at the details of frequency hopping and direct-sequence spread spectrum.

6.4.6.1 Frequency Hopping

FHSS can be divided into two classes—fast hopping and slow hopping. A fast-hopping transmission changes frequency one or more times per data bit. In slow hopping, several bits are sent per hopping frequency. Slow hopping is used for fast data rates since the frequency synthesizers in the transmitter and receiver are not able to switch and settle to new frequencies fast enough to keep up with the data rate if one or fewer bits per hop are transmitted. The spectrum of a frequency-hopping signal looks like Figure 6.19.

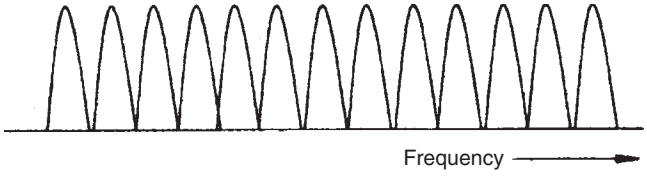


Figure 6.19: Spectrum of Frequency-Hopping Spread Spectrum Signal

Figure 6.20 is a block diagram of FHSS transmitter and receiver. Both transmitter and receiver local oscillator frequencies are controlled by frequency synthesizers. The receiver must detect the beginning of a transmission and synchronize its synthesizer to that of the

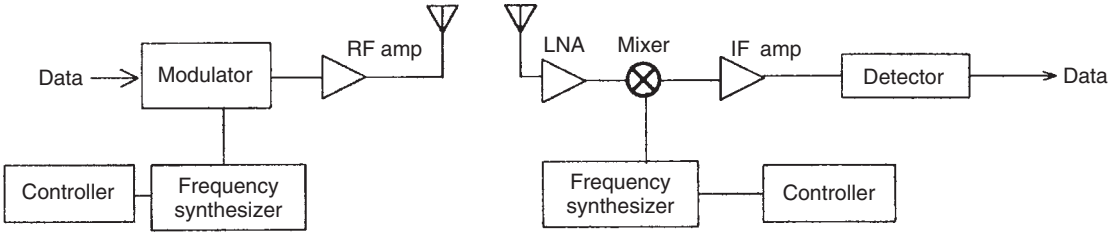


Figure 6.20: FHSS Transmitter and Receiver

transmitter. When the receiver knows the pseudo-random pattern of the transmitter, it can lock onto the incoming signal and must then remain in synchronization by changing frequencies at the same time as the transmitter. Once exact synchronization has been obtained, the IF frequency will be constant and the signal can be demodulated just as in a normal narrow-band superheterodyne receiver. If one or more of the frequencies that the transmission occupies momentarily is also occupied by an interfering signal, the bit or bits that were transmitted at that time may be lost. Thus, the transmitted message must contain redundancy or error correction coding so that the lost bits can be reconstructed.

6.4.6.2 Direct Sequence

Figure 6.21 is a diagram of a DSSS system. A pseudo-random spreading code modulates the transmitter carrier frequency, which is then modulated by the data. The elements of

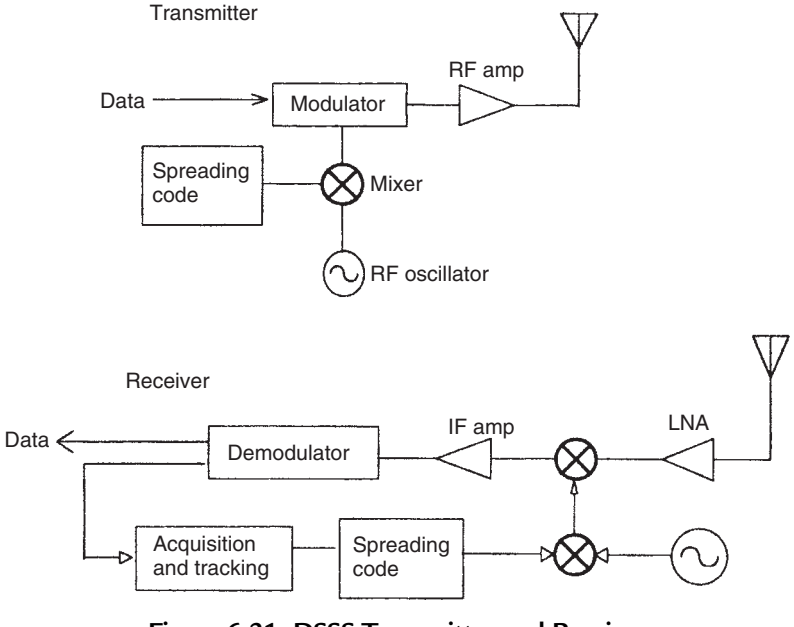


Figure 6.21: DSSS Transmitter and Receiver

this code are called chips. The frequency spectrum created is shown in Figure 6.22. The required bandwidth is the width of the major lobe, shown on the drawing as $2 \times Rc$, or twice the chip rate. Due to the wide bandwidth, the signal-to-noise ratio at the receiver input is very low, often below 0 dB. The receiver multiplies a replica of the transmitter pseudo-random spreading code with the receiver local oscillator and the result is mixed with the incoming signal. When the transmitter and receiver spreading codes are the same and are in phase, a narrow-band IF signal results that can be demodulated in a conventional fashion.

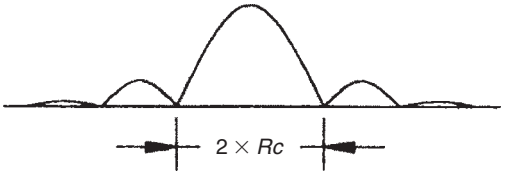


Figure 6.22: DSSS Frequency Spectrum

There are two stages to synchronizing transmitter and receiver pseudorandom codes—acquisition and tracking. In order to acquire a signal, the receiver multiplies the known expected code with the incoming RF signal. The output of this multiplication will be random noise when signals other than the desired signal exist on the channel. When the desired signal is present, it must be synchronized precisely in phase with the receiver's code sequence to achieve a strong output. A common way to obtain this synchronization is for the receiver to adjust its code to a rate slightly different from that of the transmitter. Then the phase difference between the two codes will change steadily until they are within one bit of each other, at which time the output of the multiplier increases to a peak when the codes are perfectly synchronized. This method of acquisition is called a sliding correlator.

The next stage in synchronization, tracking, keeps the transmitted and received code bits aligned for the complete duration of the message. In the method called "tau delta" the receiver code rate is varied slightly around the transmitted rate to produce an error signal that is used to create an average rate exactly equal to the transmitter's.

Other methods for code acquisition and tracking can be found in any book on spread spectrum—for example, Dixon's—listed in the references. Once synchronization has been achieved, the resulting narrow-band IF signal has a *S/N* equal to the received *S/N* plus the process gain, and it can be demodulated as any narrow-band signal. The difference in using spread spectrum is that interfering signals that would render the normal narrowband signal unintelligible are now reduced by the amount of the process gain.

6.4.6.3 Relative Advantages of DSSS and FHSS

The advantages of one method over the other are often debated, but there is no universally agreed-upon conclusion as to which is better, and both types are commonly used. They

both are referenced in the IEEE specification 802.11 for wireless LANs. However some conclusions, possibly debatable, can be made:

- For a given data rate and output power, an FHSS system can be designed with less power consumption than DSSS.
- DSSS can provide higher data rates with less redundancy.
- Acquisition time is generally lower with FHSS (although short times are achieved with DSSS using matched filter and short code sequence).
- FHSS may have higher interference immunity, particularly when compared with DSSS with lowest allowed process gain.
- DSSS has more flexibility—frequency agility may be used to increase interference immunity.

6.5 RFID

A growing class of applications for short-range wireless is radio frequency identification (RFID). A basic difference between RFID and the applications discussed above is that RFID devices are not communication devices *per se* but involve interrogated transponders. Instead of having two separate transmitter and receiver terminals, an RFID system consists of a reader that sends a signal to a passive or active tag and then receives and interprets a modified signal reflected or retransmitted back to it.

The reader has a conventional transmitter and receiver having similar characteristics to the devices already discussed. The tag itself is special for this class of applications. It may be passive, receiving its power for retransmission from the signal sent from the reader, or it may have an active receiver and transmitter and tiny embedded long-life battery. Ranges of RFID may be several centimeters up to tens of meter. Operation frequencies range from 125 kHz up to 2.45 GHz. Frequencies are usually those specified for nonlicensed applications. Higher data rates demand higher frequencies.

In its most common form of operation, an RFID system works as follows. The reader transmits an interrogation signal, which is received by the tag. The tag may alter the incoming signal in some unique manner and reflect it back, or its transmitting circuit may be triggered to read its ID code residing in memory and to transmit this code back to the receiver. Active tags have greater range than passive tags. If the tag is moving in relation to the receiver, such as in the case of toll collection on a highway, the data transfer must take place fast enough to be complete before the tag is out of range.

These are some design issues for RFID:

- Tag orientation is likely to be random, so tag and reader antennas must be designed to give the required range for any orientation.

- Multiple tags within the transmission range can cause return message collisions. One way to avoid this is by giving the tags random delay times for response, which may allow several tags to be interrogated at once.
- Tags must be elaborately coded to prevent misreading.

6.6 Summary

This chapter has examined various characteristics of short-range systems. It started with the ways in which data is formatted into different information fields for transmission over a wireless link. We looked at several methods of encoding the one's and zero's of the baseband information before modulation in order to meet certain performance requirements in the receiver, such as constant DC level and minimum bit error probability. Analog systems were also mentioned, and we saw that pre-emphasis/deemphasis and compression/expansion circuits in voice communication devices improve the signal-to-noise ratio and increase the dynamic range.

Reasons for preferring frequency or amplitude digital modulation were presented from the points of view of equipment complexity and of the different regulatory requirements in the US and in Europe. Similarly, there are several considerations in choosing a frequency band for a wireless system, among them background noise, antenna size, and cost.

The three basic modulation types involve impressing the baseband data on the amplitude, frequency, or phase of an RF carrier signal. Modern digital communication uses combinations and variations in the basic methods to achieve high bandwidth efficiency, or conversely, high signal-tonoise ratio with relatively low power. We gave an introduction to quadrature modulation and to the principles of spread-spectrum communication. The importance of advanced modulation methods is on the rise, and they can surely be expected to have an increasing influence on short-range radio design in the near future.

References

[6.1] Anthes, John, "OOK, ASK, and FSK Modulation in the Presence of an Interfering Signal," Application Note, RF Monolithics, Dallas, Texas.

[6.2] Dixon, Robert C., *Spread Spectrum Systems*, John Wiley & Sons, New York, 1984.

[6.3] Vear, Tim, "Selection and Operation of Wireless Microphone Systems," Shure Brothers Inc. 1998.

CHAPTER 7

High-Speed Wireless Data: System Types, Standards-Based and Proprietary Solutions

Ron Olexa

Wireless data networks are often divided into several categories according to how the networks are viewed by the user. Such characteristics as fixed or mobile, point-to-point (PTP) or point-to-multipoint (PTM), licensed or unlicensed, and standards-based or proprietary are used to define the network. In reality, there are only two distinct types of networks: fixed or mobile. For purposes of definition fixed networks include networks that connect two or more stationary locations as well as systems like 802.11-based networks designed to support "nomadic" users. The nomadic user is nominally a fixed user constrained by the bounds of coverage available on the network. In a truly mobile system, the service will be ubiquitously available, and support use while the user is in motion. The first systems to offer true broadband mobile data are still years away at the time this book is being written. By adding EDGE, GPRS, 1XRTT, and 1XEVDO overlays to their voice networks, cellular and PCS carriers have taken the first tenuous steps in the direction of providing true mobile data, but the speeds at which current networks function cannot yet be called broadband. That designator can be used when the average connection speed per user exceeds 2 Mbps.

The systems discussed in this book will deliver true broadband connectivity. Available equipment can support speeds in excess of 500 Mbps. The equipment utilized in a network will be impacted by the type of network being implemented as well as the costs and service expectations of the network. While more complex networks require attention to more variables, RF design tools and knowledge requirements are fairly common for all networks regardless of their type.

7.1 Fixed Networks

The simplest network is the fixed point-to-point network. As the name implies, these are facilities that connect two or more fixed locations such as buildings. They are designed to extend data communications to locations physically separate from the rest of the network. A fixed network solution could be used to connect buildings together, to provide a network connection to a home, or to connect multiple network elements together.

These links may be familiar as the traditional microwave link. They use highly directional antennas in order to achieve range and control interference. Depending on the technology selected and the frequency of operation, these links can be designed to span distances as short as several hundred feet or as long as 20 or more miles, with capacities of under 1 Mbps to nearly 1 Gbps.

These systems are designed and engineered as individual radio paths, each path connecting two points together. A network of many of these individual paths could be designed to connect a multitude of disparate locations. For example, a fixed point-to-point network constructed out of a number of unique point-to-point links could be used to extend high-speed connectivity from a central point to a number of buildings in a campus or office park. It could also be used to extend the high speed connectivity of a fiber optic-based network to buildings surrounding the fiber route, thus avoiding the cost and complexity of digging up the streets to extend lateral connections from the fiber route into those other buildings.

Another variant of point-to-point networks are point-to-multipoint networks. In these networks a master or central station no longer uses individual antennas, each focused on a single station. Instead, it uses a wide aperture antenna that is capable of serving many stations in its field of view. In this way, a single system and antenna can share its capacity with a number of users. The benefit of such a system is that a single antenna can serve multiple locations, thus eliminating the need for many individual dish antennas to be located on the roof or tower that serves as the central location. The downside of such a network is threefold. Because the central station uses antennas that cover a wider area, they have a lower gain. This reduces the distance these networks can communicate as compared to a point-to-point network. Secondly, since many users share network capacity it may not be the optimal solution for supporting multiple very high bandwidth users. As with any system, the peak capacity requirements of the users must be considered as part of the overall network design. In the case of point-to-multipoint networks, the peak usage characteristics of multiple users must be considered. The third downside is related to interference management and frequency reuse. Since the central site transmits over a wide area, the ability to reuse the same frequencies in the network becomes more limited.

Point-to-point and point-to-multipoint networks can be accomplished using the licensed or unlicensed bands that exist in frequency ranges from under 1 GHz to over 90 GHz. They can use a multitude of proprietary technologies, or can be accomplished using equipment built to standards such as 802.11 or 802.16. Your selection of operating frequency and technology will be governed by factors such as range, capacity, spectrum availability, link quality, and cost.

7.2 Nomadic Networks

Another variation of point-to-multipoint networks is the network that directly supports a user's connection to the system. By this I mean instead of connecting buildings together, these

nomadic networks connect individual computer users to the network. In the case of a laptop computer or PDA, these computing devices are somewhat mobile, and the network is designed to offer a low level of mobility to these users.

802.11b is a common standard for this type of network, although 802.11g and 802.11a also support this type of use. In order to be truly portable the RF device in the computer must be small, low powered, and the antennas used at the computer must be small and have an omnidirectional pattern. In addition, the user may be shielded from the base station by walls or other objects that attenuate the signal. This leads to a significant reduction in the area that can be effectively covered by one of these networks. Where point-to-point network range could be measured in miles, a nomadic implementation of the same technology has ranges measured in tens to hundreds of yards.

Nomadic networks are becoming quite commonplace. An 802.11b network offering Internet access in a coffee shop is one example of this type of network. Wireless office LANs, and WISP networks covering campuses or Multiple Dwelling Units (MDUs), like apartments, can also be considered nomadic networks.

These networks are the first step being taken to provide individuals with high-speed data access in many public and private venues.

These networks are not true mobile networks. While they can provide some mobility, they do not cover large areas and they do not support the high velocity mobility that would be needed to support a user in a vehicle. As with everything there are trade-offs. Localized low mobility solutions are fairly easy and inexpensive to implement. Better yet, there is unlicensed spectrum available to use for building this type of network, and a large installed base of customer equipment built to operate on the 802.11b Wi-Fi standard already exists. These factors have led to the rapid development of all sorts of nomadic networks, some as small as a home; others as large as a community.

7.3 Mobile Networks

The most complex network is one designed for true mobility. Like a voice-based cellular or PCS network, the high-speed mobile data network must provide ubiquitous coverage, and must support high velocity mobility. These requirements are not easily achieved or inexpensive. These systems will require many tens of megahertz of licensed spectrum, and will require technology that can deal with the hostile RF environment found in a truly mobile application. The 802.16e, 802.20 and CDMA2000 standards are several of the standards that may eventually bring true broadband mobile data solutions to large areas of the earth. Because of their cost, complexity, and need for interference managed dedicated spectrum, large telecom carriers, as opposed to the small businesses that offer nomadic network solutions, will be the most likely owner of these networks.

7.4 Standards-Based Solutions and Proprietary Solutions

The IEEE has a number of working groups responsible for developing open standards. These open standards are available for any manufacturer to use, hopefully ensuring competition and volume production. The IEEE has developed the 802.11x and 802.16 standards, and as of July 2003 has a working group developing the 802.20 standard.

Each of these standards is designed with a certain utility and limitations in mind. For example 802.11b was designed as a short-range wireless Ethernet replacement. While it can be used for other applications (such as community networks) it is not optimized for this type of service, and will never perform as well as a technology that was designed from the ground up to address the unique issues found in a community network.

Certain manufacturers develop equipment that is not designed to any current IEEE or other standard. These solutions sometimes become popular enough that they become a de facto standard. More often, these proprietary standards become niche market solutions, which are only available from a single source. These proprietary solutions may be technically best suited for certain applications, but often are significantly more expensive than standards-based options. In the end, it's up to you to determine whether the improved performance is worth the additional cost and single-vendor supply risks.

There are many proprietary solutions available; unfortunately in a competitive market manufacturers are not willing to release much detail about their equipment operation and performance without the recipient signing a Nondisclosure Agreement (NDA) which limits the amount of information that can be shared or published. Because of this limitation I will not be spending much time discussing proprietary solutions in detail.

7.5 Overview of the IEEE 802.11 Standard

Like many standards, 802.11 has gone through many iterations and expansions over the years. Initially encompassing a 1 Mbps throughput on a 900 MHz channel, it now supports up to 54 Mbps in the 2400 MHz and 5600 MHz bands.

802.11x, also sometimes known as Wi-Fi, is an IEEE certified wireless networking standard that currently includes the IEEE 802.11a, 802.11b and 802.11g specifications. In the U.S., the RF emission of these devices is governed by FCC Part 15 rules. These rules govern the power output, equipment and antenna configurations useable in the unlicensed bands. A copy of the FCC Part 15 rules is included on the CD-ROM that accompanies this book. The 802.11b spec allows for the wireless transmission of approximately 11 Mbps of raw data at indoor distances from several dozen to several hundred feet and outdoor distances of several to tens of miles as an unlicensed use of the 2.4 GHz band. The 802.11a spec uses the unlicensed 5 GHz band, and can handle 54 Mbps over shorter distances.

The 802.11g standard applies the 802.11a modulation standards (and therefore supports 54 Mbps just like 802.11a) to the 2.4 GHz band, and offers "backward compatibility" for 802.11b devices. The achievable coverage distances for these standards depend on impediments and obstacles to line of sight.

The 802.11b specification started to appear in consumer form in mid-1999, with Apple Computer's introduction of its AirPort components, manufactured in conjunction with Lucent's WaveLAN division. (The division changed its named to Orinoco and was spun off to the newly formed Agere Corporation with a variety of other Lucent assets in early 2001; these assets were resold to Proxim Corporation in June 2002, although Agere continues to make chips.)

802.11x is an extension of wired Ethernet, bringing Ethernet-like principles to wireless communication. As such, 802.11 is agnostic about the kinds of data that pass over it. It's primarily used for TCP/IP, but can also handle other forms of networking traffic, such as AppleTalk or NetBEUI.

Computers and other devices using Windows or Mac OS operating systems, and many flavors of Unix and Linux, may communicate over Wi-Fi, using equipment from a variety of vendors. The client hardware is typically a PC card or a PCI card, although USB and other forms of Wi-Fi radios are also available. Adapters for PDAs, such as Palm OS and PocketPC based devices, are available in various forms, and smaller ones that fit into internal Secure Digital and Compact Flash card slots started appearing in late 2002.

Each radio may act, depending on software, as a hub or as part of an ad hoc computer-to-computer transmission network; however it's much more common that a Wireless Local Area Network (WLAN) installation uses one or more Access Points (AP), which are dedicated stand-alone hardware with typically more powerful chipsets and higher gain antennas. Home and small-office APs often include routing, a DHCP server, NAT, and other features required to implement a simple network; enterprise access points include access control features as well as secure authentication support.

The 802.11b standard as implemented in the 2.4 GHz band is backwards compatible with early 2.4 GHz 802.11 equipment. 802.11b can support speeds of 1, 2, 5.5 and 11 Mbps on the same hardware. Multiple 802.11b access points can operate in the same overlapping area over different channels, which are subdivisions of the 2.4 GHz band available.

Internationally, there are 14 standard channels, which are spaced at 5 MHz intervals, from 2.4000 to 2.487 GHz. Only channels 1 through 11 are legal in the U.S.A. The 802.11 channel is 22 MHz wide, so it occupies multiple 5 MHz channels (see Figure 7.1). Only channels 1, 6, and 11 can be assigned to an 802.11 network with no overlap among them. If closer spaced channels are assigned, there will be inter-carrier interference generated. Such overlapping systems can still work, but the interchannel interference will effectively raise the noise floor in the channel, which will have a negative impact on the throughput and range of the systems.

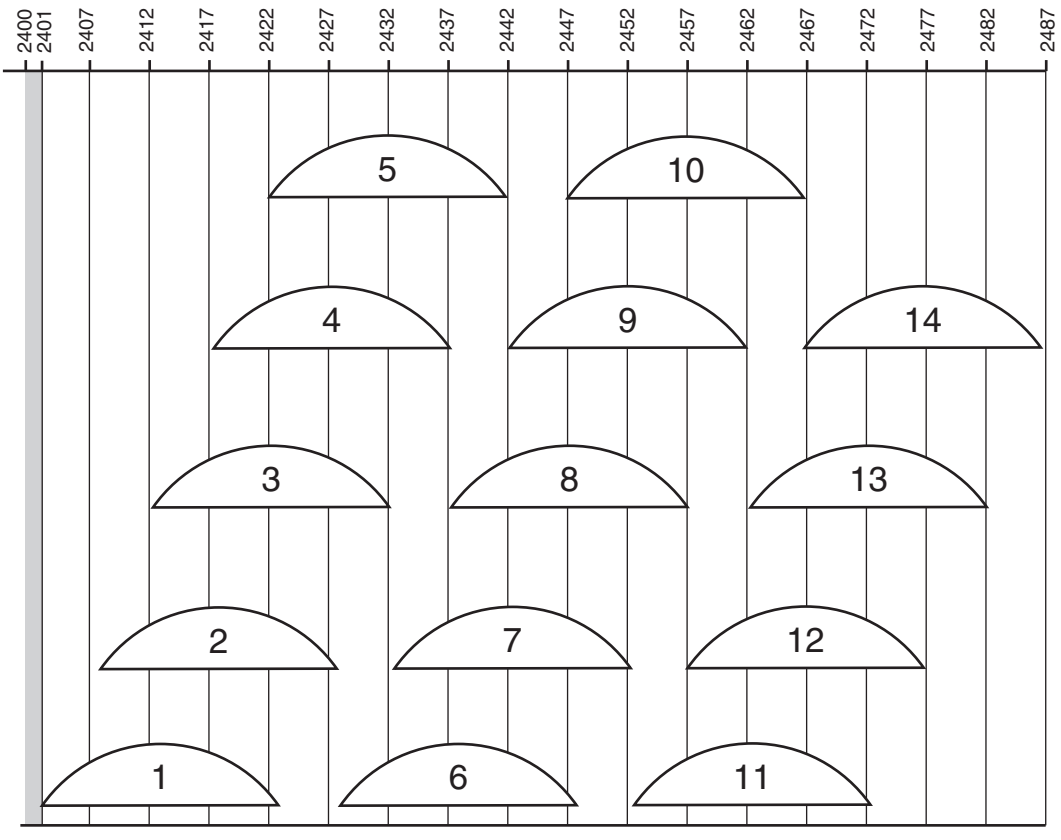


Figure 7.1: 802.11b Frequency vs. Channel Allocation

802.11b uses several types of modulation. Barker Code Direct Sequence Spread Spectrum with BPSK or QPSK modulation is used to transmit at 1 and 2 Mbps respectively, while Complimentary Code Keying is used to support speeds of 5.5 and 11 Mbps. Multiple users are supported by the use of Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA). A new higher speed standard called 802.11g features complete backwards compatibility with 802.11b, but it offers three additional encoding options (one mandatory, two optional) that boost its speed to 54 Mbps, although two 22 Mbps "flavors" are part of the specification as well. The higher speed connections use the same modulation as 802.11a: Orthogonal Frequency Division Multiplexing (OFDM). Future speed improvements achieved through the use of more efficient modulations are expected in 802.11 products operating in both the 2.4 and 5 GHz bands.

802.11a specifies the use of OFDM modulation only, and supports data rates of 6, 9, 12, 18, 24, 36, 48, or 54 Mbps of which 6, 12, and 24 Mbps are mandatory for all products. OFDM operates extremely efficiently, thus leading to the higher data rates. OFDM divides the data signal across 48 separate sub-carriers to provide transmissions. Each of the sub-carriers uses

phase shift keying (PSK) or Quadrature Amplitude Modulation (QAM) to modulate the digital signal depending on the selected data rate of transmission. In addition, four pilot sub-carriers provide a reference to minimize frequency and phase shifts of the signal during transmission.

Multiple users are supported by the use of CSMA/CA, so the same limitations inherent in this access methodology for 802.11b will be present in 802.11a as well. The operating frequencies of 802.11a fall into the U-NII bands: 5.15–5.25 GHz, 5.25–5.35 GHz, and 5.725–5.825 GHz. As shown in Figure 7.2, within this spectrum there are twelve 20-MHz channels (eight allowable only for indoor use and four useable for indoor or outdoor use) that do not overlap, thus allowing denser installations. Additionally, each band has different output power limits that are detailed in the FCC rules Part 15.407. 802.11a's range is less due to both its frequency of operation and more complex modulation, but in a closed environment like an office or a home it can often transmit higher speeds at similar distances as compared to 802.11b.

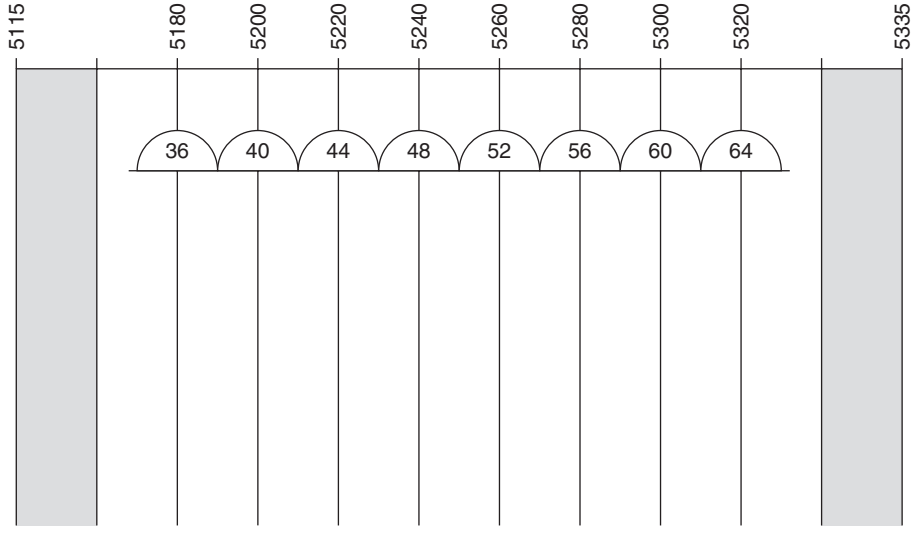


Figure 7.2a: 5.2 GHz 802.11a Channels

As 802.11 systems have proliferated, a number of issues surrounding its limitations have been raised. The security provisions of 802.11 are notoriously weak, and it does not inherently support Quality of Service packet prioritization. New working groups at the IEEE are addressing these limitations. The 802.11e, h, and i standards will improve the capabilities of 802.11, and make it more robust and more useful in a number of situations.

Even though standards exist, they do not guarantee that equipment from different manufacturers will interoperate. To assure interoperability, and thereby assist adoption by the consumer,

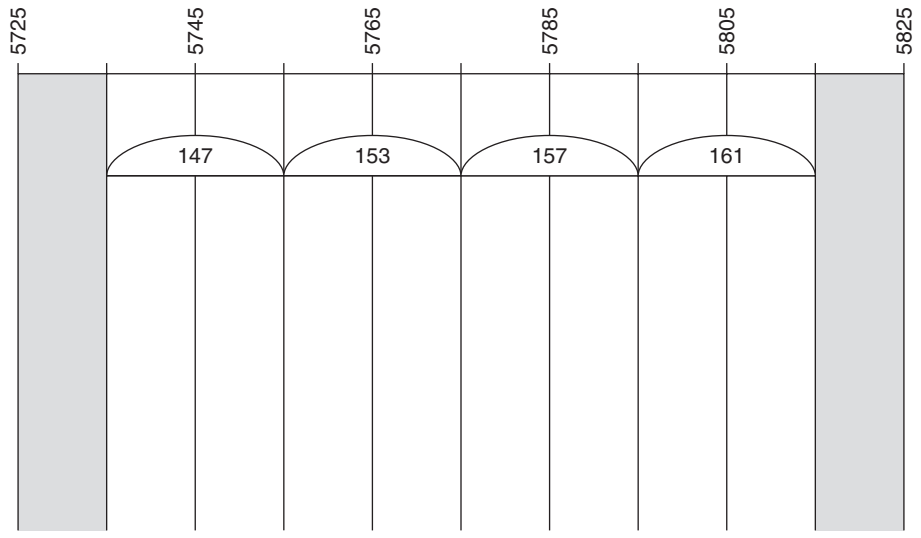


Figure 7.2b: 5.7 GHz 802.11a Channel Allocations

an industry group known as the Wi-Fi Alliance (formerly known as the Wireless Ethernet Compatibility Alliance) certifies its members equipment as conforming to the 802.11a and b standards, and allows compliant hardware to be stamped Wi-Fi compatible. The Wi-Fi seal of approval is an attempt at a guarantee of compatibility between hundreds of vendors and thousands of devices. (The IEEE does not have such a mechanism, as it only promulgates standards.) In early October 2002, the group modified the Wi-Fi mark to indicate both a and b standards by noting 2.4- or 5-GHz band compatibility.

802.11b was the first standard deployed for public short-range networks, such as those found at airports, hotels, conference centers, and coffee shops and restaurants. Several companies currently offer paid hourly, session-based, or unlimited monthly access via their deployed networks around the U.S. and internationally.

802.11a and b are a great way to extend a data or Internet connection to a site that does not have one through point-to-point operation, or to build a point-to-multipoint system which could provide a high speed data connection shared by a number of fixed and nomadic users.

As delivered, 802.11 products conform to FCC Part 15 rules, which limit both the device RF power and EIRP achieved by use of a gain antenna. The most stringent restrictions are placed on omnidirectional operations, since those operations result in the highest overall interference contribution to the surrounding area. In the case of omni operation, the EIRP is limited to 1 watt. If a directional antenna is used, the allowable EIRP jumps to 4 watts. If a fixed point-to-point link is implemented, even higher EIRP is available, in fact with 30 dBi

gain antennas, over 100 watts EIRP can be achieved. EIRP of this magnitude will support a point-to-point link over 15 miles long, given the right conditions in the path.

802.11 does have a significant impediment when used in a WISP or MAN type deployment: like the Ethernet standard upon which it is based, it uses Carrier Sense Multiple Access as its access protocol. In the case of 802.11, the full specification is Carrier Sense Multiple Access/ Collision Avoidance (CSMA/CA), which anticipates that all stations are able to hear each other, and thus to have the ability to listen for activity on the channel prior to transmitting. If another carrier is heard, the station knows the channel is in use and "backs off" for a random time. At the end of the back off interval, the station listens again, and transmits if the channel is clear.

In a widespread outdoor system the end users can hear the base station or AP, but they may not be able to hear each other, so CSMA/CA does not work. As the system gets loaded it is possible for multiple users all to transmit at once, leading to interference and packet loss. Another effect is the "near far" issue in which users closer to the base station get better access to the bandwidth than stations further away. This is due to the higher signal strength of the close-in user swamping the weaker signal of the remote user. Enabling Request to Send/Clear to Send (RTS/CTS) can diminish these effects. With RTS/CTS enabled, the stations ask permission before transmitting, and wait to get an all clear before they actually start their transmission. Still, there is no way for a station to know whether a CTS was in response to his RTS or another sent by another station, so collisions can still happen. Additionally, RTS/CTS adds additional overhead, thus reducing the bandwidth available to carry real traffic.

Several manufacturers have addressed these issues by adding some semblance of central control to the systems. By queuing the users and specifying who gets to transmit next, there is an improvement in usability, quality and throughput for all users. Of course, such equipment uses proprietary standards, and some even require the radio hardware to be their proprietary solution.

802.11 was never designed for WISP or Metropolitan Area Network (MAN) deployment. It has become a de facto standard for such use because it is inexpensive, has freely available spectrum, and works well enough on lightly loaded systems. The IEEE recognized the limitations of 802.11 and the need for a MAN solution. The 802.16 standard was promulgated in order to provide a solution for the WISP and metropolitan area network provider. Because 802.16 was designed to cover large areas, its MAC layer does not use CSMA/CA, so it does not exhibit 802.11's CSMA/CA limitations.

7.6 Overview of the IEEE 802.16 Standard

The IEEE 802.16 Air Interface Standard is a state-of-the-art specification for fixed broadband wireless access systems employing a Point-to-Multipoint (PMP) architecture. The initial

version was developed with the goal of meeting the requirements of a large number of deployment scenarios for Broadband Wireless Access (BWA) systems operating between 10 and 66 GHz. As a result, only a subset of the functionality is needed for typical deployments directed at specific markets.

An amendment to add support for systems operating between 2 and 11 GHz was added to the initial specification. Since the IEEE process stops short of providing conformance statements and test specifications, in order to ensure interoperability between vendors the WiMAX forum was created. In much the same manner as the Wi-Fi forum assured equipment interoperability to the 802.11 standard, the WiMAX forum will provide the testing and certification necessary to assure vendor equipment interoperability for 802.16 hardware.

Task Group 1 of the IEEE 802.16 committee developed a point-to-multipoint broadband wireless access standard for systems in the frequency range 10–66 GHz. The standard covers both the Media Access Control (MAC) and the PHYsical (PHY) layers. Task groups a and b jointly produced an amendment to extend the specification to cover both the licensed and unlicensed bands in the 2–11 GHz range.

A number of PHY considerations were taken into account for the target environment. For example at frequencies above 6 GHz, line of sight (LOS) paths between stations are a must. By taking the need for line of sight paths as a design requirement the PHY can be designed for minimal effects related to multipath. This allows the PHY to accomodate wide channels, typically greater than 10 MHz in bandwidth, thus giving IEEE 802.16 the ability to provide very high capacity links on both the uplink and the downlink.

At the lower frequencies, line of sight is not required for link operation, although the lack of line of sight (LOS) forces other design trade-offs. Adaptive burst profiles (changing both modulation and forward-error correction (FEC)) are used to further increase the typical capacity of 802.16 systems with respect to older technology. The MAC was designed to accommodate different PHYs for the different environments. The single-carrier PHYs are designed to accommodate either Time Division Duplexing (TDD) or Frequency Division Duplexing (FDD) deployments, allowing for both full and half-duplex terminals in the FDD case.

The MAC was designed specifically for the PMP wireless access environment. It is designed to carry any higher layer or transport protocol such as Asynchronous Transfer Mode (ATM), Ethernet or Internet Protocol (IP) seamlessly, and is designed to accommodate future protocols that have not yet been developed. The MAC is designed for the very high bit rates (up to 268 Mbps each way) of the truly broadband physical layer, while delivering ATM compatible Quality of Service (QoS) to ATM as well as non-ATM (MPLS, VoIP, and so forth) services.

The frame structure allows terminals to be assigned uplink and downlink burst profiles dynamically according to their link conditions. This allows a trade-off between capacity and robustness in real-time, and provides an approximate twofold increase in capacity

on average when compared to nonadaptive systems, while maintaining appropriate link availability.

The 802.16 MAC uses a variable length Protocol Data Unit (PDU) along with a number of other concepts that greatly increase the efficiency of the standard. Multiple MAC PDUs may be concatenated into a single burst to save PHY overhead. Additionally, multiple Service Data Units (SDU) for the same service may be concatenated into a single MAC PDU, thus saving on MAC header overhead. Variable fragmentation thresholds allow very large SDUs to be sent piece meal to guarantee the QoS of competing services. Additionally, payload header suppression can be used to reduce the overhead caused by the redundant portions of SDU headers.

The MAC uses a self-correcting bandwidth request/grant algorithm known as Demand Assigned Multiple Access/Time Division Multiple Access (DAMA/TDMA) that eliminates the shortcomings of the CSMA/CA technique. DAMA adapts as needed to respond to demand changes among multiple stations. With DAMA, the assignment of timeslots to channels varies dynamically based upon need. For transmission from a base station to subscribers, the standard specifies two modes of operation, one targeted to support a continuous transmission stream (mode A), such as audio or video, and one targeted to support a burst transmission stream (mode B), such as IP-based traffic. User terminals have a variety of options available to them for requesting bandwidth depending upon the QoS and traffic parameters of their services. Users can be polled individually or in groups. They can signal the need to be polled, and they can piggyback requests for bandwidth.

7.7 10–66 GHz Technical Standards

In the same manner as the Wi-Fi consortium managed compatability for 802.11 devices, the WiMAX forum is working with 802.16 products and standards to assure broad compatability. Since the 10–66 GHz standard was the first to be released, WiMAX initially created a 10–66 GHz technical working group. The technical working group created equipment operating profiles and test specifications, but an authorized, independent laboratory does actual testing. For each system profile, functions are separated between mandatory and optional feature classes. There can be differences from one equipment manufacturer to another in implementing optional features, but mandatory features will be the same in every vendor's product.

WiMAX is currently defining two MAC system profiles, one for basic ATM and the other for IP-based systems. Two primary PHY system profiles are also being defined: a 25 MHz-wide channel (typically for U.S. deployments) for use in the 10–66 GHz range and a 28 MHz wide channel (typically for European deployments) also for use in the 10–66 GHz range. The PHY profiles are identical except for their channel width and their symbol rate, which is proportional to their channel width. Each primary PHY profile has two duplexing scheme sub-profiles one for Frequency Division Duplex (FDD) and another for Time Division Duplex

(TDD). Additionally, because these systems were designed for operating over LOS paths, traditional multistate QAM modulation is used.

7.8 2–11 GHz Standards

In early 2003, the IEEE 802.16 standard was expanded with the adoption of the 802.16a amendment, focused on broadband wireless access in the frequencies from 2 to 11 GHz. Given the charter of the WiMAX forum, to promote certification and interoperability for microwave access around the globe, WiMAX has expanded its scope to include the 802.16a standard.

The 802.16a standard is designed to operate over both LOS and NLOS paths. Because of the multipath effects present in NLOS paths, QAM as used in the 10–66 GHz 802.16 variant, was not a suitable modulation. 802.16a instead uses OFDM as its modulation technique.

The WiMAX 2–11 GHz working group is currently defining MAC and PHY System profiles for IEEE 802.16a and HiperMAN standards. The MAC profiles that are being developed include IP-based versions for deployment in both licensed and unlicensed spectrum.

While the IEEE 802.16a amendment has several physical layer profiles, the WiMAX forum is focusing on the 256-point FFT OFDM PHY mode as its initial and primary interoperability mode. Various channel sizes that cover typical spectrum allocations in both licensed and license exempt bands around the globe have been chosen. All selected channel sizes support the 256-point FFT OFDM PHY mode of operation.

In February 2003, the IEEE instituted another working group, the 802.16e working group. The 802.16e extension adds vehicular speed mobility in the 2 to 6 GHz licensed bands. At the time this book is being written, this extension is still in committee. It is anticipated that the standard will be released in mid 2004.

7.9 Overview of the IEEE 802.20 Standard

The 802.20 standard focuses on true high velocity mobile broadband systems. The 802.20 interface seeks to boost real-time data transmission rates in wireless metropolitan area networks to speeds that rival DSL and cable modem connections (1 Mbps or more). This will be accomplished with base stations covering radii of up to 15 kilometers or more, and it plans to deliver those rates to mobile users even when they are traveling at speeds up to 250 kilometers per hour (155 miles per hour). This would make 802.20 an option for deployment in high-speed trains. The standard is focused on operation in licensed bands below 3.5 GHz.

The 802.20 Working Group was actually established before the IEEE gave the go-ahead to 802.16e. The IEEE originally intended to have the 802.20 standard in place by the end of 2004, but the group has been mired in conflict and has made little progress to date.

802.20 may become a direct competitor to third-generation (3G) wireless cellular technologies such as CDMA2000 and GMRS. Instead of using TDMA or CDMA technology, 802.20 is expected to use OFDM as its modulation technique.

7.10 Proprietary Solutions

In addition to the standards-based solutions, there are numerous vendor proprietary systems available. Proprietary solutions are normally designed to best suit a particular deployment scenario, and may operate in licensed bands, unlicensed bands, or in some cases both.

Just like standard solutions, proprietary solutions continue to evolve in order to keep a competitive edge and to better meet the needs of a growing business opportunity. Because of the financial dynamics associated with companies providing proprietary solutions as well as the changing requirements of the marketplace, there is no guarantee that the equipment or manufacturers discussed next will still be available by the time you read this. Table 7.1 lists some of the proprietary manufacturers and publicly available product specifications.

Because proprietary solutions are just that: proprietary, it is often extremely difficult to obtain specific information about the operation of the hardware without signing a nondisclosure agreement with the vendor. Of course this is only possible if the vendor will commit to such an agreement with you. Lacking these particulars about the equipment can make it difficult to compare operating characteristics of the equipment, and analyze how a particular solution might fulfill your particular requirements. Happily, all is not lost. Generally, information about capacity and throughput is generally publicly available. The missing information usually relates to the actual RF operating characteristics of the hardware.

In order for any radio transmitting equipment to be sold in the U.S., it is required to go through an FCC certification process. This certification is accomplished by an independent testing lab, which conducts tests and measurements on the equipment to assure that it meets the FCC's technical requirements for the band in which it operates. The result of passing this certification process is that the FCC grants an authorization number, which is used by the manufacturer to show that the equipment is legally operating within the FCC rules, and that it can legally be sold for operation in the U.S.

The FCC publishes the results of these tests as public record. They can be found at the FCC Equipment Authorization System Generic Search web page. As of February, 2004 this page is located at: https://gullfoss2.fcc.gov/prod/oet/cf/eas/reports/GenericSearch.cfm. Knowing as little as the manufacturer's name and the band of operation will allow you to use this search engine to identify certified equipment. Once you've identified the particular equipment you're interested in, read the test results and other documentation on file. While some information may be held in confidence, important information like spectrum analyzer plots of the output waveform and the output power will be part of the public record. Knowing power output is

Table 7.1: Some proprietary manufacturers and publicly available product specifications

Company	Product	Operating Band	Total Speed	Maximum number of Sectors	Maximum number of users per Sector	LOS/NLOS	Modulation	Encryption Levels	Output Power	Receiver Sensitivity	Point to Point	Point to Multipoint	Duplex
AIRAYA	A|108 Wireless Bridge	5.25–5.35 GHz	108 Mbps	1	1	LOS	OFDM	152-bit	EIRP 29.6 dB		Yes	No	Half
AIRAYA	A|108 Wireless Bridge	5.25–5.35 GHz	108 Mbps	1	1	LOS	OFDM	152-bit	EIRP 29.6 dB		Yes	No	Half
Alvarion	Breeze-ACCESS	2.4, MMDS, 3.5, 5.15–5.35, 5.4, 5.7 UNII, 5.7 ISM	3 Mbps for GFSK, 12 Mbps for 3.5 GHz	depends on band, up to 12 in 2.4, up to 36 in 3.5 (with multibeam)	1000	LOS and NLOS options	FHSS, DSSS, & OFDM	128-bit, triple DES option	26 dBm, with APC		Yes	Yes	TDD
Aperto	PacketWave	2.5 GHz, 3.5 GHz and 5.8 GHz	20 Mbps per sector	6	1000	LOS/NLOS		Other	Varies depending on band		No	Yes	Half
Aperto	PacketWave Point-to-Point Bridges	5.8 GHz	20 Mbps per sector	1	1	LOS/NLOS		Other	Varies depending on band		Yes	No	Half
Axxcelera	AB-Access	5.7 UNII band	25 Mbps PMP, 12.5 Mbps PTP	12	256	LOS	TDMA/TDD	Yes, proprietary	32 dBm	−86.1 dB for 10−4 BER	Yes	Yes	Full or half
BeamReach	BeamPlex	2.3 GHz		Not Applicable	16,000 per cell	NLOS	Adaptive Multi-Beam OFDM	Yes				Yes	
BeamReach	BeamPlex	2.3 GHz		Not Applicable	16,000 per cell	NLOS	Adaptive Multi-Beam OFDM	Yes				Yes	
Ceragon	FibeAir	6, 7, 8, 11, 13, 15, 18, 23, 26, 28, 28, 31, 32, 38 GHz	622 Mbps	N/A	N/A	LOS		DES	MAX 24 dBm	−73 dBm	Yes	No	Full
Cirronet	WaveBolt	2.4 GHz and 5.8 GHz (5.8 available in Q1 '03)	1 Mbps	5	240	LOS/NLOS	FHSS	Proprietary	+18 dBm	−88 dBm	Yes	Yes	Full

Dragon-wave	AirPair	18, 23, 28 GHz	50–100 Mbps	N/A	N/A	LOS	Single Carrier QAM	REL 3 will include DES encryption	+13/ +17 dBm	−77/ −80.5 dBm	Yes	N/A	Full
Innowave	MGW	0.8; 1.5; 1.9; 2.4; 3.4–3.8 GHz	850 Kbps	6	1000	NLOS	FHCDMA/ TDMA/TDD	Other	27 dBm	−90 dBm	No	Yes	Full
Innowave	eMGW	1.5; 1.9; 2.4; 3.4–3.8; 5.7 GHz	1.5 Mbps	6	2000	NLOS	FHCDMA/ TDMA/TDD	Other	27 dBm	−90 dBm	No	Yes	Full
Innowave	WaveGain	3.5 GHz	15 Mbps	4	128	NLOS	WCDMA/ FDD	Other	37 dBm	−110 dBm	No	Yes	Full
Mesh-Network	MEA IAP6300	2.4 GHz	6 Mbps	N/A	250	NLOS	Multi-Hop DSSS	*VPN, IPSEC	22+ dBm		Yes	Yes	Half
Motorola	Canopy	5.2 GHz	10 Mbps	6	200	LOS		single DES	30 dB	−83 dB	Yes	Yes	Half
Motorola	Canopy	5.7 GHz	10 Mbps	6	200	LOS		single DES	30 dB	−83 dB	Yes	Yes	Half
Navini	Ripwave	2.4 GHz; 2.5/ 2.6 GHz; 2.3 GHz	48 Mbps (3.3 sec-tored cell). 72 Mbps in 2003	3	1000	NLOS, Zero-install plug-and-play	Phased-array smart antennas, Multi-Carrier Synchronous Beamform-ing (MCSB), TDD	Patented CDMA encoding+ spatial isolation and nulling with beam-forming provides for a high level of security and can be overlayed.	Depends on the fre-quency of operation		Yes	Yes	Full
Nokia	Nokia Roof Top	2.4 GHz	12 Mbps aggregate	6	40	NLOS	FHSS	none	12 dBm to 27 dBm	−82 dBm	No	No	Half duplex
P-Com	AirPro Gold. Net	2.4 & 5.8 GHz	11 Mbps	4	127	Both	FHSS	MAC security	+28 dBm −2.4 GHz, +275.8 GHz				
Proxim	Tsunami Multipoint 20 MB Base Station Unit	5.8 GHz	20 Mbps	6 typical max per hub site	1,023	Near Line of Sight		Proprietary	36 dBm		No	Yes	Half

(*Continued*)

Table 7.1: (Continued)

Company	Product	Operating Band	Total Speed	Maximum number of Sectors	Maximum number of users per Sector	LOS/NLOS	Modulation	Encryption Levels	Output Power	Receiver Sensitivity	Point to Point	Point to Multipoint	Duplex
Proxim	Tsunami Multipoint 20 MB Subscriber Unit	5.8 GHz	20 Mbps	6 typical max per hub site	N/A	Near Line of Sight		Proprietary	35 dBm		No	Yes	Half
Proxim	Tsunami Multipoint 60 MB Base Station Unit	5.8 GHz	60 Mbps	6 typical max per hub site	1,023	Near Line of Sight		Proprietary	36 dBm		No	Yes	Half
Proxim	Tsunami Multipoint 60 MB Subscriber Unit	5.8 GHz	60 Mbps	6 typical max per hub site	N/A	Near Line of Sight		Proprietary	35 dBm		No	Yes	Half
Proxim	Quick-Bridge 20	5.8 GHz	18 Mbps	N/A		LOS		16 Char Security ID	+36 dBm EIRP	−89 dBm	Yes	No	Half duplex
Proxim	Quick-Bridge 60	5.8 GHz	54, 36, 18 Mbps aggregate capacity	N/A		LOS		16 Char Security ID	+36 dBm EIRP	−77 dBm	Yes	No	Half duplex
Proxim	Quick-Bridge 20 +T1	5.8 GHz	12 Mbps aggregate capacity	N/A		LOS		16 Char Security ID	+36 dBm	−89 dBm	Yes	No	Half duplex
Proxim	Quick-Bridge 20 +E1	5.8 GHz	12 Mbps aggregate capacity	N/A		LOS		16 Char Security ID	+36 dBm EIRP	−89 dBm	Yes	No	Half duplex
Proxim	Tsunami 10 2.4 GHz Wireless Ethernet Bridge	2.4 GHz	10 Mbps full duplex with way-side T1 channel	N/A		LOS	DSSS	8 bit Security Address	+27 dBm	−86 dBm	Yes	No	

Proxim	Tsunami 10 5.8 GHz Wireless Ethernet Bridge	5.8 GHz	10 Mbps full duplex with wayside T1 channel	N/A		LOS	DSSS	8 bit Security Address	+20 dBm	−84 dBm	Yes	No	
Proxim	Tsunami 45 5.8 GHz Wireless Fast Ethernet Bridge	5.8 GHz	45 Mbps full duplex with way-side T1 channel	N/A		LOS		12 char Security Code	+17 dBm	−79 dBm	Yes	No	
Proxim	Tsunami 45 5.3 GHz Wireless Fast Ethernet Bridge	5.3 GHz	45 Mbps full duplex with way-side T1 channel	N/A		LOS		12 char Security Code	+13 dBm	−79 dBm	Yes	No	
Proxim	Tsunami 100 5.3/ 5.8 GHz Wireless Fast Ethernet Bridge	5.3 and 5.8 GHz	100 Mbps full duplex with way-side T1 channel	N/A		LOS		12 char Security Code	+10 and +17 dBm	−77 dBm	Yes	No	
Proxim	Tsunami 100 5.8 GHz Wireless Fast Ethernet Bridge	5.8 GHz	100 Mbps full duplex with way-side T1 channel	N/A		LOS		12 char Security Code	+16 dBm	−71 dBm	Yes	No	
RadioLAN	Campus BridgeLINK-II	5.775 GHz 5.3 GHz 5.2 GHz	10 Mbps	6	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	
RadioLAN	Campus BridgeLINK-Lite	5.775 GHz	10 Mbps	1	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	

(*Continued*)

Table 7.1: (Continued)

Company	Product	Operating Band	Total Speed	Maximum number of Sectors	Maximum number of users per Sector	LOS/NLOS	Modulation	Encryption Levels	Output Power	Receiver Sensitivity	Point to Point	Point to Multipoint	Duplex
RadioLAN	Campus BridgeLINK-II (RMG-377-EA1)	5.775 GHz 5.3 GHz 5.2 GHz	10 Mbps	1	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	
RadioLAN	Campus BridgeLINK-I (RMG-377-25P) I	5.775 GHz 5.3 GHz 5.2 GHz	10 Mbps	1	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	
RadioLAN	Campus BridgeLINK-I (RMG-377-RW1) I	5.775 GHz 5.3 GHz 5.2 GHz	10 Mbps	1	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	
RadioLAN	Campus BridgeLINK-II (RMG-377-RW2)	5.775 GHz 5.3 GHz 5.2 GHz	10 Mbps	1	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	
RadioLAN	Campus BridgeLINK-II (RMG-377-RW3)	5.775 GHz 5.3 GHz 5.2 GHz	10 Mbps	1	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	
RadioLAN	Campus BridgeLINK-II (RMG-377-S90)	5.775 GHz 5.3 GHz 5.2 GHz	10 Mbps	4	128	LOS	D-PPM	*WEP 128-bit *WEP 256-bit *Other	+17 dBM		Yes	Yes	

Redline Communi-cations	AN 100	3.4000–3.800 GHz	Up to 70 Mbps			NLOS	OFDM		+"23 dBM"	−88 dBm@ 1E-09 BER in a 7 MHz channel	Yes	Yes	TDD
Remec	ExcelAir® 70	3.5 GHz	Up to 300 Mbps	6	200–400	LOS	SCQAM	None	+40 dbi EIRP		No	Yes	Full
Solectek	SkyWay-NET	2.4 GHz	11 Mbps	6	64	LOS	DSSS	Proprietary	26 dBm	−83 dBm	Yes	Yes	Half
Solectek	SkyWay-LINK	2.4 GHz	11 Mbps			LOS	DSSS	Proprietary	26 dBm	−83 dBm	Yes	No	Half
Solectek	SkyMate CPE	2.4 GHz	11 Mbps	6	64	LOS	DSSS	Proprietary	23 dBm	−80 dBm	Yes	Yes	Half
Solectek	AIRLAN Bridge Kit	2.4 GHz	11 Mbps			LOS	DSSS	Proprietary	23 dBm	−80 dBm	Yes	No	Half
Solectek	AIRLAN Bridge 5	5.8 GHz	11 Mbps			LOS	DSSS	Proprietary	23 dBm		Yes	No	Half
Wi-LAN	AWE 120–24 Wireless Ethernet Bridge	2.4 GHz	12 Mbps raw, up to 9 Mbps effective	3	1000	LOS	DSSS	Proprietary up to 20 Byte	20 dBm	−81 dBm	Yes	Yes	Half Duplex
Wi-LAN	AWE 45–24 Wireless Ethernet Bridges	2.4 GHz	4.5 Mbps raw, up to 3.4 Mbps effective	3	250	LOS	DSSS	Proprietary up to 20 Byte	20 dBm	−83 dBm	Yes	Yes	Half Duplex
Wi-LAN	VIP 110–24	2.4 GHz	11 Mbps raw, up to 8 Mbps effective	4	500	NLOS	DSSS	Proprietary up to 20 Byte	0 to +23 dBm	−82 dBm	Yes	Yes	Half Duplex
Wi-LAN	Ultima3 RD (Rapid Deployment)	5.8 GHz	12 Mbps raw, up to 10 Mbps effective	N/A	N/A	LOS	DSSS	Proprietary up to 20 Byte	−10 dBm to +21 dBm	−80 dBm	Yes	No	Half Duplex
Wi-LAN	Ultima3 ER (Extended Range)	5.8 GHz	12 Mbps raw, up to 10 Mbps effective	N/A	N/A	LOS	DSSS	Proprietary up to 20 Byte	−10 dBm to +21 dBm	−80 dBm	Yes	No	Half Duplex

(*Continued*)

Table 7.1: (Continued)

Company	Product	Operating Band	Total Speed	Maximum number of Sectors	Maximum number of users per Sector	LOS/NLOS	Modulation	Encryption Levels	Output Power	Receiver Sensitivity	Point to Point	Point to Multipoint	Duplex
Wi-LAN	Ultima3 AP (Access Point)	5.8 GHz	12 Mbps raw, up to 10 Mbps effective	4	1000	LOS	DSSS	Proprietary up to 20 Byte	−10 dBm to +21 dBm	−80 dBm	No	Yes	Half Duplex
Wi-LAN	Ultima3 CPE (Customer Premises Equipment)	5.8 GHz	12 Mbps raw, up to 10 Mbps effective	N/A	N/A	LOS	DSSS	Proprietary up to 20 Byte	−10 dBm to +21 dBm	−80 dBm	No	Yes	Half Duplex
Wi-LAN	LIBRA Series (Access Point)	3.5 GHz	16 Mbps raw, up to 12 Mbps effective (7 MHz)	6	2047	NLOS	W-OFDM		avg/peak +22/ +32 dBm	−82 dBm/ −80 dBm (3.5/7 MHz)	Yes	Yes	Full Duplex
Wi-LAN	LIBRA Series (CPE)	3.5 GHz	16 Mbps raw, up to 12 Mbps effective (7 MHz)	N/A	N/A	NLOS	W-OFDM		+17/ +27 dBm	−79 dBm	No	Yes	Half Duplex

key to analyzing the coverage you can expect from a particular solution. Receiver sensitivity is the other factor you will need to learn. In many cases this is not a published part of the FCC certification tests, so you will need to derive it from the information available. The waveform is a useful tool for this investigative work. It will offer clues about the modulation in use and the spectral occupancy of the signal. Couple that with information about the capacity or throughput of the solution, and you can begin to understand the operating characteristics of the receiver. An RF hardware design expert could use this information to derive the receiver sensitivity based on the characteristics. Alternately estimation could be made by assuming that the sensitivity will be similar to that of known equipment having similar modulation characteristics.

Now that you have a brief overview of the types of solutions available for use in deploying a wireless data network, it's time to move on to gaining an understanding about how radio works and what issues must be considered in designing a radio-based network.

This page intentionally left blank

CHAPTER 8

Propagation Modeling and Measuring

Ron Olexa

When designing a system or network, it is helpful (if not imperative) to know the coverage area provided by each site. This is important for two reasons: first, you want to assure that the users in the desired coverage area are served with a high quality signal, and second, you need to know how each transmitter adds to the interference levels in surrounding areas.

In order to evaluate the coverage that will be provided by the selected hardware, either propagation prediction models or physical surveys can be used. Propagation modeling is accomplished with a software tool, while physical surveys are accomplished by temporarily installing hardware then measuring the resulting coverage. If you are building few transmitter locations, or are only constructing systems inside buildings, the site survey may be the quickest and is certainly the most accurate method. If, on the other hand, you are planning multiple outdoor sites in various areas, the time and expense associated with acquiring and learning a software-based predictive model may prove valuable.

8.1 Predictive Modeling Tools

In the early 1980s, the first large scale cellular telephone networks began to be planned and constructed. In order to support these projects it was necessary to develop tools that would allow reasonably accurate prediction of RF propagation. Without such tools the only way to assess the coverage of a site was to perform a lengthy, complex and costly "drive test" on each site alternative. This involved erecting an antenna at the appropriate height (often 100 or more feet in the air), connecting a transmitter operating at the appropriate power, and then driving around the desired coverage area with a special receiver capable of recording signal strength and location. Obviously, this was a massive amount of effort when contemplated for thousands or tens of thousands of locations.

Luckily, during the 1970s a significant body of work defining the statistical properties of RF propagation was accomplished worldwide. This work led to the development of a series of algorithms that described the mean behavior of RF over a varying environment, terrain, and morphology. In their simplest forms, these algorithms described a series of curves that identified the propagation loss per decade over the various environments.

As cellular systems matured, the complexity of cellular systems increased and additional bands at higher frequencies became available. This led to additional refinement of the

algorithms used to predict propagation. These propagation models have names like TIREM, HATA, Longley-Rice, and Walfish-Ikegami. All are differing attempts to characterize propagation in different environments accurately.

It is important to realize that these propagation models are, at best, estimations of real-world propagation. The models are based upon statistical behavior, and the data that represents the terrain and morphology that the signal propagation is evaluated over is rather coarse and incomplete. Still, the models are useful for analyzing coverage and interference if you understand, accept, and design around their limitations.

8.2 Spreadsheet Models

In their simplest form, these algorithms can be implemented in a spreadsheet that is used for calculating the average coverage area of a communication site for the purpose of estimating the number of sites necessary to provide service in the designated area. Though they provide a more realistic estimation of coverage than simple free space loss calculations, they still do nothing more than allow you to draw a circle describing the range of a system. Figure 8.1 shows the different ranges achievable based on varying the path loss slope. Obviously, it is important to understand the average loss characteristics of the environment before selecting a

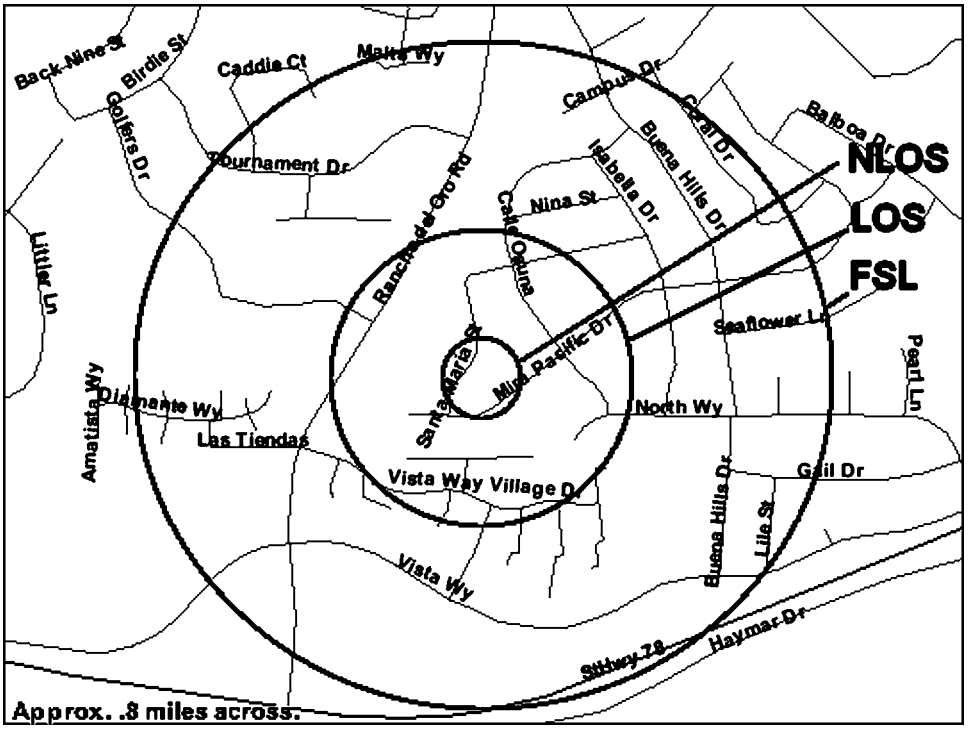


Figure 8.1: Map with Different Circles Describing Coverage Based on Various Loss Slopes

loss slope. Selecting an inappropriate slope will lead to significant over or under estimations of the actual propagation in the area.

This simplistic approach can be useful as a financial planning tool because it allows the approximate system costs to be known, but it is still not useful for designing a network that would provide known coverage in a known area, as well as predict interference levels outside the desired coverage area. For this, more complex models are necessary.

8.3 Terrain-Based Models

These models needed to take into account the actual terrain and morphology (land use) within the coverage area, and calculate coverage based upon those characteristics. These models begin to represent the actual behavior of the RF signal based on what obstacles it encounters while propagating outward from the transmitter. The same algorithms discussed above form the basis for these more complex models. The difference is that the models are applied over known terrain and (maybe) morphology. This is accomplished by using digital terrain data. Terrain data is available from several sources, including the USGS. Such publicly available data has a resolution of 1 km, 100 meters, 30 meters, and 10 meters. This means that the data is averaged into a block of the size shown. 100-meter data averages all the terrain in a 100×100 meter square, and represents it as a single elevation. 10-meter data averages the terrain contained in a 10×10 meter square. This averaging does lead to some inaccuracy in coverage prediction, so it is important to acquire the highest resolution data available. Morphological data is harder to find. It is normally custom digitized from high altitude stereometric photographs. Morphological data is also somewhat time limited in its utility, because trees grow and new buildings are constructed.

Modeling is accomplished by placing the simulated radio base station in a modeled environment representative of the actual area to be covered so signal strength could be predicted. This is done by looking out across the digital landscape represented by the terrain and morphological data, and calculating the mean signal strength based upon the environment present in that single slice. Each model predicts the mean loss using different parameters, but the results are the same: a plot identifying the expected propagation overlaid on a terrain map, or road map, or both.

In order to accomplish this, large scale computing power is necessary. The first of these models were run on mainframe computers, and as processing power increased, minicomputers. By 1984, programs were developed that could run on a PC. Today many programs exist and are available for purchase or license from various sources. All of these programs can run effectively on a modern desktop or laptop computer.

8.4 Effectively Using a Propagation Analysis Program

Acquiring a software package and operating it is not the hard part—effectively using it is. There are a number of issues that you must consider in order to assure the propagation

predictions reflect the real world. You must make sure you have accurate terrain information and morphological information. In addition, you should do field measurements from a number of sites in various settings and compare the measured results to the predicted results using the technology and band you are implementing. This comparison will show whether the model over or under predicts, and will allow you to "tweak" variables in the model in order to make the predictions line up with reality. By doing this you gain confidence in the model, and eventually you will be able to rely on propagation modeling to predict the behavior of new sites without always resorting to field tests and site surveys.

Propagation prediction software is a tool most effectively used by engineers with some experience in RF propagation. Without a certain level of experience, the tool provides little value. For example, propagation prediction software is only as accurate as the underlying information in the database it uses for predicting coverage. It is important to acquire the highest resolution terrain data available. In addition to terrain, the land has usage, or morphological features, such as buildings, roads, and trees. If these factors are not considered in the database, then accuracy suffers. Take, for example, New York City. Because the terrain is relatively flat, if you were to use terrain alone to predict coverage from a site with antennas at 100 feet elevation, you would predict large coverage area from any site in Manhattan. The actual propagation would be far less due to the high density of buildings in the area.

The coverage shown in Figure 8.2 shows the difference between predictions based on terrain only, and terrain plus average building clutter. While the propagation prediction is more accurate with average building clutter data added to the terrain, it is still inaccurate. The actual coverage from the site in question would look more like a "+" sign. The site is located at a street intersection, so there is no blockage along the street in the north-south and east-west orientations. Therefore the signals will propagate significantly further in those directions than in any other orientation. The predictive model did not identify this because the average building clutter data and terrain data did not have enough resolution to identify roads vs. buildings. Manhattan is an area where the use of data acquired from stereo photography may be relevant. Because the digitized image contains information about the actual location and orientation of streets and buildings, as well as the actual height of buildings and the level of the streets between them, it provides a much more accurate representation of the real world. The propagation software would be able to "see" the actual environment as a series of deep narrow canyons (streets) among tall dense canyon walls (buildings). Using this type of data instead of average terrain and morphology will allow significantly more accurate propagation prediction. The downside of this approach is cost. The database generated from the stereo photography is quite expensive to obtain.

While Manhattan is an extreme example, the same inaccuracies propagated by the use of average terrain and morphology will exist in any area where there are natural or manmade objects on the ground. These objects affect propagation by both blocking the radio path and by

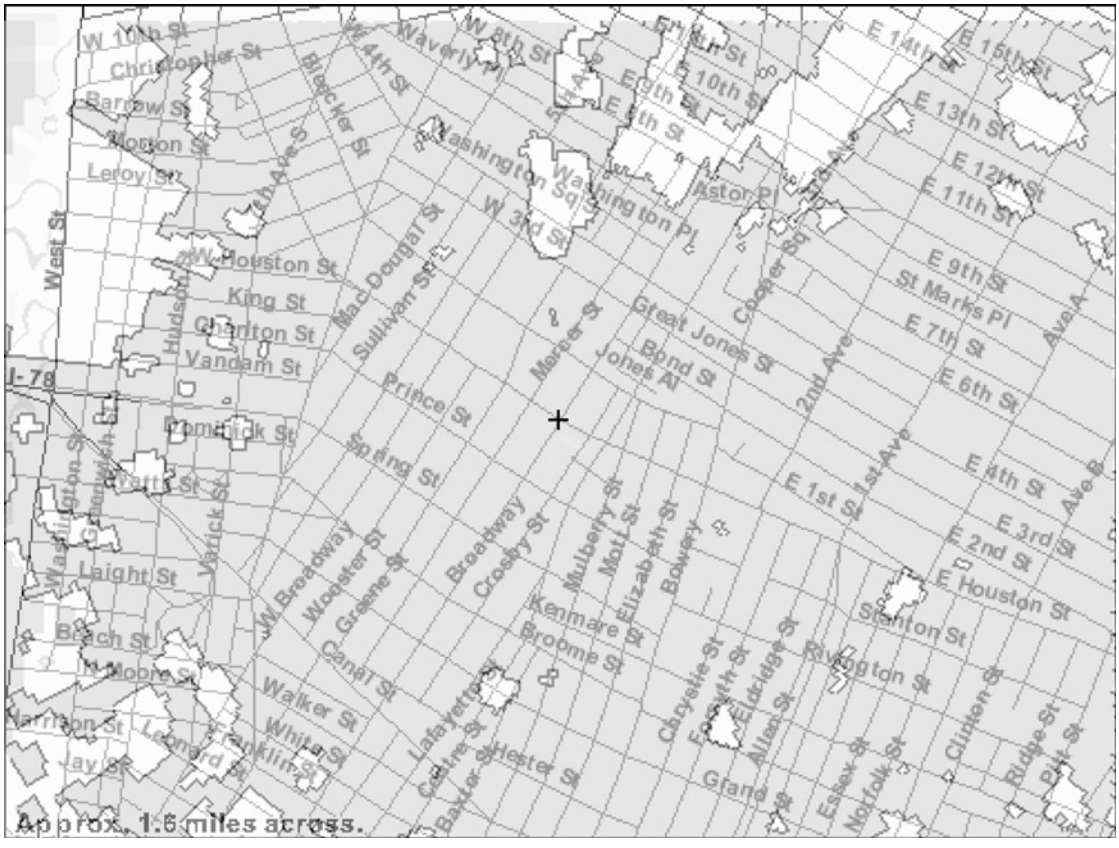


Figure 8.2a: Predicted Manhattan Coverage Based Upon Terrain Only

providing reflective surfaces that cause multipath. Both these effects lead to signal attenuation in the real world that, depending on the accuracy of the terrain and morphological databases, may not be appropriately considered by the propagation model.

This is why field surveys are necessary. The survey provides actual coverage information that can be compared to the predictive model. By comparing the two results you can begin to understand the average propagation of radio signals in different environments. In fact, this is how the predictive models were initially created. Millions of measurements were taken in different environments, and the results were graphed as signal strength vs. distance from the transmitter. Statistical analysis was done on the results, and a determination of the average loss over distance and impact of diffractive and reflective objects was determined. These statistical results were then represented as a set of algorithms that are used in the predictive models to characterize propagation across a given environment. Of course, you will not need millions of data points. We're not trying to create a new algorithm, just verify the accuracy of the one we've chosen.

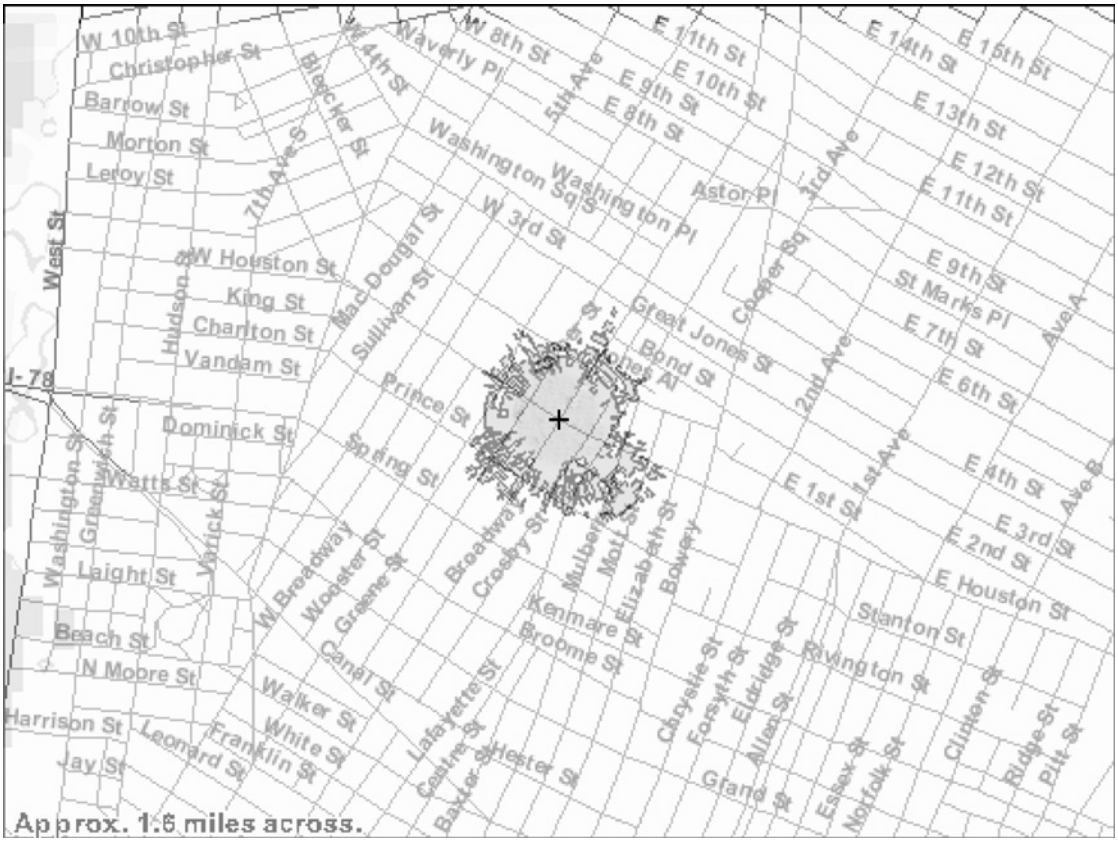


Figure 8.2b: Manhattan Coverage Based Upon Morphology

By using a combination of predicted propagation, field surveys, plus intuition and logic, you can estimate the coverage available from a site. Let's take the Manhattan example. The prediction based on terrain alone is useless for determining general coverage. It is, however, representative of the coverage achieved on the streets in direct view of the site; thus the "+" shape defines the extended coverage area associated with the immediate cross streets. The buildings surrounding the site will attenuate the signal by over 25 dB each, so you can assume that there will be some in-building coverage in the buildings directly adjacent to the site, but the intervening buildings will attenuate the signal so much that the signal will not appear in buildings a block away. Given these assumptions, you could estimate the coverage to be similar to Figure 8.3.

The astute reader may have noticed that I have shown coverage expectations for the above example without representing a number of key factors. What frequency was used? What antennas? How much transmit power? What receive sensitivity? What cable loss?

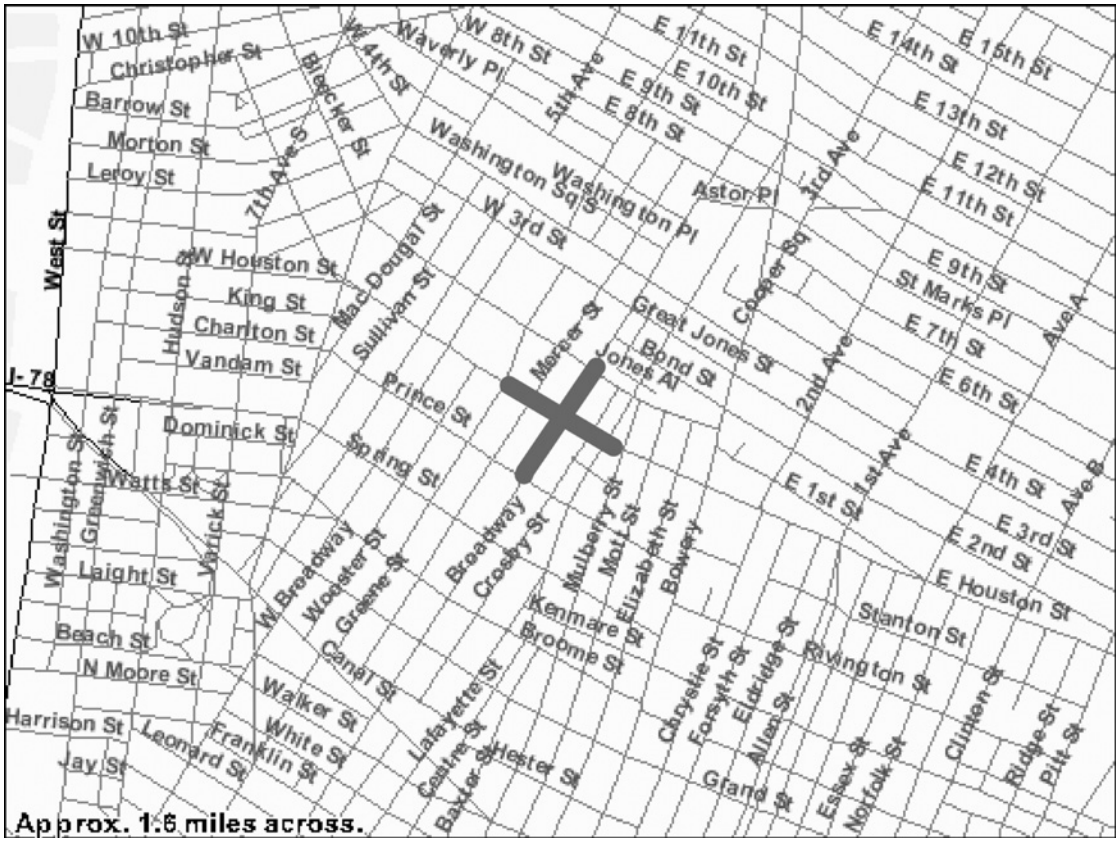


Figure 8.3: Real-World Coverage Estimate Drawing

In addition to terrain and morphology, these factors need to be known in order to effectively use any propagation model. These are the critical factors for determining maximum path loss, and are among the key variables used by the predictive modeling software to calculate coverage.

8.5 Using a Predictive Model

Let's take a look at some actual propagation modeling software, and load it up with real-world data and system variables. The software that generated the propagation plots in this book is a program called Radio Mobile Deluxe, written by Roger Coudé and made available as freeware. A copy of this program has been included on the CD-ROM accompanying this book. Roger is constantly updating the program and adding new improvements, so you may want to check and see if a newer version is available. Radio Mobile is based upon the Longley-Rice propagation model, and adds several morphological correction factors which can be used to change the model to more closely predict the actual propagation in an area based upon land

use and terrain. As shown in Figure 8.4, accurately setting these morphological parameters is critical. The best way to determine the right parameters is by doing site surveys in areas similar to those you wish to cover. Once this real-world data is available, it can be compared to the modeled results, and the morphological parameters associated with the different areas can be "tweaked" until the predicted behavior closely approximates the measured data. To do this, look at area averages and try to get them to match as closely as possible. Remember the model does not have actual morphology, so it does not know where individual roads, trees, and buildings are located. Because of this the model cannot accurately predict local shadowing caused by these objects. The best it can do is predict the average propagation behavior in an area based upon an average morphological density. Therefore you should not expect absolute accuracy of predictions against measurements. The accuracy of the model increases the further you get out of local clutter. If base station and far end equipment is located above the level of local buildings and foliage, the model will predict more accurately because the clutter is now nothing more than a source

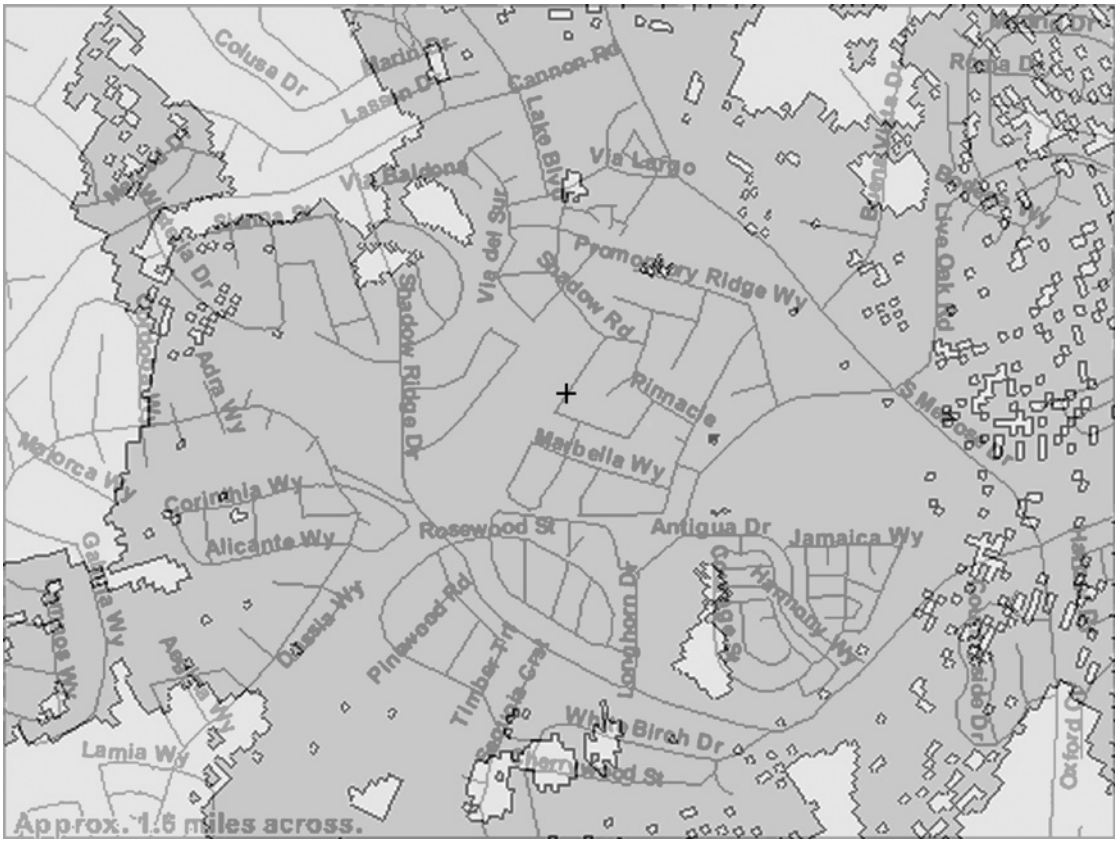


Figure 8.4a: Propagation Model Output Plots Terrain Only

Figure 8.4b: Propagation Model Output Plots Terrain Plus Forest Morphology

of multipath. The closer to the ground you place one of the stations, the more local morphological features will begin to impact the accuracy because of the local shadowing they generate.

Nonetheless, modeling is a valuable tool. Though it may not be able to tell all, it can often tell enough about the propagation in an area to give a level of comfort about what, on average, to expect as the coverage provided by a base station. This can be useful for ranking locations, or simply checking to see if the coverage of a site appears sufficient to justify its costs.

The first thing to do with this or any propagation prediction software is to gather the appropriate terrain and morphological databases that the model needs for prediction. Also gather any digital street maps that you may want to use as base maps for plotting your coverage on. These tasks are straightforward when using Radio Mobile, because the program is designed to access publicly available Internet databases containing terrain data as well as street maps.

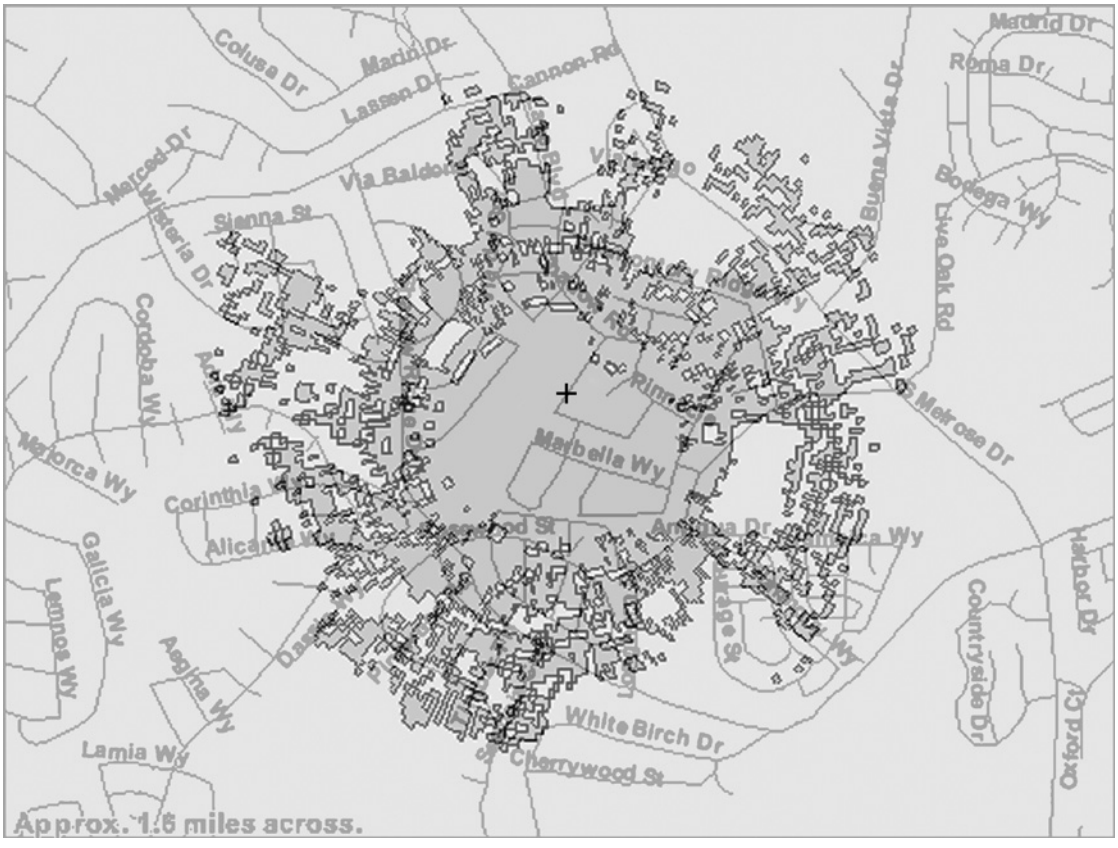


Figure 8.4c: Propagation Model Output Plots Terrain Plus Low Density Suburban Morphology

The next requirement is to know the RF performance characteristics of your equipment. These include frequency of operation, RF power output, receive sensitivity, and antenna characteristics for both the base station and the client end equipment. For this example I'll use the following RF characteristics, which are similar to those used for unlicensed 802.11b equipment operated in a WISP system:

- Frequency = 2450 MHz
- Base Station Tx power = 0.1 W
- Base Station Rx sensitivity = −94 dBm
- Base Station Antenna Pattern = 360 degrees
- Base Station Antenna Gain = 14 dBi
- Base Station antenna azimuth = 0 degrees

- Base Station Antenna Height = 60 feet
- Cable loss = 4 dB
- Far end Tx Power = 0.1 W
- Far end Rx Sensitivity = −94 dBm
- Far end Antenna pattern = 30 degrees
- Far end Antenna gain = 5 dBi
- Far end Antenna Height = 15 feet
- Cable loss = 0 dB

The final requirement is an estimation of the morphological correction factors to be added to the model. These can be determined by site survey, or by estimation from known propagation behavior in similar areas.

With the above information loaded into the model and applied to terrain data only it generates the plot in Figure 8.4a. This is probably an accurate representation of coverage in an open desert or coastal plain, but since populated areas do not have this open-terrain-only characteristic, additional losses due to the morphology of the area must be considered. Adding loss associated with forested morphology to the terrain yields the plot shown in Figure 8.4b. As you see, this significantly reduces the predicted coverage, as would be expected when trying to cover an area with dense foliage at these frequencies. Figure 8.4c shows the predicted behavior when the model is tuned for the morphology associated with a typical newly developed suburban subdivision, which is the actual environment being covered by this site. The correction factors selected were based upon data collected in this and similar environments, and provide a reasonably accurate estimation of the site's real-world behavior. This illustration shows the importance of accurate terrain and morphological data. It also shows why you should do some field-testing to validate predictions before blindly believing the output.

Gathering accurate field measurements is a task that can be accomplished in association with collecting other information about the area where the system will be deployed. It can be done as part of a comprehensive site survey.

8.6 The Comprehensive Site Survey Process

Since we are dealing with a radio-based wireless technology, it exhibits the irregular propagation characteristics of an RF-based service. As discussed in previous chapters, the RF signal is subject to fading, multipath, and many attenuation variables along its propagation path. These variables cause both the covered area and the spot coverage within the covered area to be difficult to predict accurately. One way to determine coverage accurately is by

performing a site survey. The site survey involves the temporary installation of equipment and the use of measurement tools to actually measure the signal in the desired coverage area.

Since the RF site survey requires such a significant effort, it should be conducted as part of a comprehensive site survey. A comprehensive survey will consider many factors that need to be known in order to deploy a system that meets the coverage, capacity and cost requirements set out for the network. In addition, the site survey can be an invaluable tool in determining what needs and limitations (like availability of power and network connectivity) exist vis-à-vis the system installation. The following site survey outline identifies the issues that need to be addressed during the survey.

8.7 Survey Activity Outline

- Location Identification
  - Latitude/Longitude
  - Location Address
  - Owner/Manager Contact Information
  - Structure Type
    - Structural Material
  - Area of desired coverage
- Identification of customer requirements
  - Coverage
  - Capacity
  - Security
- Identification of RF Zones
  - Based on Desired Coverage
  - Based on desired capacity
  - Based on interference management
  - Based on power and interconnect availability
- Equipment and Technology Selection
- Spectrum Analysis
  - Identify existing interferers

- RF Survey
  - Measure coverage from test locations
  - Plot actual vs. desired coverage in area
- Network design
  - Interconnect to radios
  - Interconnect to rest of network
  - Capacity
  - Security
  - Access control
  - IP addressing
- Equipment Selection
  - Transceivers
    - Make
    - Model
    - Frequency
    - Vendor
  - Antenna
    - Make
    - Model
    - Pattern
    - Vendor
  - Mounting Equipment
    - Type/Description
    - Vendor
  - Network
    - Interconnect

- Cable
- Radio
  - Routers
  - Switches
  - Hubs
  - Access control
- Conceptual Design
  - RF
  - Network
  - Interconnect
    - RF site interconnect
    - Network interconnect
      - To internal data network
      - To Internet
  - Costs
    - Equipment
    - Labor
    - Recurring
- Facilities lease
  - Availability and cost
- Site lease
  - Availability and cost

8.8 Identification of Requirements

Before equipment can be placed and measurements taken, it is important to determine what is expected of the network and how it will be used. The area to be covered, the number of users, and the services used all have an intertwined relationship. Understanding them will be of great help in determining the best equipment solution and locations to effectively serve the users' needs.

For example, the design requirements of a system to provide Internet access to a 20′ by 20′ coffee shop and provide service to five simultaneous users is significantly different from a system designed to cover an entire 20,000 square foot office area and all the computer users in it. Differing even more are the requirements of a Wireless ISP (WISP) that wants to provide ISP services to a town or community.

In the first case above, the room is small and the user community is also small. It is realistic to expect that a single-access point located within the room will provide sufficient coverage and capacity. The outcome of the site survey in this case is determination of the best location for the access point based upon RF coverage and ease of getting power and Ethernet cabling to that location. The survey should also identify what other hardware, like a router or gateway, is required to implement the service.

The second case is much more complex. How is the office space laid out? What construction materials were used in interior walls, ceilings and floors? How are the users distributed? What concerns does the customer have regarding signal leakage out of the building? What cost does the customer have in mind to provide this solution?

All of these issues will affect not only the RF issues surrounding the location, power level, antenna configuration, and channel reuse of the system, but also the network requirements surrounding traffic segmentation, security, routing, and switching.

The WISP case has different complexities. What are CAPEX, OPEX, and revenue expectations for the system? What are the reliability expectations? How big an area is to be covered? What are the locations of available sites in which to install the access point or similar equipment? Does sufficient Internet access bandwidth exist at these sites? If more than one site is needed for coverage, what are the available backhaul methods? What is the terrain of the area? What is the morphology of the area? How many users will be in the covered area? What will their usage patterns be like? How will the users be distributed across the coverage area? How will the system provide access control? What security concerns does the WISP have? Will the user have Customer Premise Equipment (CPE) located outside their home, or is the service expected to provide RF coverage to CPE located within the residence?

As you can see, the larger the system deployment, the more complex it becomes. Knowing what is expected of the system is the first step in resolving the best method of either meeting those expectations or setting new, more realistic expectations.

8.9 Identification of Equipment Requirements

After gaining an understanding of the customer's needs and expectations, the implementer should be able to select equipment that best suits the needs of the environment. For example, a cheap consumer quality 802.11b AP may be the perfect solution for the coffee shop, because

it is inexpensive and includes a low-end router, DHCP server, and provides NAT functionality. Thus it is a one-box solution for this particular environment.

This consumer grade AP solution would be a poor choice for the office or WISP examples. The office solution has need of an AP with features like remote management capability, power control, the ability to use external antennas to customize the area covered by each AP, the best encryption available, and the ability to be upgraded with new firmware so it can offer state of the art capabilities for the longest period of time.

The WISP, on the other hand, is probably not best served by a traditional 802.11b AP. The requirement of large area coverage from minimum locations means that the equipment will need to be tailored to high EIRP devices with high gain antennas. Moreover, the equipment will be mounted outdoors, thus requiring weatherproofing to be a design consideration. There are numerous companies who offer such products (Motorola, Alvarion, Proxim, Navini, and Vivato to name a few) and new ones seem to enter the market every month. Some of these solutions are 802.11 compatible, others use proprietary air interface solutions.

Each solution has its place in the market. Understanding the requirements of the system will assist in selecting the manufacturer and solution that is best suited to serving those needs.

For the remainder of this chapter, it is assumed that a working knowledge of the capabilities of the equipment being contemplated already exists. If it does not, then one should become familiar with the capabilities and expected coverage of the equipment before embarking on the site survey. Many of the techniques outlined below can be utilized to determine the coverage and capabilities of equipment, and can be used to evaluate the equipment in a known environment.

8.10 The Physical Site Survey

Once system requirements are understood, a physical site survey can commence. Obtain as much existing information as possible. Items like topographic maps, satellite images, building blueprints, and so forth will be invaluable in planning the survey.

With these documents in hand, you can begin to physically survey the property. Walk or drive around the area to be covered to get a visual understanding of the area to be covered, noting any major obstacles to coverage.

If outdoor coverage is planned, one should look for and note dense trees, buildings, and hills between the radio site and the desired service area. Note how far away you can physically see the radio site from as many locations within the desired service area as possible.

For indoor systems note the location of metal or cement walls and floors, as well as the location of large metal objects like refrigerators. Also note the location of "utility walls," i.e., those walls that contain dense runs of piping and or electrical cables.

Determine where the equipment needs to obtain its data and power connections. If the survey is of an office building, and the equipment needs to be connected to an existing computer network, note where this network equipment is located, and how new cables will need to be routed to get there. If remote connections are needed—in other words the connection to a data source does not reside on the same site being surveyed—note where the telco facility room is located on the property and where the other end of the connectivity must go. Also note how cables or wireless facilities can be routed from a central point of interconnect to the radio site locations, and whether there is a secure space where the network and interconnect hardware can be located.

8.11 Determination of Antenna Locations

Determining optimal antenna locations is the key to a successful deployment. An optimal location serves a multitude of needs: it provides optimal RF coverage; meaning it can be optimized to provide sufficient coverage of the area without leading to significant interference elsewhere in the system, it has easy access to power, it has easy access to network interconnect facilities, it can be easily installed and secured, and it has reasonable access for future service needs.

Since 802.11 hardware is easily available and has a large base of testing tools, I'll use 802.11 as the basic technology to discuss the decisions and tools required for system design. Even if the system you are designing is not 802.11-based, you can use the same procedures and criteria in designing a network based on 802.16, 802.20, or any other standard or proprietary solution.

The first step is to select an equipment solution based upon the needs of the customer and the environment to be covered. Select solutions that will most easily or most cost effectively meet the coverage and capacity requirements of the area.

Once the equipment is selected you have a baseline for the RF transmit power, receive sensitivity, and antenna options. As previously discussed, these numbers are used to determine the available path loss using the following equation:

$$L\,[\text{dB}] = Ptx\,[\text{dBm}] + Gtx\,[\text{dBi}] - Prx\,[\text{dBm}] + Grx\,[\text{dBi}] - M\,[\text{dB}] - Ca\,[\text{dB}] \quad (8.1)$$

where L is the link budget in dB, Ptx is transmit power, Prx is receiver sensitivity, M is fading margin, Ca is the attenuation of area construction material, and Gtx and Grx are antenna gains on the transmit side and receive side respectively.

Using the conservative power levels and antenna gains associated with common AP equipment used for indoor office LAN type deployments yields the following:

$$L = 15\,\text{dBm} + 0\,\text{dBi} - (-82\,\text{dBm}) + 0\,\text{dBi} - 10\,\text{dB} - 8\,\text{dB} = 79\,\text{dB}$$

Using the graph in Figure 8.5, you can see that in an office environment where the propagation will consist of some line-of-sight and some non-line-of-sight paths, the expected coverage of a single AP location could range from 60 to 150 meters depending on the actual conditions

of the path. If no interior walls block the path, the signal will propagate further. High-density walls will attenuate the signal more severely.

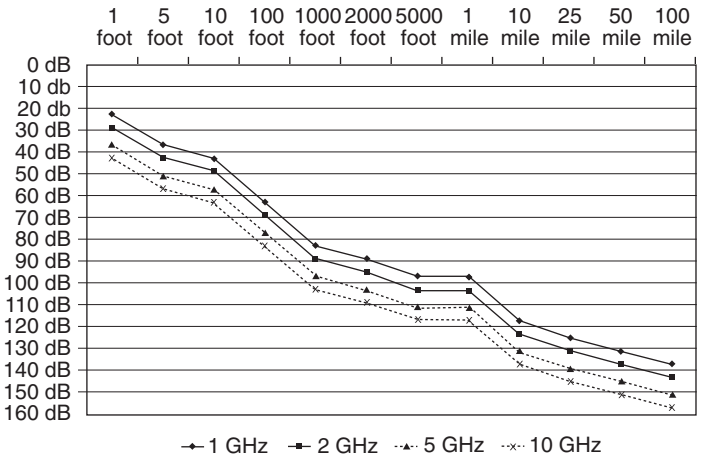


Figure 8.5: Chart Showing Loss vs. Distance at 1, 2, 5 and 10 GHz

Remember, this is a simple graph, and does not take into account all the propagation variables that will be found in the field. The distances derived from the graph are average numbers. There will be areas of a building (like central cores and elevator shafts in a high rise building) that exhibit far greater attenuation than the average in the office environment. This is why a site survey is helpful: it allows you to measure the actual propagation environment so you can decide precisely where radio sites should be located in order to provide best coverage, capacity and interference management.

Depending on the physical layout of the space to be covered and the availability of power and interconnect, a number of location options are viable: you could use an omni antenna located on the ceiling in a central location, or you could use a directional antenna located high up in a corner or along an outside wall and pointing toward the space to be covered.

Possible solutions are shown in Figure 8.6:

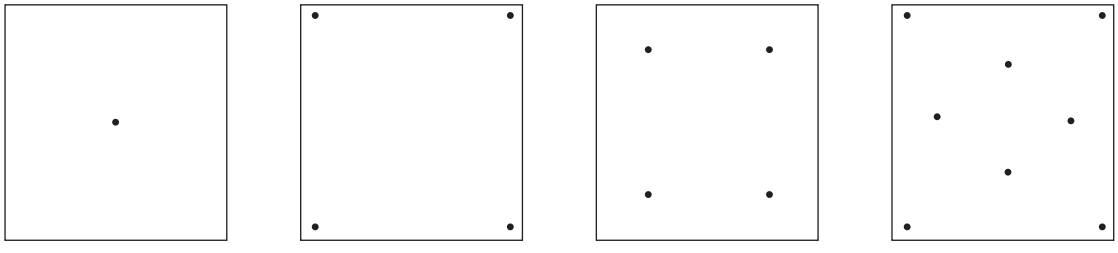
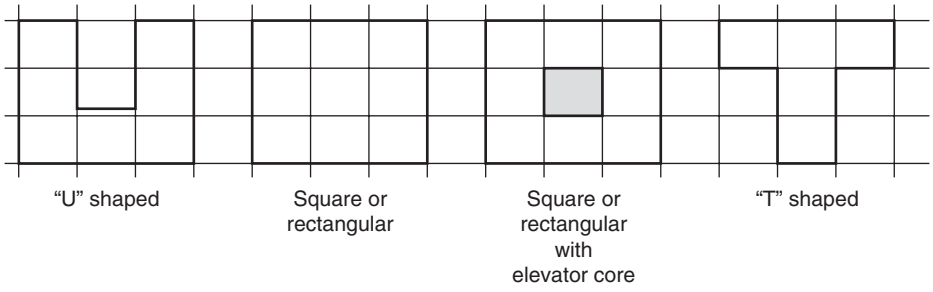


Figure 8.6: Base Station Location Option Diagrams

Real buildings may not be square or rectangular; however, the same location opportunities exist regardless of the building's shape. As shown in Figure 8.7, one method of breaking up a simple or complex floor plan is to grid it into squares that approximate the coverage you expect from each base station. By gridding the floor, you get a sense of how the space is organized, and you can analyze the user density within the grid. If you find too many users in a square, break it down further so you can see the actual area that needs to be served by each base station in order to accommodate coverage and capacity needs. With the area so divided, the selection of technology and the ideal placement of base station equipment can become much clearer.



By breaking down the area to be covered into a series of grid squares, sized to approximate the reliable coverage area of the technology, each area can be designed with the appropriate base station locations to serve it while minimizing interference in other areas.

Figure 8.7: Building Layouts and Gridding

8.12 RF Site Survey Tools

Now that possible AP or transmitter site locations have been identified, it's almost time to do some testing and measurement of the signal strength, noise, and interference in the environment. Before you can begin measuring these values, you need to acquire test equipment to do the measurement and data collection. Luckily such hardware and software is readily available and in some cases free.

If you are using proprietary hardware it will be up to the equipment provider to supply the measurement software and procedures. If you are using a standard 802.11 solution there are numerous software choices.

There are three classes of measurement and test software available for 802.11 RF testing and monitoring, each having its own benefits and limitations. The first class is the client manager that is included with most client cards. It is the simplest tool and has few or no features. Some can display the signal and SNR of the access point to which you are connected, others only show a bar graph of signal strength and the speed at which you are connected to the AP. Still

others, like the Lucent Orinoco client manager shown below, can show all active channels in the vicinity so long as they are associated with the same network and have the same SSID In addition they show the MAC address of the AP, the signal strength, noise level, and SNR. The Orinoco client manager also has rudimentary data logging capability. It can save the measurement results on 1 second or greater intervals automatically, or it can be set up for manual logging where the user must tell it to log a measurement and provide a text explanation to go with the measurement. This manual mode can be useful for taking "way-point" measurements, in other words measurements that are correlated to a known point in space.

This class of software can be useful in small-scale site surveys, like the coffee shop example mentioned earlier, or to use in spot-checking coverage in a larger deployment. It can also be a useful troubleshooting tool because each user of the network will have this software installed in their computer when the wireless card is installed. A first echelon of troubleshooting user problems would be to have the user open the client manager program and look at the information displayed.

The next class of software contains the free solutions like Kismet if you use Linux, or Netstumbler if you use Windows. Both have some limitation vis-à-vis the client cards and GPS formats they support, so care must be taken to assure compatibility with the rest of the test setup. The big benefit of these software packages is their improved feature set. They can be used to monitor all AP activity on all channels simultaneously. In addition they have a GPS interface, which makes them much more useful if outside measurements are contemplated. They log data to a file, and have the ability to export these files in a number of formats for post-processing and analysis.

The final class of software is the commercial package like Airopeek, AirMagnet, and Ekahau Site Survey. These packages are significantly more functional than the freeware packages. They also have compatibility issues with the Client Cards and GPS formats they work with. They are also expensive: $1000 to $2500 per copy of the software. It is also worthy of note that they have all been designed with certain purposes in mind, and they do not have 100% overlap of capabilities.

For example all of the above solutions are capable of collecting data, but only the Ekahau Site Survey software has the built in ability to create coverage maps directly from the software to a map image. To map the output of the other packages requires exporting the information and manually creating a coverage map with another software package.

There will continue to be new developments in the field of software and hardware for site surveying, monitoring and evaluation. It is well worth your time to search out currently available options. Evaluate several choices, and select the one that seems to best fit your particular requirements.

8.13 The Site Survey Checklist

Before you head out on your survey, take the time to assure that you have all the common items you may need on site. The obvious items are such things as:

- The selected radio hardware solution
- Your portable computer configured for measurements
  - Computer
  - Client card matching the chosen hardware solution
  - Measurement software
  - GPS
  - A cart or sling to carry the computer
  - Extra batteries and a battery charger/AC adapter
  - Any cables needed to connect external devices
- Spectrum analyzer

Less obvious items include:

- Mounting hardware for temporarily installing the radio equipment
- Tools to accomplish the temporary installation
- Extension cords to reach power
- Network cables to reach existing network
- Duct tape to tape down these cables
- Wire, tie-wraps
- Antennas appropriate to the initial design analysis
- High-quality coax jumpers to connect the antenna to the hardware
- Stepladder

8.14 The RF Survey

The survey is accomplished by temporarily installing the selected hardware solution at one or more of the predetermined locations, powering it up, getting it configured and operational, then using a client device and special software to collect information on signal strength, noise and SNR ratio.

Determine the best way to mount the hardware temporarily in the locations you've predetermined from studying the area to be covered. You want it secure, but do not want to permanently damage the area where you are mounting the hardware.

Once mounted, power it up and perform any configuration necessary to get it operating. Turn on your survey device and look for the signal from the hardware you just installed. If you are very close (within 10 feet) to the hardware you should see signal strengths ranging from −40 to −60 dBm. If you see appropriate signal strength from the desired equipment, you are ready to begin surveying the coverage area.

It is important to remember that 802.11 as well as any number of other technologies operate in unlicensed spectrum allocations. If the technology you are deploying is operating under FCC part 15 rules, a few initial tests are in order. First, use a spectrum analyzer to look for existing carriers in the band. Because Part 15 devices use different modulations, the only way to see and characterize the use of the band is to look at the spectrum analyzer plot and identify all carriers occupying the band. Next, check your survey software to see if it has identified any other equipment using the same standard as your equipment working on or near the channel selected for your equipment. If you see other operating hardware, make sure you set your equipment to operate on a nonconflicting channel. Also check the noise floor on the chosen channel to assure it's below −90 dBm. If the noise floor is over −90 dBm, there is a good possibility that another device using noncompatible modulation is operating on the channel. Because this noncompatible system will be seen as noise or interference by the new network, it is best that this channel also be avoided at this location in order to assure the best coverage and capacity from your device.

Now that you have the equipment functioning on a clear channel the RF survey can commence. Begin by moving around the desired coverage area and noting the signal strength and SNR at as many locations as possible. This is where your environment and selection of measurement software becomes critical. If you are measuring outside, GPS can be used for positioning, and with compatible measurement software GPS can be used to log location and signal strength and SNR at that location. Without GPS you will have to manually log as many points as feasible, as well as keep accurate track of your path. With the manually logged points and knowledge of how you got from point-to-point, you can manually create a coverage map from the information collected. Measuring indoors presents a situation similar to having no GPS. Since GPS does not generally work indoors, it cannot be relied upon for positioning in the indoor environment.

Continue to move away from the test node until the signal falls below noise level. Move back into the coverage area until you again acquire signal. If you have measurement software that is capable of collecting location data, take advantage of this capability and move randomly around the periphery in and out of coverage. The software should collect sufficient data to define the site boundary. Now move randomly inside this boundary, collecting as much data as feasible in the area covered by the test location. Use your mapping/plotting software to generate a coverage map of the area for further review and analysis.

If you do not have the ability to collect accurate positional data with your software, try the following procedure. Keeping the signal 1 to 3 dB above the noise floor (SNR = 1 to 3 dB), move around the entire periphery of the covered area. As you will begin to notice, there may

be significant changes in the location of this outer periphery. You may in some cases notice that a movement of 5 feet at the periphery requires you to move 20 feet closer to the test site in order to maintain the signal, while in other areas the opposite will be true. Continue to move about the periphery and accurately draw this contour on a map, picture, or blueprint of the site.

You will now have a map defining the limits of coverage of the test location. This is not the same as the useable limit of the site, defined as an area with sufficient SNR and fade margin to provide solid connectivity to the user, but does define the interference limit of the location. This will become important in considering the positioning of other RF locations and the channel selection for these locations.

Now repeat the measurement process with new SNR levels. 5 dB, 10 dB, and 15 dB levels are reasonable starting points. These contours should be plotted on the same map as the first measurement.

Perform a quick evaluation of the data you've collected. An example evaluation is provided by Figure 8.8. Does the test site cover the desired area? Are there any coverage holes in

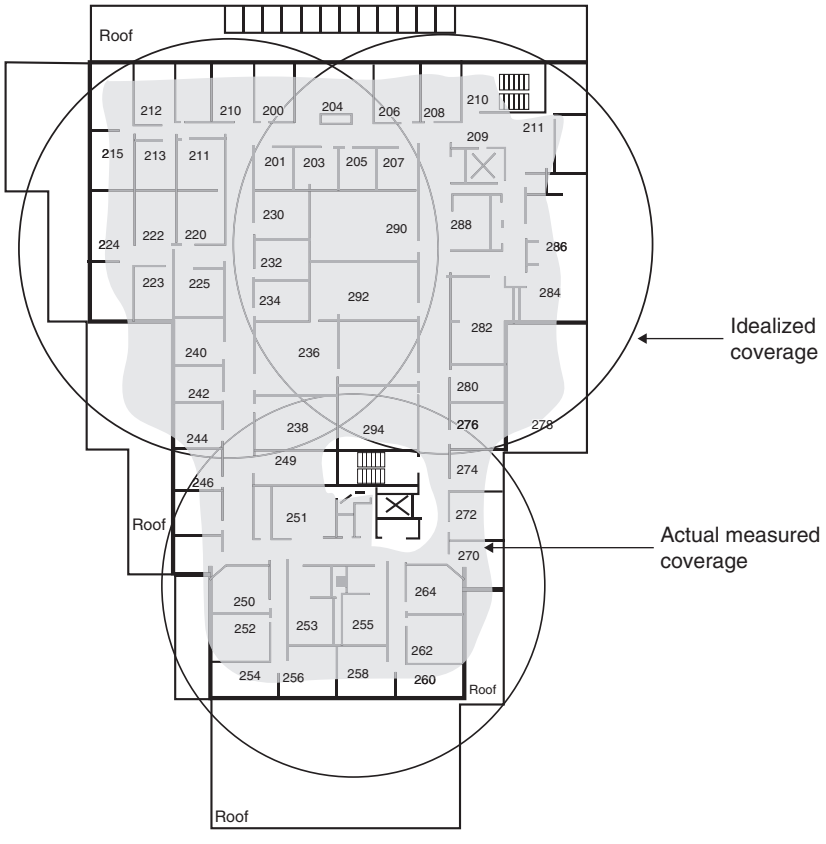


Figure 8.8: Coverage Plot

critical areas? If the coverage is not as expected, or there are critical coverage gaps, try to identify why the test site is not behaving as anticipated. Look carefully at the coverage contours; is there a clearly identifiable shadow in the coverage? If so, there is most likely a construction anomaly or other obstruction in the path. Having identified the location of this blockage, determine if it can be avoided by moving the test site to a new location that avoids the obstacle and perform the survey again. If, for some reason the test site cannot be moved, then the coverage holes will have to be accepted as areas of poor coverage by the system, or they may be correctable with a signal repeater located in the weak area.

Once you are satisfied with the coverage, repeat this procedure on all remaining test locations.

8.15 Data Analysis

With the data collected and visually plotted on a map or other image representative of the area to be covered, numerous details will become evident. These details will be useful in finalizing a system design that best meets the needs of the customer.

Because the analysis, and the changes necessary to conform the network to real-world needs is an iterative process, the review is best conducted utilizing a flowchart methodology. The flowcharts in Figure 8.9 are representative of the approach used to analyzing the data and make changes as needed to conform the solution to the real world as characterized by the survey. The first review should be conducted while the survey equipment is still mounted and operational. Upon seeing the initial survey results, you may decide that some optimization needs to take place. It's easier to do the additional tests now, rather than try to recreate the survey installation a second time.

The first analysis should be to review the coverage of the system. The first flowchart is used to analyze whether the site(s) provide coverage to the intended area, and if not, offer a number of alternatives to correct the coverage.

Once coverage is deemed acceptable, the second flowchart is used for managing excess coverage. Excess coverage is a problem on several fronts: Depending on how big the coverage extension is, it may be a security issue. For example if the wireless application is to provide connectivity for an office LAN, and the coverage of the WLAN extends outside the building into the parking lot or into the street, there is an opportunity for unauthorized access or monitoring of the WLAN and all the traffic it contains. Such an unintended coverage extension should be addressed first by minimizing the unintended excursion. If it cannot be completely eliminated, then additional security measures may be desirable on the WLAN. The second problem generated by excess coverage is overlap into the primary coverage area of another radio site. Because the client card normally identifies its primary radio site

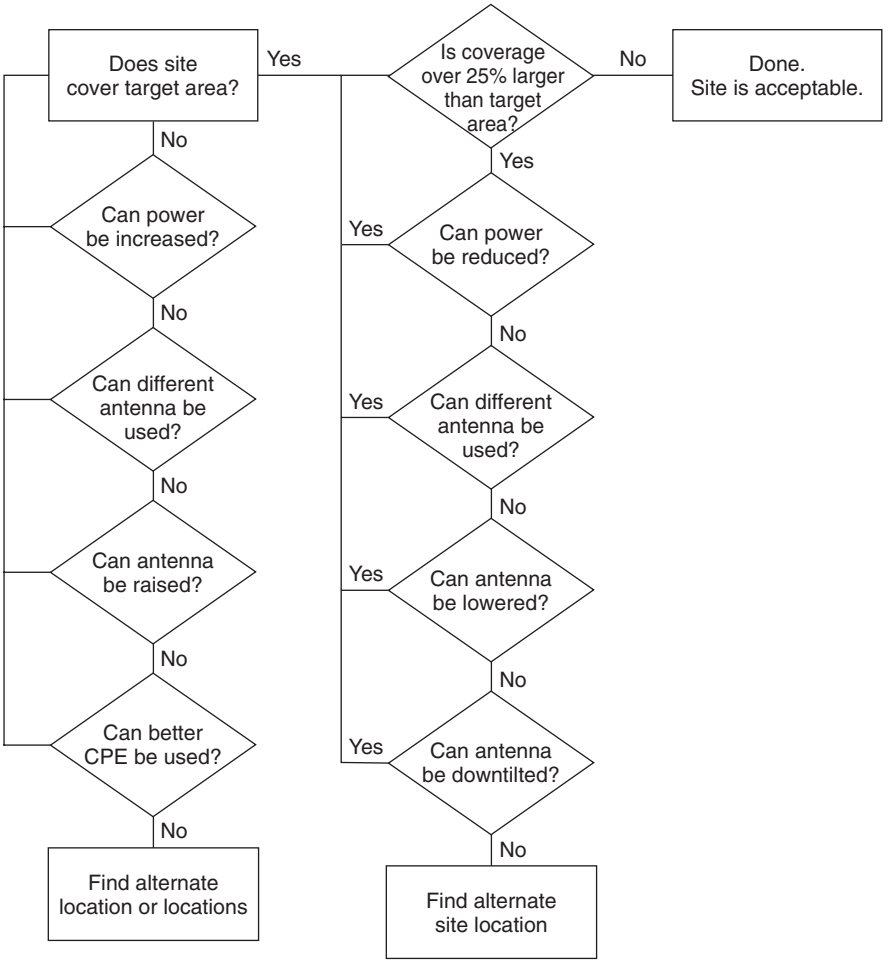


Figure 8.9: Site Analysis Flowchart

based on signal strength, there will be areas in the overlap where there is contention for the strongest signal. Fading and multipath exacerbate this problem, and lead to moment-by-moment changes in signal strength between the contending signals. The result of this can be the inability of the client to maintain communication with a single radio site. The client may "bounce" between competing signals at random, leading to throughput issues, or in some cases complete loss of data or connectivity.

More importantly from a system design standpoint, this excess coverage leads to interference with other radio sites using the same channel. This issue is critical in deployments using multiple radio sites, since there are a limited number of channels. This limited channel set will need to be reused by the radio sites over and over again within the coverage area. Interference is avoided by physical separation of co-channel reuse sites.

Power reduction, antenna selection, antenna downtilt, and site placement can all have a profound effect on controlling coverage of a site. The first corrective action to reduce the coverage should be power adjustment. In order to select an appropriate power level, you must refer to the link margin used when starting the survey exercise. The published receiver sensitivity plus a fade margin was used in the link budget as the base signal strength necessary to maintain a communication link at the desired throughput. Using this value (receiver sensitivity plus fade margin), reduce the power of the radio site until the edge of the desired coverage area is provided signal at this level. This is a straightforward process utilizing the data you've already collected. Your coverage map already shows signal strength across the coverage area. Look at the measured signal strength at the edge of the defined coverage area, and subtract the measured signal from the required signal. The result will be the number of dB the transmitter can be power reduced.

If the coverage area can be conformed with a power reduction, next make sure that the power reduction has not created any coverage holes inside the desired coverage area. Do this by identifying any areas on the survey map that show shadows or weak coverage. Subtract the number of dB you've reduced the power, and ensure that these weak areas still have sufficient signal to meet expectations. If they do not, then you could either increase the power until they do, consider adding an external antenna to the clients in the weak area, or add a signal repeater to the weak area.

You should keep in mind that no RF-based system is perfect. Even a well-designed system has coverage gaps. The best that can be expected is for the coverage to be useable over most (85 to 90%) of the covered area.

If power alone is insufficient to reduce the covered area, using a lower gain antenna, or a directional antenna downtilted toward the center of the desired coverage area may solve the problem. Lowering the antenna placement may also help. Unfortunately, all of these solutions will probably require additional survey time, since the previously collected data cannot be easily utilized to analyze changes of this magnitude.

Now that your design has been through enough iterations to assure maximized coverage inside the desired coverage area and minimized coverage outside that area, it's time to select channels.

Regardless of the technology selected, there will be a limited number of channels available for use. The first limitation on available channels will be those assigned by the government regulator in charge of spectrum allocation; other users of the spectrum in the area will cause the second limitation. In the case of technology using unlicensed spectrum (such as 802.11 products), the available channels might be used by devices as different as cordless phones and video transmitters.

If the system you are constructing has fewer transmitter locations than available channels, the deployment is simple: just assign unique available channels to each transmitter. If the number of locations exceeds the number of channels, then a frequency reuse plan will need to be designed and implemented. The coverage maps generated during the survey and corrected for power level are of great value in accomplishing this task.

The background provided in this chapter becomes the basis for deploying effective networks.

This page intentionally left blank

CHAPTER 9

Indoor Networks

Daniel M. Dobkin

9.1 Behind Closed Doors

Indoor environments are defined by walls, floors, ceilings, and, where applicable, windows and doors. Inside these constraining structures we find incidental obstacles such as the human occupants. The structural features and interior contents determine the propagation characteristics of indoor environments. Because major features don't change rapidly within a building, it makes some sense to give thought to their effects on propagation, but after thinking is done we shall certainly need to measure. Finally, buildings are occupied by people, often users of electronic equipment that can interfere with wireless local area networks (WLANs).

In view of these observations, we pursue an understanding of the RF side of indoor networks by first examining how buildings are built and the implications for propagation at microwave frequencies. After examining some surveys of signal strength for various sorts of facilities, we consider the properties of common sources of interference and conclude with some examples of tools to assist in setting up indoor coverage networks.

9.2 How Buildings Are Built (with W. Charles Perry, P.E.)

9.2.1 Some Construction Basics

We divide buildings into residential and commercial construction. Within commercial construction, we consider low-, mid-, and high-rise and very large structures.

Buildings are designed to do three things: stand up, serve the needs of their occupants, and burn slowly if at all. All other features of a building are subservient to these key requirements.

The first obligation of a building is to remain standing. This requirement is fulfilled in rather different fashion in residential or low-rise construction and in larger buildings. Residential buildings, and commercial buildings up to a few floors, rely on some of the building walls to carry part of the load of the roof and upper floors. Walls that bear such structural loads are known as *load-bearing* or *shear walls* (from the requirement to tolerate significant shear stresses without collapse); other walls serve merely to divide the interior space into convenient regions and are commonly known as *partition walls*. Shear walls are an integral

part of a building structure, not subject to modification without risk to the structural integrity of the building. Partition walls are generally lightly built and lightly removed or moved. The exterior walls of a building are almost always shear walls; some interior walls may also be load bearing. (Note that interior columns may also be provided for additional support without excessive consumption of space.) The construction materials and techniques used in shear walls are different from those used for partition walls, particularly in commercial construction; their microwave properties are also somewhat different (see section 3).

Mid-rise and high-rise construction are based on load-absorbing steel or reinforced-concrete frames. In a frame-based design, the exterior walls bear no significant loads and are not required to provide structural support but merely to protect the interior from the weather and provide privacy to the occupants. Thus, a wide variety of materials may be used for the exterior walls of such buildings, constrained by material cost, construction cost, appearance, and maintenance requirements.

Commercial construction practices are generally similar the United States, western and central Europe, and much of modern Asia, with certain minor exceptions related primarily to the cost of local assembly labor relative to local materials. Residential construction practices differ rather more noticeably across the world, because they are more subject to the vagaries of culture and personal preference.

9.2.2 Low-Rise Commercial Construction (One to Three Stories)

Low-rise commercial buildings typically share the load of supporting the structure between exterior walls, interior shear walls, and interior columns. A representative structure is shown in Figure 9.1. The exterior walls were traditionally hand-laid brick or concrete masonry units (known as cinder blocks to former college students) with steel reinforcing bars. This approach is still used on small buildings or where assembly labor is inexpensive and easily available. In most modern construction, exterior walls are made of reinforced concrete panels, which are fabricated on site and lifted into place: this approach is known as *tilt-up* construction.

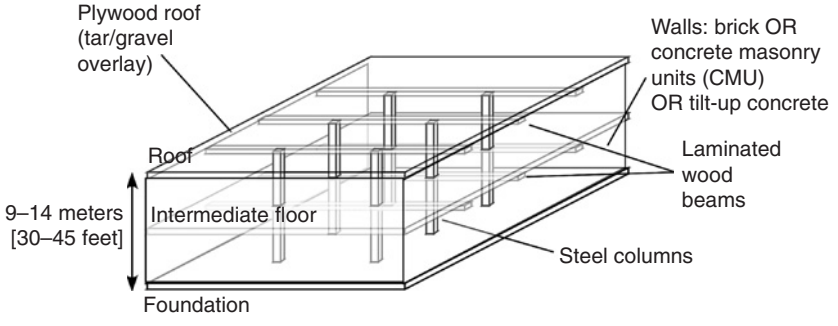


Figure 9.1: Typical Low-Rise Commercial Construction

A typical tilt-up panel is shown in Figure 9.2. Panels are fabricated by populating a wooden mold with reinforcing bars and then pouring concrete; the panels incorporate such major building features as openings for windows and doors and attachment points to facilitate connection to the cranes that lift them into place after cure. A single panel will usually form one complete exterior wall of a generally rectangular building.

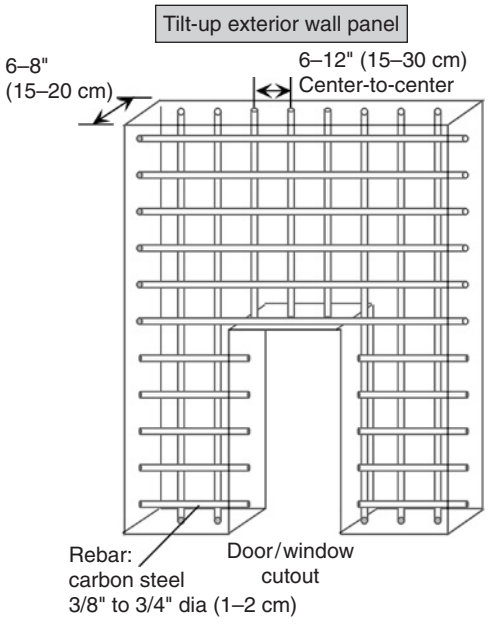


Figure 9.2: A Segment of a Tilt-Up Concrete Exterior Wall Panel

The lowest floor of the building is often laid directly on top of a concrete slab foundation. Beams that provide strength for the intermediate floors are themselves supported by ledges built into the tilt-up panels and by saddles attaching them to interior support columns (Figure 9.3).

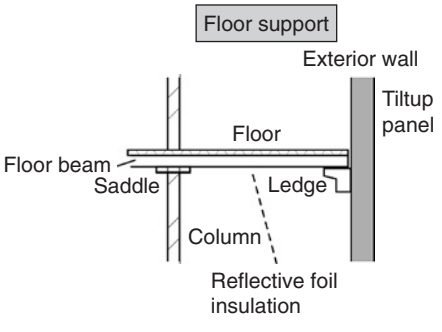


Figure 9.3: Intermediate Floors Are Supported at the External Walls and Internal Columns

The floor beams for small buildings are generally wood or wood laminates, nominally 4 × 12 inches (10 × 30 cm) in cross-section, though slightly smaller in actual dimensions. The intermediate floors are typically constructed of panels that rest on the floor beams (Figure 9.4). Each panel is built of plywood (thin layers of wood laminated together with misoriented grain for improved strength), with intermediate nominal 2 × 4-inch wood support members (*joists*), hanging from metal joist hangars, between the main floor beams providing local support. The plywood floor may optionally be covered with a thin coating of lightweight concrete, typically reinforced with a coarse wire grid.

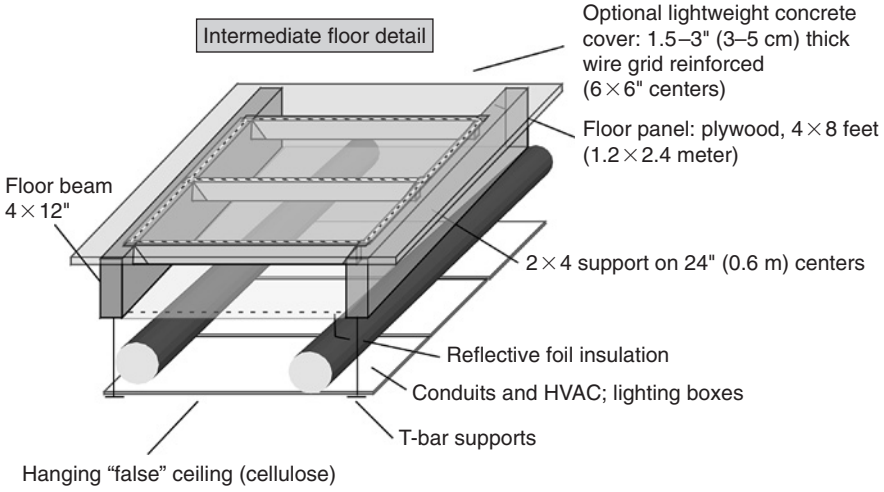


Figure 9.4: Intermedicat Floor and Structures Supported by the Floor

Intermediate floors form the ceiling of the story below them but rarely is this fact visible. A false ceiling, often constructed of pressed cellulose (i.e., paper) about 2–3 cm thick, is generally hung from metal T-bar supports attached to the floor beams to provide an aesthetically pleasing interior. These cellulose panels are easily popped out of their support frames to provide access to the space above. Between the false ceiling and the intermediate floor one will often find metal ductwork that provides heating, ventilation, and air conditioning services for the interior of the building. These ducts are typically constructed of thin sheet steel, though plastic ducting may also be encountered. Both round and rectangular cross-sections are used, and sizes vary from around 10 inches (25 cm) to 30–40 inches (around 1 m). The ducts may be wrapped with fiberglass or other thermal insulation materials. The author has found that these ducts are not always particularly close to where building plans say they are supposed to be; if in doubt, it is a good practice to pop out the ceiling panels and look. Fluorescent lighting panels, typically sheet metal boxes 0.5–2 m wide and 1–3 m long, also hang from the floor beams.

An alternative approach to construction of intermediate floors, with important consequences for microwave propagation, uses a thin corrugated steel deck covered with poured concrete to form the floor surface (Figure 9.5). Open-web steel joists may be used in this case instead of wooden joists, though the beams are still often wood laminates. As we discuss in section 3, conventional plywood/concrete floors cause a modest microwave loss (5–10 dB), whereas a corrugated steel floor in a low-rise building is a very effective shield, such that propagation between floors can be considered negligible. Thus, the type of floor used in a building is an important factor in determining whether neighboring floors represent common or distinct network coverage areas.

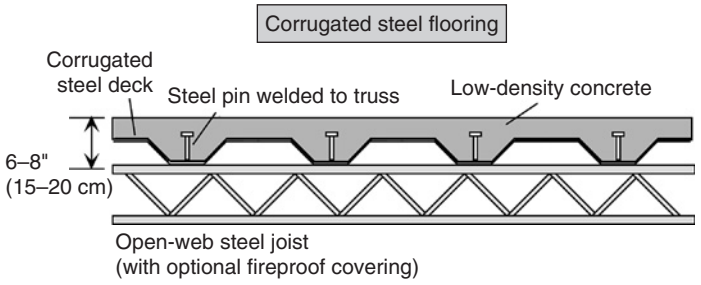


Figure 9.5: Corrugated Steel Intermediate Floor Construction

Interior shear (load-bearing) walls, if present, are typically constructed in the same fashion as exterior walls, using tilt-up concrete or hand-laid reinforced masonry. Interior partition walls are generally formed using gypsum (calcium sulfate dihydrate, $CaSO_4 \cdot 2H_2O$) wall boards assembled onto sheet-metal studs (Figure 9.6). Gypsum is popular because it is inexpensive, not readily flammable, and releases moisture upon heating. The gypsum is generally laminated with paper covering, and the wall board is usually painted or covered with patterned paper for aesthetic reasons. Studs are a few centimeters across and laid at 0.4-m (16-inch) intervals, so that they provide some scattering but little impediment to propagation at ISM or Unlicensed

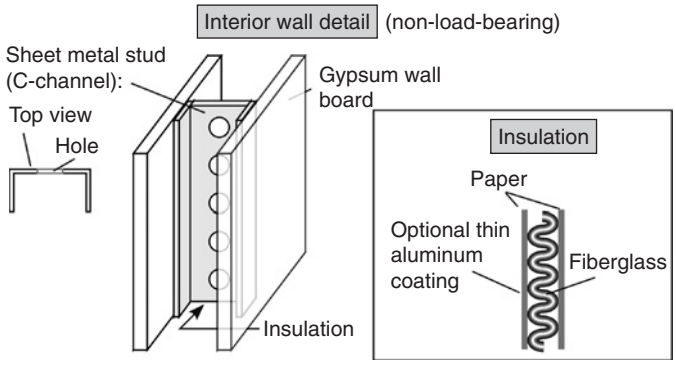


Figure 9.6: Interior Wall Construction, Commercial Buildings

National Information Infrastructure (UNII) bands. Interior walls may be insulated with fiberglass or similar material, typically wrapped in a paper cover. The paper can be coated with a layer of vacuum-deposited aluminum or aluminum foil; roof insulation is always aluminized, but the use of aluminized insulation for exterior walls is less frequent and is optional for interior walls. The thickness of the aluminum appears to vary from about 6 mm (deposited film) to as much as 50 mm (foil).

Roof construction is similar to intermediate floor construction (Figure 9.7). Wooden-laminate beams and joists support plywood roof panels, on which is laid a waterproof covering. The latter is typically composed of a layer of protective felt covered with tar-and-gravel-treated paper in commercial construction, though conventional shingled roofs using asphalt, tile, or ceramic-coated steel shingles may also be encountered. Conduits and ducting for heating, venting, and air conditioning may be suspended from the roof in the manner used for lower floors. Roofs are almost always thermally insulated, and aluminized insulation is generally used.

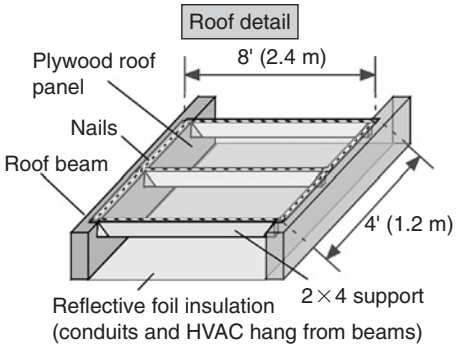


Figure 9.7: Roof Construction, Low-Rise Commercial Buildings

9.2.3 Mid-Rise and High-Rise Commercial Construction

Buildings taller than three floors rarely rely on load-bearing external or internal walls. Instead, structural integrity is provided by a frame composed of welded or bolted steel girders (Figure 9.8). In some structures, reinforced concrete columns are used instead of or in addition to steel framing. A *shear tower* made of reinforced concrete is often present at or near the center of the building, providing key services such as plumbing access, electrical services, and the elevator shafts and stairways that allow user access. The shear tower provides a relatively fireproof and mechanically robust escape path in the case of emergency.

Because the external walls are no longer load bearing, a wide variety of materials may be used in their construction. Interior construction techniques and finishings are generally similar to those used on low-rise buildings; intermediate floor construction is also similar to that described in section 2.2.

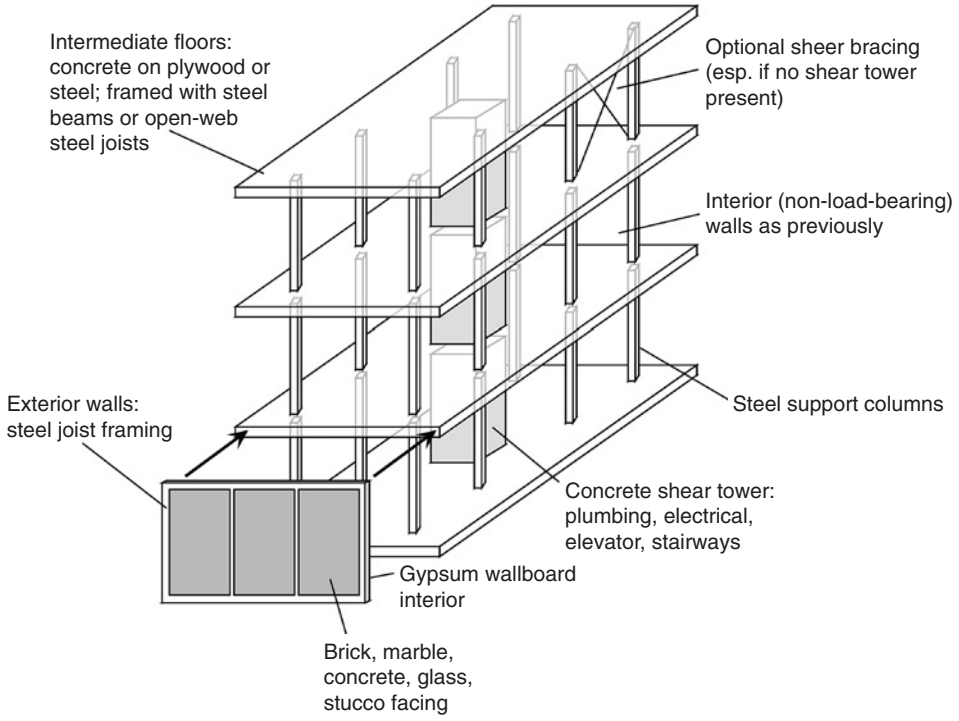


Figure 9.8: Steel-Framed Mid-Rise Construction (3–15 Stories)

Very tall buildings—skyscrapers—use elaborations of the steel framing approach used in mid-rise construction (Figure 9.9). Very large steel beams greater than 1 m in extent are required at the base of large structures, tapering to a more modest size at the top. Beams in large buildings are always provided with fire protection, either using gypsum or concrete sheathing. Concrete columns, formed from high-density concrete poured around a core of steel reinforcing bars wrapped in a steel confinement structure, are used instead of steel I-beams in some structures. Interior partitions and intermediate floor construction are similar to those used in low-and mid-rise buildings.

Exterior walls of tall buildings are often dominated by glass window area. The glass is often coated (*glazed*) to control insolation and consequent heating of the building interior. A wide variety of films is used. Commonly encountered materials are zinc oxide (ZnO, an insulator), tin oxide (SnO_2, a semiconductor with typical conductivity of around 3–10 mΩ-m or 300–1000 mΩ-cm), Ag, Ti, silicon nitride (Si_3N_4, an insulator), porous silicon dioxide (SiO_2, also an electrical insulator), titanium nitride (TiN, a fair conductor similar to tin oxide), stainless steel, and titanium dioxide (TiO_2, a good insulator). An example coating structure uses 40-nm (400 Å) layers of tin oxide separated by 10-nm layers of silver; thus, the whole structure is less than 140 nm (0.14 μm) thick, much less than a skin depth even for a good conductor. It

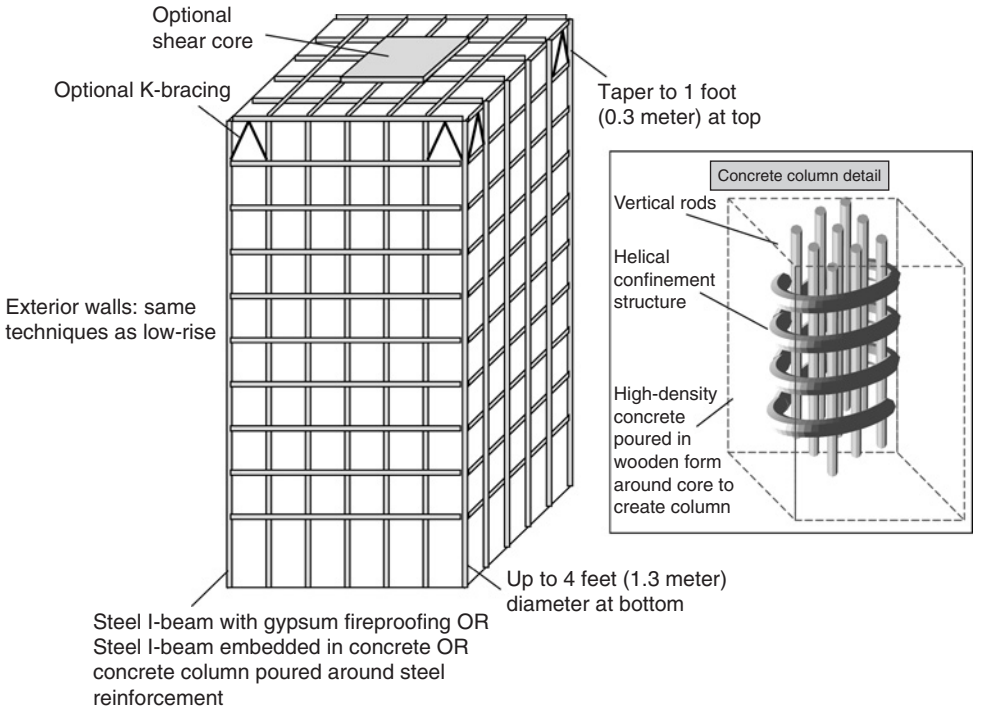


Figure 9.9: High-Rise Construction

is reasonable to expect that such windows are modestly reflective at microwave frequencies, but such thin layers should not result in strong absorption within the layer. The author is aware of anecdotal reports of glass coatings specifically designed to provide high absorption at microwave frequencies, purportedly used in airports and other facilities to impede cellular telephony coverage (and thus increase airport revenues through repeater charges), but has not been able to verify these reports to date.

9.2.4 Residential Construction

Residential construction practices around the world are more disparate than commercial construction, being more strongly guided by culture, history, and traditional preferences. The discussion here focuses on practices common in the United States.

In the United States, small single-family homes and small apartment buildings have traditionally been constructed using wooden frames for structural strength. Wood is also used for studs and rafters to support interior and exterior walls and roofing. Exterior walls are reinforced with plywood facing to improve shear strength of the framing. In seismically active regions like California, additional shear bracing of various kinds is used to ensure structural integrity in the event of an earthquake. Exterior wall materials are chosen for decorative value

and robustness and vary widely. Wood paneling, stucco (about which we have a bit more to say in a moment), brick, and aluminum siding are all common. Brick may be laid unreinforced in seismically stable areas.

Interior walls are typically gypsum (*dry wall* or *wall board*) over wooden studs. Services (electrical wiring, plumbing, vents) are provided in the interwall spaces. Ceilings are generally constructed of plaster laid over gypsum wall board. Roofs are plywood laid on wooden beams and rafters. Roofs are generally constructed of shingles (individual tiles overlaid so that the point of attachment of the shingle is shielded from rain water). Shingle materials include asphalt mixed with gravel, treated wood, and metal-backed ceramic. Tar-and-gravel roofs are rarely used in residential construction (though they are encountered in "Eichler" homes common in the southern Peninsula region of the San Francisco Bay area, where the author lives).

New U.S. construction is moving rapidly to sheet-metal studs and beams for interior and exterior walls, because the cost has fallen to below that of typical wood framing. The studs and beams are configured as C-shaped sheet metal channels as shown in Figure 9.6. Exterior walls are stiffened with metal straps or sheets. The total reinforced area is small because of the excellent tensile strength of metal beams. Metal plates are used at the windows to distribute lateral stresses. The exterior walls are covered with polycarbonate insulation followed by decorative coatings. Interior walls continue to use gypsum wall board to ensure fire resistance. Roof construction practices resemble those used in low-rise commercial construction, with plywood laid over sheet-metal rafters.

Plaster exterior wall construction has evolved somewhat over the last century (Figure 9.10). In the early part of the twentieth century, stucco exteriors were formed by laying plaster over

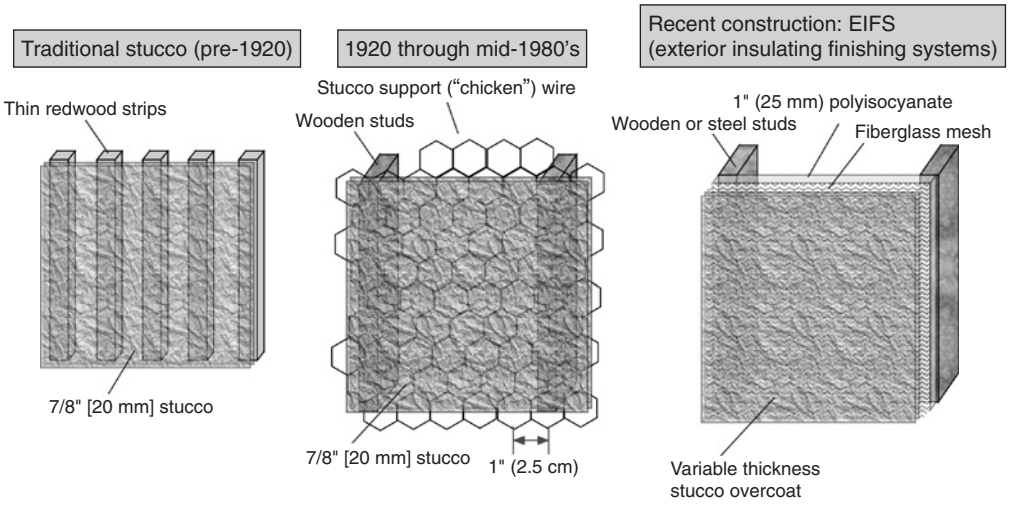


Figure 9.10: Stucco Through the (Modern) Ages

thin strips of redwood. From about 1920 through the 1980s, it was more common to place a layer of hexagonal stucco support wire as reinforcing material for the plaster layer. In recent construction, *exterior insulating finishing systems* are generally used. These consist of a polyisocyanate insulating layer on which is laid a fiberglass mesh, forming a bed for the plaster layer. Wire support may still be used for stucco layers of thickness greater than 1 cm (3/8 inch).

9.2.5 An International Flavor

Significant differences in construction practices (and construction quality, but that's another story) exist around the world, conditioned mainly by the local costs of labor and materials. In Russia and Poland, exterior walls are masonry and fill, with masonry interior walls common, because of low local labor cost and excellent thermal insulation properties of masonry. Lightweight concrete walls are becoming popular as well. Intermediate floors and ceilings are often a composite of fired clay tiles, which are shaped to provide conduits for services and ventilation, and reinforced concrete.

Western European construction uses brick, masonry, and concrete in internal walls to a much greater extent than is typical in the United States, both in commercial and residential construction. Concrete masonry units are also very popular in Mexican construction.

Taiwanese commercial construction is very similar to U.S. practice. Japanese commercial construction uses reinforced concrete extensively for its robustness under seismic stress. Residential construction varies widely. Japanese homes are generally wood framed, often with heavy tiled roofs. Hong Kong folks generally live in large apartment buildings constructed along commercial lines.

9.2.6 Very Large Structures

Structures that must have a large unsupported interior span, such as auditoriums, airport terminals, theaters, and convention centers, face special structural problems. The basic challenge is the support of a roof span without intervening columns. Small structures can use *cantilevered* beams (the term refers to the fact that both the position and slope of the ends of the beam are fixed to minimize sagging), but large structures exceed the shear strength of available materials.

Various solutions for large-span roofs exist. Ancient engineers used masonry arches and domes to construct cathedrals and public buildings; these rounded structures effectively convert gravitational forces to compressive stresses in the roof and walls, which stone and masonry are well fit to resist, but at the cost of heavy expensive roofs. Modern solutions often use long-span *truss* structures, composed of triangles of steel wires or beams arranged in a plane, so that stresses applied to the truss are converted into tensile stress on the members of the constituent triangles. A similar approach uses a *space frame*, in which tetrahedra formed from steel beams are assembled to fill space with minimal-length structures. Again, shearing

is not possible without tensile stress on the members of the truss. Another approach used for very large structures is to suspend a fabric roof from a network of support cables in the manner of a suspension bridge.

Tensile stresses from the weight of the roof must be transferred to the walls. Walls are formed of reinforced concrete or steel columns. In a technique known as *infill* construction, the space between the columns is taken up with non–load-bearing, decorative, weather-resistant materials such as brick, concrete masonry blocks, or glass blocks. Tilt-up construction is not practical for large structures, nor are wood beams used in modern practice.

9.3 Microwave Properties of Building Materials

From our examination of construction practices, it is apparent that the microwave properties of several common materials are needed to understand likely propagation results in indoor environments. We should like to know about the behavior of (at least) concrete, wood, gypsum, glass, and masonry.

A number of individual studies of particular materials are available in the literature, but the most complete single examination of building material properties at microwave frequencies the author has been able to obtain was performed by Stone and coworkers at the U.S. National Institute of Standards and Technology's Gaithersburg laboratories. Stone and colleagues measured attenuation after transmission through varying thicknesses of concrete with and without reinforcement, brick on concrete, concrete masonry units and brick walls, wet and dry soft lumber, plywood, drywall, and glass, all with a fixed test setup. Thus, the data set forms a valuable reference for relative attenuation of a wide variety of materials, measured in a consistent fashion.

There are some important limitations of the data that should be mentioned. No attempt was made to measure or correct for reflection from the samples, although the data were time gated to remove any reflections from the ambient around the test setup. Because the measurements were made in normal incidence, reflection coefficients of most materials are less than about 0.5, so the correction is of the order of 2 dB or less for a single interface. A complete correction would involve accounting for multiple reflections from both interfaces, but in the cases where this might be relevant, absorption is also high so that the correction is not of great significance. This correction is thus of modest import except in the case of samples with very little attenuation, where the distinction between absorption and reflection is of modest import in any case. Data were taken using separate systems in the 0.5- to 2-GHz and 3- to 8-GHz range; the low-frequency data appear consistent and sensible, but the higher frequency data show a complex frequency dependence and are quantitatively inconsistent with the lower frequency data. Thus, here we examine only the 0.5- to 2-GHz results, from which we must optimistically extrapolate to the frequency ranges of interest for WLAN applications.

The low-frequency results of Stone et al. are summarized in Figure 9.11. We can generally conclude that drywall provides minimal attenuation (on the order of 1 dB at normal incidence) and that dry plywood also has very little absorption or reflection. Wet plywood incurs an additional dB of attenuation (and/or reflection). Brick and soft lumber in typical thicknesses involve attenuations of 5–10 dB, dependent on thickness. Again, wet lumber incurs a decibel or two of additional absorption over dry lumber.

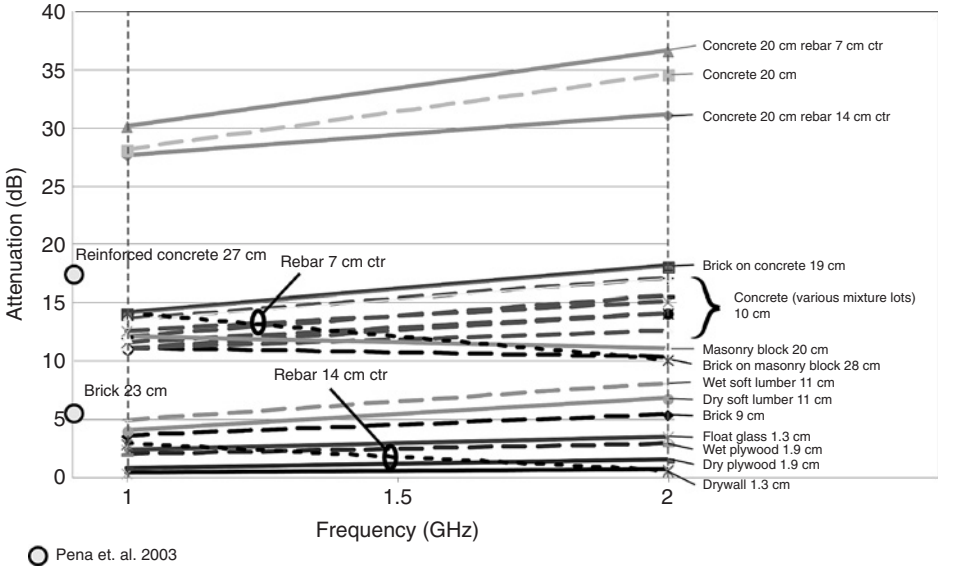


Figure 9.11: NIST Study of Attentuation of Various Building Materials at Microwave Frequencies; "Rebar" Denotes the Presence of Reinforcing Steel Bars Within Concrete or the Properties of Grids of Reinforcing Bars in Air (After Stone, 6055)

Concrete, by itself or in combination with brick, appears to represent a very strong attenuator, reducing the incoming signal by as much as 35 dB at 2 GHz for a 20-cm-thick layer. The presence or absence of reinforcement ("rebar") has a modest effect on concrete attenuation.

Results are also shown for grids of reinforcing bars in air, using 7- and 14-cm spacings. Recall that a wavelength at 1 GHz is about 30 cm and the rod diameter is about 19 mm, so $7 - 2 = 5$ cm holes are less than a quarter of a wave at 1 GHz. We see that half-wave grids provide little impediment to propagation, but quarter-wave grids are more significant. The reflection/scattering of the 7-cm grid extrapolated to 2.4 GHz (0.4 λ openings) is still 3 dB larger than that of the 14-cm grid at 1 GHz (also 0:41 openings), showing that the size of the steel rods is not negligible for the smaller grid. Nevertheless, we can conclude that conductive grids with openings significantly larger than half a wavelength at the frequency of interest introduce only a few decibels of reflection and scattering. Note that the effect of

rebar spacing within concrete is much less than that which would have been expected if the bare grid attenuation was added to that of unreinforced concrete; the large refractive index of the concrete means the wavelength within the medium is less than that in air, thus making the openings larger in terms of local wavelengths.

Also shown on the graph are measurements of the properties of brick and concrete obtained by Peña and coworkers in a separate study reported in 2003. These measurements were performed at roughly 900 MHz. The researchers obtained and modeled angle-dependent data, with full correction for reflections, to extract the real and imaginary dielectric constants and thus reflection and absorption behavior of the materials. Their result for a 23-cm brick wall (about 6 dB absorption) is generally consistent with that obtained for a 9-cm brick wall by Stone and coworkers. However, their results for a reinforced concrete wall show much lower (albeit still significant) absorption than that reported by Stone for similar thickness.

Recommended attenuation values for transmission through various materials provided by Wireless Valley, Inc., a commercial organization involved in both propagation modeling and installation consulting for indoor networks, are summarized in Table 9.1. These recommendations are generally consistent with the measurements of Stone et al. for glass, wall board (recalling that one wall penetration encounters two layers of drywall), wood, and brick, although Peña et al.'s results for brick suggest rather lower absorption. Reifsneider's recommendations for concrete walls are generally consistent with the data of Peña et al. (recalling that a 27-cm wall is unusually thick for U.S. construction) but not with that of Stone et al. In the course of some related work, I measured the absorption of a typical tilt-up reinforced concrete wall 20 cm thick at only 5–6 dB at 5.3 GHz.

Table 9.1: Recommended Values of Attenuation for Various Interior Structures

Obstacle	900 MHz	1.8 GHz	2.4 GHz
Interior wall (drywall)	2	2.5	3
Brick, concrete, masonry block wall	13	14	15
Cubicle wall	1	1.5	2
Wooden door	2	2.5	3
Glass window	2	2.5	3
Glass window, insolation "doping"	10	10	10
From Indoor Networks Training Course, Reifsneider, September 2003.			

Thus, we can see that multiple sources report more or less consistent results for common materials, but large discrepancies occur in reported behavior of concrete walls. What's going on? To understand why different authors report such differing behavior for this material in particular, we must undertake a brief examination of the nature of this ubiquitous modern building material.

Concrete, more formally *Portland concrete cement* (sometimes abbreviated as PCC), is a mixture of Portland cement and *aggregate*. The aggregate is crushed stone or other nonreactive material and plays an important role in the physical properties of the material but is probably not particularly microwave active. The cement is a powder consisting of tricalcium silicate $3CaO \cdot SiO_2$, and dicalcium silicate $2CaO \cdot SiO_2$, with smaller amounts of tricalcium aluminate $3CaO \cdot Al_2O_3$ and tetracalcium aluminoferrite $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$. When mixed with water, the tricalcium silicate forms a hydrated calcium silica gel, $3CaO \cdot 2SiO_2 \cdot 3H_2O$, and releases Ca^+ and OH^- ions ($Ca(OH)_2$, which is dissociated in aqueous solution). It is the hydrated gel that provides mechanical strength to the mixture after curing. The reactions that form the gel are complex, and the structure of the result is not well understood. It is well known that in the related process of formation of gels from pure silica, the nearly reversible hydrolysis of water by silica bridge bonds,

$$\equiv Si - O - Si \equiv + H_2O \Leftrightarrow \equiv Si - OH + HO - Si \equiv \tag{9.1}$$

(where $\equiv Si - O$ denotes a silicon atom attached to three other oxygen atoms, presumably within the bulk of the gel or particle), plays an important role in the formation of gel particles and in the redissolution and merging of particles to form an extended gel. It seems plausible to infer that an analogous set of nearly reversible reactions support the creation of a connected if porous gel from the initial powders during the curing of concrete. It is also well known from research into the interaction of water and silicon dioxide that water in the presence of silica can exist in three forms, distinguishable by their infrared absorption: free water molecules, hydrogen-bonded water, and silanol groups (Figure 9.12). The key relevance is that silanols, although still polar to a similar extent to water molecules, are not free to rotate in response to an electromagnetic potential. The vibrational frequencies of the attaching bonds are greatly in excess of microwave frequencies. Thus, silanol will not contribute to microwave absorption. Hydrogen-bonded water molecules are also relatively unable to move, attached as they are to the fairly rigid silica lattice, and will contribute little absorption at microwave frequencies. Thus, microwave absorption in concrete cement seems likely to be dominated by the residual free water molecules and should slowly fall as the concrete cures and water is taken up into bonded sites.

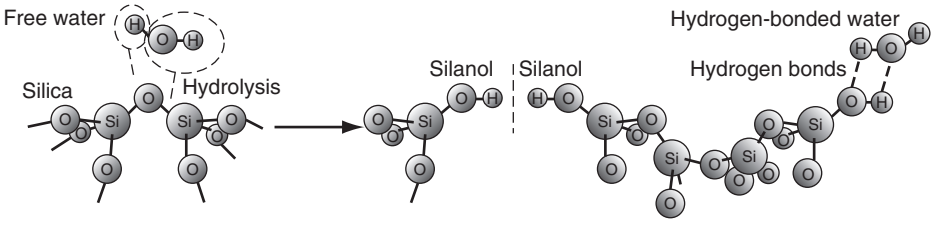


Figure 9.12: Interaction of Water With Silica: Silanol Groups Are Produced by Hydrolysis of Bridge Bonds and May Thereafter Form Hydrogen Bonds to Each Other or to Nearby Water Molecules

This supposition appears to be borne out by the microwave absorption data at 300 MHz obtained by Pokkuluri, shown in Figure 9.13. If we use Stone et al.'s data as a rough guide for the frequency dependence of absorption in concrete, we infer a change of about 20%/GHz, so that we should expect an absorption of about 5 dB/20 cm after 3 years of cure at 2 GHz, in qualitative agreement with the data of Dobkin and Peña (and perhaps also that of Wireless Valley, because they do not provide a reference thickness).

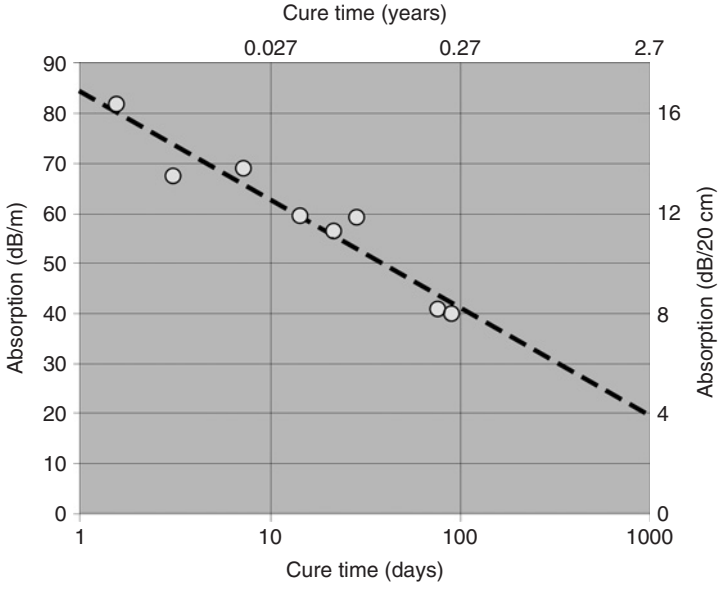


Figure 9.13: Microwave Absorption at 300 MHz in Concrete as a Function of Cure Time (Based on Data From Pokkuluri Thesis, Virginia Polytechnic Institute, 1998)

More to the point, we find that for samples cured some tens of days, we would expect an attenuation roughly two times higher than observed in samples cured for years, approximately accounting for the discrepancy between the results of Stone et al. and those of other workers. Stone et al.'s data include various mixtures of concrete incorporating differing amounts of the constituents and show significant (30%) variations in absorption resulting. The corresponding variation in fully cured concrete might be expected to be at least as large and perhaps larger, because absorption depends on the residual water available after hydration and thus on details of porosity and stoichiometry, likely to be influenced by initial composition and curing conditions. This supposition is supported by the data of Kharkovsky and coworkers taken on 15-cm-thick samples at 8–12 GHz. They found that two mixtures with water-to-cement ratios of 0.4 and 0.7 had absorption of 16 and 28 dB, respectively, after 6 months of cure. Kharkovsky et al.'s data also show slow continuous changes in absorption at long times: they found about 2 dB/month at 5–6 months of cure, roughly consistent with Pokkuluri's data at

much lower frequency. Thus, we can conclude that a typical 20-cm-thick concrete wall is likely to represent about 5–10 dB of absorption at 2.4 GHz and perhaps 2–3 dB more at 5 GHz, with the exact value significantly dependent on composition and curing conditions.

The author has been able to find only one brief reference to plaster walls (Ali-Rantala et al.) that suggested the absorption of plaster is slightly less than that of concrete. Referring to Figure 9.10, one can estimate that 2 cm of stucco would represent 0.5–1 dB of absorption, comparable with that of similar layers of plywood or gypsum. The use of wire reinforcement of typical dimensions, noted in Figure 9.10, appears likely to represent little impediment to propagation at 5–6 GHz; the present author has verified that a layer of wire composed of 2-inch (5 cm) hexagons attenuates a 5.3-GHz signal by only about 1.5 dB, and 1-inch (2.5 cm) hexagons impose 3.5 dB attenuation. At 2.4 GHz, 2-inch hexagonal screen represented an almost-insignificant 1.4-dB loss, but 1-inch hexagonal screen caused a decrease of 7.1 dB in signal strength from open space. Because stucco walls are typically very thin, the dielectric effects on wavelength would be expected to be modest, so one might propose that a residential wall using stucco reinforced by the typical support wire can represent as much as 7–8 dB of transmission loss in the 2.4-GHz ISM band.

Refractive indices and the corresponding reflectivity at normal incidence for various building materials are summarized in Table 9.2. It is clear that most building materials have refractive indices around 2–3. The reflection coefficient of a real slab of material will vary depending on its thickness if the material absorption is small (e.g., plywood or glass or gypsum in typical thicknesses) due to multiple reflections from the two interfaces. Estimated reflectivity versus

Table 9.2: Refractive Index and Reflection at Normal Incidence (From a Single Interface) for Various Building Materials

Material	*n*	Γ (Normal Incidence)	Source; Remarks
Brick	2	0.33	Pena et al. op. cit.[1]; Landron et al. IEEE Trans Ant Prop 44 p. 341, 1996
Concrete	2.5	0.43	Pena et al. op. cit.[1]; similar to CRC Handbook value for $CaSO_4$-$2H_2O$ of 2.3
Glass	2.5	0.43	CRC Handbook values for Corning 0080, 0120 (soda-lime glass, soda-lead glass)
Coated glass		0.7[2]	Landron et al. op. cit.
Limestone	2.7	0.46	Landron et al. op. cit.
Gypsum	2.2	0.37	Tarng & Liu IEEE Trans Vehic Tech 48, no. 3, 1999
Wood	2.2	0.37	Tarng & Liu op. cit.

[1]Pena data allows a range of $n = 1.7–2.2$ for brick and 2.3–2.7 for concrete.
[2]Landron data measured on exterior (coated) side, presumably dominated by coating.

slab thickness over the range of refractive index expected is shown in Figure 9.14. For $n = 2$ appropriate for those materials most often used in thin slabs (wood, glass, and gypsum), the reflectivity is at a maximum $\Gamma \sim 0.6$ at typical thicknesses of 1.3 to 1.9 cm (0.5–0.75 inches) versus the normal reflectivity from a single interface of about 0.4. Note, however, that even this maximum reflection coefficient only corresponds to a loss in the transmitted signal of about 2 dB ($10 \log(1 - 0.6^2)$). Thus, reflection plays a modest role in decreasing transmitted signal strength except at glancing angles of incidence. On the other hand, the high reflection coefficients suggest that reflected signals will be common in indoor environments, leading to multipath delay and fading.

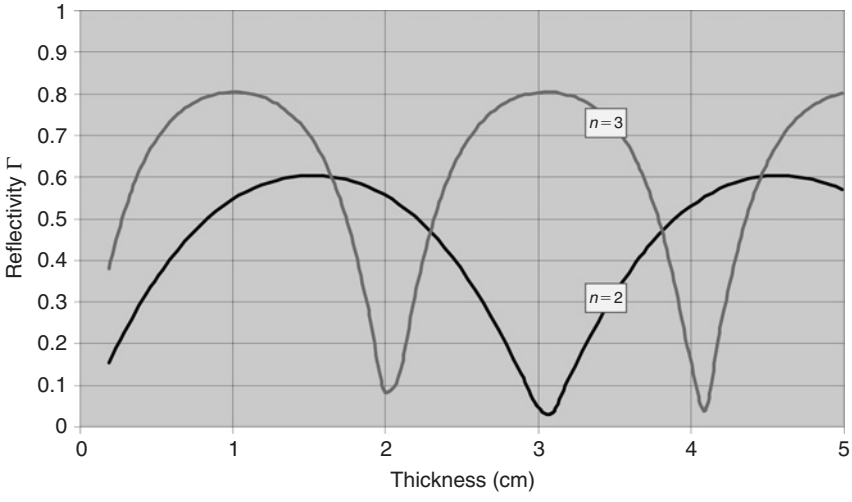


Figure 9.14: Reflectivity vs. Slab Thickness for Refractive Index $n = 2$ and 3; Slab of Refractive Index n Surrounded by Air ($n = 1$)

Some empirical data for attenuation resulting from transmission between floors are shown in Figure 9.15. Seidel et al. does not describe the means of construction of the floors studied, but by comparison with the data obtained by the present author, it is reasonable to infer that building 1 used concrete-on-wood or wood floors. The fact that loss does not increase linearly with the number of floors is ascribed to diffraction and/or reflection from neighboring buildings, which are not simply related to the number of floors separating transmitter and receiver. As one might expect, a corrugated steel floor of the type depicted in Figure 9.5 acts as an effective shield, leading to negligible direct propagation between floors. Seidel and coworkers also reported higher losses (on the order of 35 dB for one floor) for a second building, suggesting that they were also working with corrugated steel flooring in some cases. Buildings with steel floors are likely to be readily partitioned into separate WLAN domains by floor; buildings with conventional wooden floors will allow significant propagation between floors.

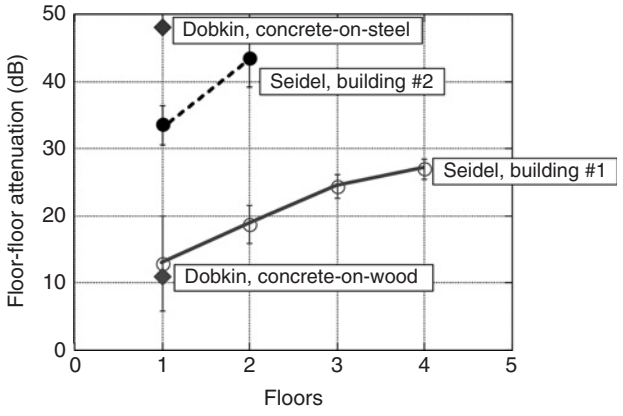


Figure 9.15: Attenuation Between Floors of a Building (Data From Seidel, IEEE Vehicular Tech Conf., 1992, p. 814, at 915 MHz; Unpublished Data by the Present Author at 2.4 GHz)

We can summarize the discussion of building materials with a few observations:

1. Absorption of materials commonly used in construction at thicknesses typically used can be classified as follows:
   - Lossy: thick concrete, masonry, or solid wood; 10–20 dB/wall
   - Modest loss: single-layer brick walls, thin masonry, wooden paneling, stucco walls; 5–10 dB/wall
   - Low loss: plywood, glass, wooden doors, cubicle walls, dry wall panels: less than 5 dB/obstacle
2. Most materials have refractive indices around 2–3, giving normal incidence reflection around $\Gamma = 0.4$ for thick layers.
   - Thin layer reflection depends strongly on thickness, varying from about 0.1 to 0.6; however, even at the peak this represents modest transmission loss of a couple of decibels.
   - Reflection loss will become significant for vertically polarized radiation on walls at angles of incidence more than 70 degrees.
   - Reflections from floors and ceilings are subject to Brewster's angle for vertically polarized antennas.

Thus, the received power in most cases will be dominated by the loss and reflection along the direct path between the transmitter and receiver. A first estimate of signal strength

(in the absence of large metal obstacles) can be obtained by counting walls along the direct ray. Glancing-angle wall reflections will subtract significant power from the direct ray and add to average power along long narrow corridors at the cost of increased fading.

9.4 Realistic Metal Obstacles

In this section we examine how realistic metallic obstacles affect propagation and compare the results with theoretical predictions.

Measurements were performed at 5.216 to 5.228 GHz in an open room with a concrete floor using a microstrip array antenna with approximately 20 dB power gain relative to an isotropic antenna to minimize the effects of reflections from ceiling features, walls, and floor on the results. Two obstacles were tested: a single 19-inch rack 0.65 m wide, 0.56 m deep, and 2 m high and a triple 19-inch rack 1.8 m wide, 0.85 m deep, and 1.6 m high. (These racks are very commonly used to hold server computers and other electronic gear in commercial and industrial facilities and provide representative examples of typically sized metallic obstacles.) The racks were unpopulated and thus consist of frames with sheet-metal front doors perforated by ventilation slots and a mostly open back side with support members crossing the access area at the top and center height. The test arrangement is depicted schematically in Figure 9.16.

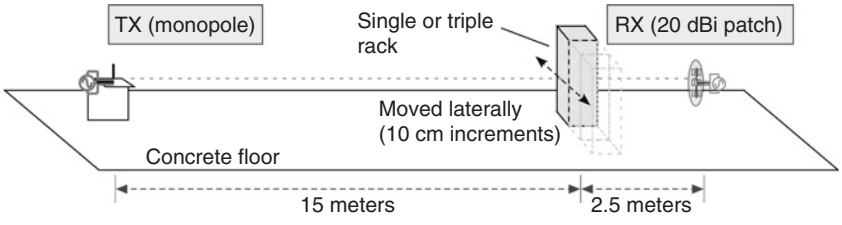


Figure 9.16: Test Range for Diffraction/Shadowing From Realistic Metallic Obstacles

The signal strength at four frequencies was averaged and compared with that obtained in the unobstructed case to produce a measure of shadow depth in decibels for each position of the obstacle. The results are summarized in Figure 9.17 and compared with estimates of signal strength obtained by numerical integration of the scattering equations for ideal flat plates of the same lateral dimensions.

It is apparent that the qualitative features of the shadows cast by the racks are well represented by the thin-plate models. The narrow deeply shadowed regions expected from the flat-plate model of the triple rack are not observed in the data; it is not clear if this is an artifact of the limited lateral resolution of the measurements.

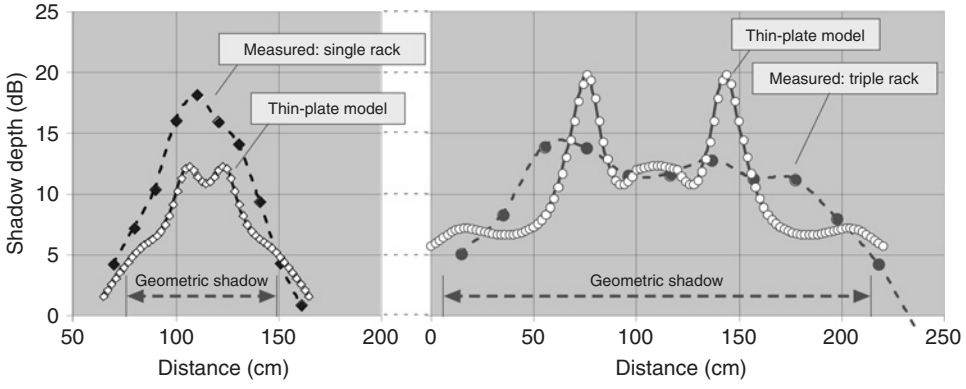


Figure 9.17: Shadow Depth vs. Obstacle Position for Single and Triple Racks; Average Over Geometric Shadow: 11.0 dB (Single Rack) and 10.1 dB (Triple Rack)

In normalized terms (equation [9.2]) the single rack is $u_o = 2.4$ wide at 5 GHz for the given configuration ($R_{\text{av}} = 2.1$ m) and can be reasonably regarded as tall. Equation [9.3] predicts a shadow depth half-way from the center line of 8 dB, in reasonable agreement with the observed shadow averaged over the geometrically shadowed region. Note, however, that the shadow depth in the central region is 5–7 dB deeper than expected for a flat plate and almost 10 dB deeper than the estimate of equation [9.3], presumably due to the complex actual geometry and finite depth of the rack.

$$u = \sqrt{\frac{2}{R_{\text{obs}}\lambda}}(x \text{ or } y) \tag{9.2}$$

$$(S)_{\text{dB}} \approx 20 \log\left(\frac{0.33}{2}\left[\frac{1}{u_{\text{o,l}}} + \frac{1}{u_{\text{o,r}}}\right]\right) \tag{9.3}$$

The triple rack is about 7.4 normalized units wide by 6.6 normalized units high. Equation [9.3] provides shadow estimates of 18.5 and 17.5 dB using, respectively, the width and height of the obstacle. This object cannot be reasonably regarded as a tall rectangular object, and so the approximations of equation [9.3] are suspect. A plausible approximation is to add the received power from two ideal shadowing objects (one very wide and one very tall), with each typical shadow calculated according to equation [9.3]; because the two shadows are comparable, we might as well just add 3 dB of signal power or remove 3 dB from the shadow depth, predicting a typical shadow of about 15 dB, slightly deeper than what is observed half-way from the edge.

We can summarize by saying that the observed shadows of complex objects of realistic dimensions are on the order of 10–20 dB deep, as we had predicted based on simple idealized models. The detailed behavior of obstacles of finite depth is not accurately predicted by simple flat-plate models, but the differences are comparable with fade margins of ±5 dB that must be

assigned to all estimates of field strength in any case. In practice, the effective shadow depth of obstacles on average will be reduced by the added power due to reflected rays following indirect paths within the indoor environment.

9.5 Real Indoor Propagation

We've now reviewed the sort of materials and obstacles we might encounter in an indoor environment and what their likely effects on a microwave signal might be. What does all this imply for operating a radio in an indoor environment? We first look at a simplified example of indoor propagation to get some idea of what to expect and then compare the results of the conceptual exercise with real data obtained in real buildings.

In Figure 9.18, the indoor propagation cartoon shows some exemplary propagation paths (rays) added. Along each path we've added a simplistic loss budget, accounting for 3 dB (reflection and absorption) in passing through interior partition walls, 5- to 7-dB loss upon reflection depending on angle, and 15 dB from thick concrete exterior walls. At the end of the process, each ray has accumulated a certain path loss due to distance traveled (not shown in the figure) and a certain additional delay relative to the direct ray; the result can be plotted on a graph of power versus excess delay, assuming a typical room size, as is done in the inset. (Note that the delay here is much too imprecise to be used to specify the relative phase of the various ray paths, so we can add powers to get an average power, but we can't specify an actual power with any confidence.) Even though this example appears somewhat complex, it

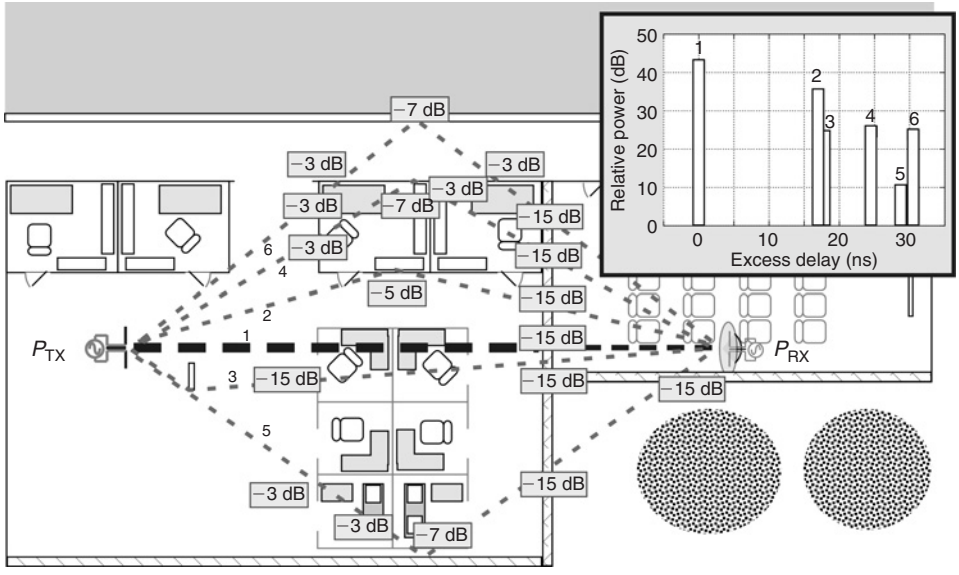


Figure 9.18: Typical Multipath Propagation in a Building With Reflection, Scattering, and Absorption Losses; Power (Arbitrary Units) vs. Delay Is Shown in the Inset

is hardly realistic: only a few reflected rays and one scattered ray are included, and we have ignored multiple reflections and reflections from the floor and ceiling or obstacles therein.

From the inset graph, we can see that the direct ray is the largest contributor to signal power, even though the configuration is hardly line-of-sight (LOS). This is because the large concrete shear wall attenuates the direct and indirect rays equally. The lower power rays will contribute both to fading, as their relative phase varies, and intersymbol interference, the latter being determined by their delay relative to a symbol time. Here, because we have included only a few rays in a small pair of rooms, the maximum delay of 30 nsec would be unobjectionable for 802.11b: the delayed rays would muddy the symbol transitions but have little effect in the middle of the symbol where sampling is typically done. However, we can see that higher symbol rates and higher modulations might start to encounter significant intersymbol interference even in this small propagation example.

From this simple example, we tentatively conclude that in a real environment we expect to see a number of rays of decreasing relative power, with characteristic delays set by the path length between the various reflecting and scattering features in the environment. Paths with longer delays will have both more path attenuation and additional loss from reflections and absorptions and so appear at lower relative power. Because buildings used by humans usually are partitioned into human-sized areas of a few meters, we would expect delays to occur in rough multiples of 10–20 nsec (3–6 m or 10–20 feet). When a clear line of sight is present or all indirect and direct rays encounter the same main obstruction, the received power will be dominated by the direct ray, leading to Rician-like distributions of received power; when an indirect path suffers the least absorption, we might see many rays of equal power and expect Rayleigh-like behavior.

In Figure 9.19 we show some examples of real indoor propagation data, from Hashemi and coworkers, in a facility described as a medium-sized office building. The qualitative features are similar to those of our simple example, though we see many more rays at low power and long delay, presumably resulting from multiple reflections. There are many fairly distinct peaks, presumably corresponding to discrete ray paths, with characteristic separation of some tens of nanoseconds. Figure 9.19 also depicts the signal-to-noise ratio (S/N) required to demodulate 64 quadrature-amplitude-modulation (QAM) symbols. In clear line of sight, the power is obviously dominated by the direct ray, and a 64QAM symbol would see a multipath delay of about 60 nsec. The non-LOS measurement shows about 18 dB attenuation of the direct path relative to the best indirect path; in this case, because of attenuation of the direct ray, many more indirect rays are comparable in power with the largest ray, and a 64QAM symbol would need to deal with delays up to about 200 nsec. We can also see why Rician-like behavior would seem reasonable in this case for the LOS received power, where the received power is dominated by the direct ray with three or four small rays about 20 dB down, and Rayleigh-like behavior would be reasonable for the non-LOS case where about 11 roughly equal rays (counting down to 20 dB relative attenuation or an amplitude of 0.1) make contributions to the received power. The LOS result has a beam at 15-nsec delay about

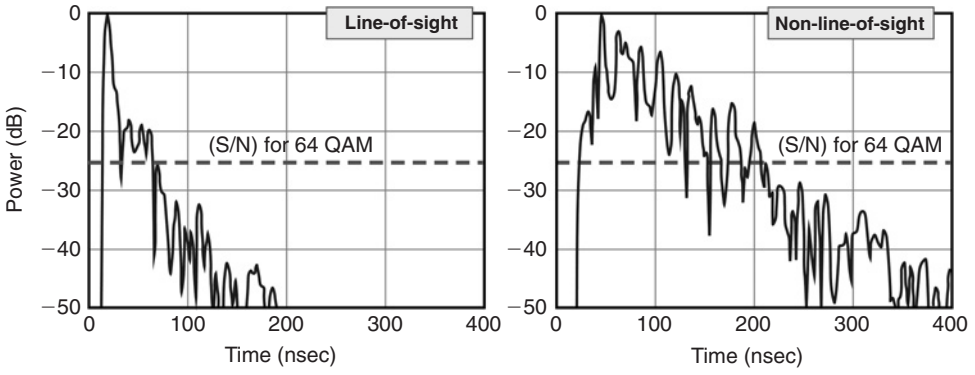


Figure 9.19: Indoor Power-vs.-Delay Data; Medium-Sized Office Building, 5 m TX-RX (From "The Indoor Radio Channel," H. Hashemi, Proc. IEEE 81, p. 943 (1993), Data by David Tholl of TR Labs); based on original image © 1993 IEEE

18 dB larger than the largest delayed beams, corresponding to a Rician v_m of about 10σ. A gradual decrease in beam power extends out to about 200 nsec, after which very little power is observed: One might guess that this corresponds to the size of the building in which the measurements were taken, with longer delay times resulting only from multiple interior reflections or objects beyond the building walls and thus significantly attenuated.

To provide some idea of how delay varies over several different facilities, Table 9.3 shows summarized data at 1900 MHz from several different buildings, measured by Seidel and coworkers. They found characteristic delays of around 100 nsec, similar to the average behavior seen in Figure 9.19. The longest RMS delays, as large as 1500 nsec, were observed when transmitter and receiver were on differing floors; the authors suggested that the long delays are the result of reflections from neighboring buildings.

Table 9.3: RMS Delay Spread and Variation in Received Power at Numerous Locations in Three Large Office Buildings

Building	Median RMS Delay Spread (nsec)	Maximum RMS Delay Spread (nsec)	No. of Locations
1	94	440	91
2	77	1470	83
3	88	507	61
From Seidel et al. (Op. Cit.).			

The distinction between LOS and non-LOS conditions is not as clear-cut as it may appear. It is apparent from the discussion of section 3 that an interior partition wall made of gypsum wall boards has little effect on signal strength, so that an LOS path may exist between neighboring

rooms for microwaves even though people cannot see through the relevant walls. On the other hand, as the path length increases, the direct ray becomes less dominant and more variation in signal strength and effective delay may be expected, even when an unobstructed path exists between the transmitter and receiver. This effect is demonstrated in Figure 9.20, where received signal strength is shown for measurements performed in a large nominally open area, with a 5m-wide unobstructed path from transmitter to receiver in all cases. The average received power shows deviations of as much as 8 dB from the free-space prediction for the direct ray. At long distances, the average power appears to be increasing over the free-space prediction due to added power from the floor reflection. The minimum signal as a function of frequency is often much lower than the average power, as shown in detail in the inset at right; locations with lower average power have more variations in power with frequency, suggesting that in those locations near cancellation of the direct ray by reflections from nearby obstacles enables rays with much longer path lengths to significantly contribute to the received power. For example, at 9.15 m, variations of as much as 18 dB are seen in signal power versus frequency. Note that the lateral

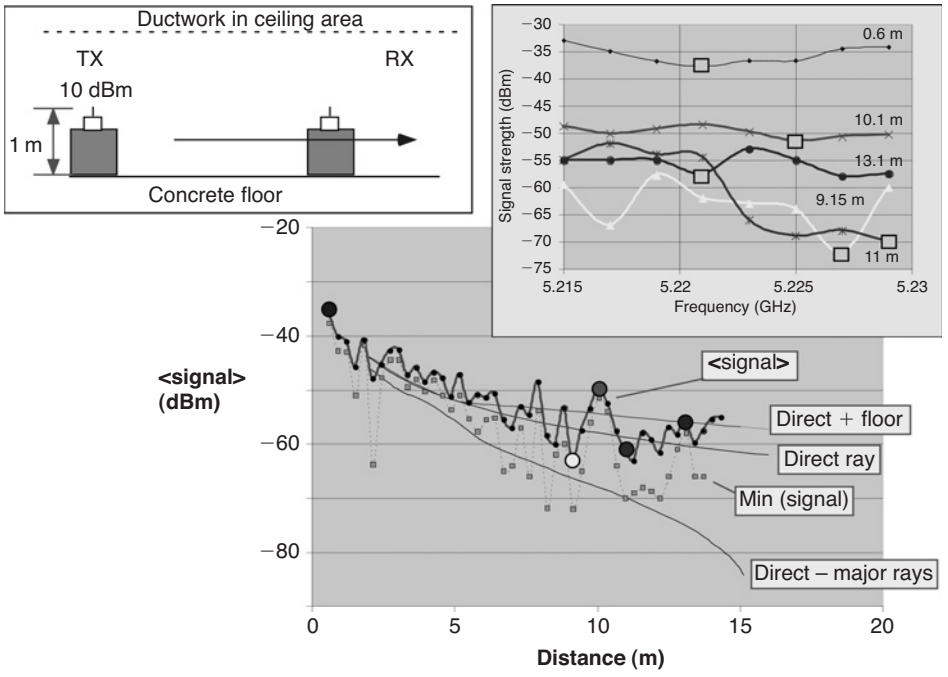


Figure 9.20: Received Signal Power Averaged over 5.215–5.228 GHz and Minimum Signal Power over Frequency vs. Direct Path Distance Compared with Free-Space Propagation of Direct Ray, Sum of Direct and Floor Reflection Power, and Worst-Case Power Where Major Rays (Floor, Walls, Ceiling) Subtract from Direct Ray; Insets Depict Measurement Configuration and Power vs. Frequency at Several Path Lengths (Indicated by Enlarged Symbols in Signal Power Plot)

resolution of the measurements is inadequate to show small-scale fading that may result from, for example, counterpropagating reflected rays from the walls of the room.

The distribution of average signal strength that results from these complex interactions with the indoor environment is shown in Figure 9.21 for a typical small wood-framed residence and in Figure 9.22 for a moderate-sized two-floor tilt-up office building with concrete-over-plywood floors. These data were obtained using an uncalibrated radio card, so relative signal levels may be meaningful but absolute signal power should not be taken too seriously. In the residence data were taken at 1-to 2-m intervals and in the larger building at roughly 5-m intervals; the contours are a smooth interpolation of the measured data points. The reported values at each location average a number of packets, with the measurement position varied over a few tens of centimeters.

Turning first to Figure 9.21, in the region where only one internal wall interposes itself between transmitter and receiver (the horizontal dotted arrow), the signal strength is quite close to that expected in free space. In directions where multiple internal and external stucco walls lie on the direct path (the inclined dashed arrow to bottom right), the detected signal strength is as much as 40 dB below the free space value. If we account for the fact that in this direction the direct path may involve as many as three stucco walls (8 dB each) and four internal partition walls, each involving two layers of dry wall ($8 \times 2 = 16$ dB loss, for a total

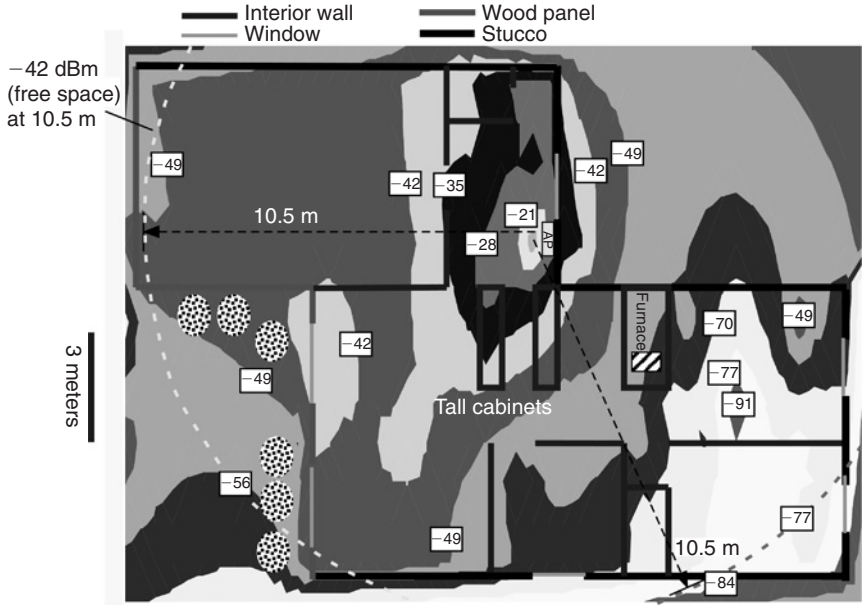


Figure 9.21: Measured Signal Power from Consumer Access Point (Nominal Output ≈ 14 dBm, Marked AP) in Wood-Framed Residence; Boxed Numbers Denote Value at Contour Edge and Dashed Lines Show Circle of Radius 10.5 m from Access Point, at Which Estimated Power for Free Space Propagation Is −42 dBm

of about 40 dB excess loss), simple wall counting leads to results in reasonable agreement with observed signal strength. The deep shadow at −91 dB from a milliwatt (dBm) is plausibly accounted for by a large metal obstacle—a central-heating furnace and associated ductwork. A first approximation to the distribution of signal strength in real complex environments is obtained just by counting losses along the direct path, allowing for diffraction.

Figure 9.22 shows similar data for a larger commercial building. On the upper floor, where the access point is located (at about 2 m height above the floor), signal strength through much of the measured area is remarkably close to that expected for free space propagation (to within the limitations of this simplistic measurement). We can infer that the cloth-and-pressboard cubicles that occupy most of the upper floor constitute little impediment to propagation at 2.4 GHz, despite the sheet-metal shelves they are equipped with. The region at upper right of the upper floor shows reduced signal strength due in part to measurement points located within rooms contained in partition walls (appearing as regional reduction in signal strength

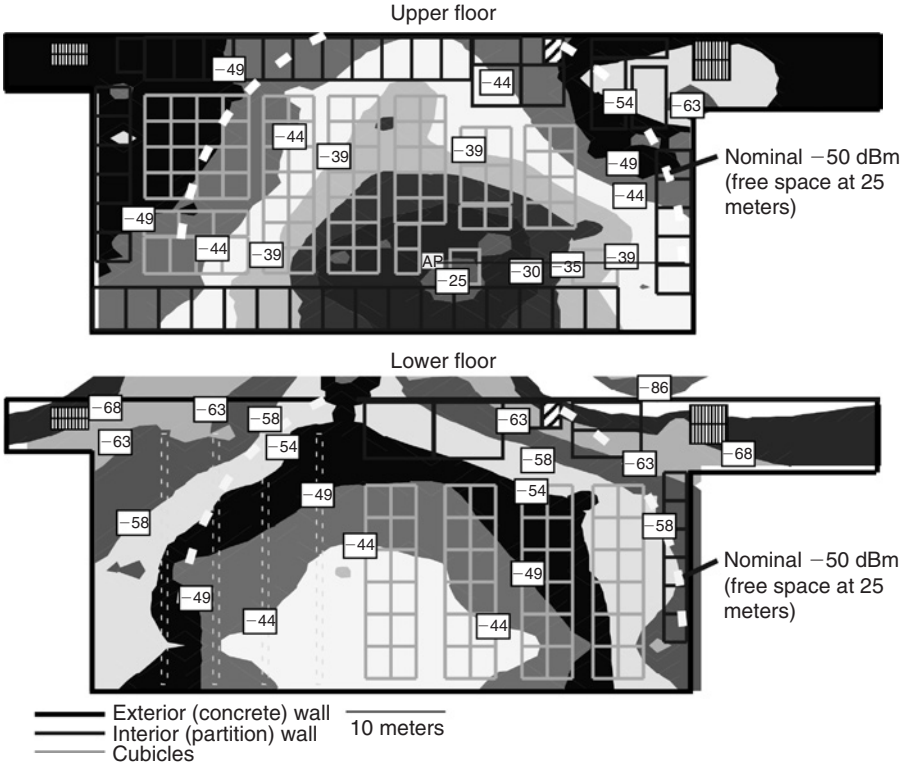


Figure 9.22: Measured Signal Power From Enterprise-Grade Access Point (Nominal Output ≈ 100 dBm, Marked AP) in Tilt-Up Two-Floor Commercial Building; Boxed Numbers Denote Value at Contour Edge and Dashed Lines Show Circle of Radius 25 m From Access Point, at Which Estimated Power for Free Space Propagation Is −50 dBm

on this low-lateral-resolution contour plot) and measurement points at which the direct ray passes through external walls.

The lower floor, which is also occupied by cubicles on the right and only some exposed conduit on the left, is also close to free space once one corrects for a floor loss of about 9 dB (measured separately). There is some suggestion of enhanced signal strength near the lower wall, perhaps due to reflections from the partition walls on the upper floor or the concrete external wall. Signal strength is reduced at the upper right where again interior partition walls and exterior walls interpose themselves onto the direct ray from the access point to the measurement point. The interior concrete shear wall at top left of the lower floor leads to modest additional attenuation. Thus, we can generally conclude that attenuation and obstacles along the direct ray once again provide decent guidance to signal strength. A caution to this sanguine conclusion is the deep minimum (86 dBm) at top right, which is not in line with the elevator shaft (striped box) or any other known metallic obstacle and is a bit of a mystery. (Complete access to all the rooms and interior wall spaces was not available.) Incomplete knowledge of building features, a likely circumstance in many cases, provides a limit on one's ability to model the results: measured data are always needed to complement theory.

Figure 9.23 depicts survey data for a much larger facility (a convention center). Convention halls are typically very large open spaces with concrete walls and floors and ceiling height of

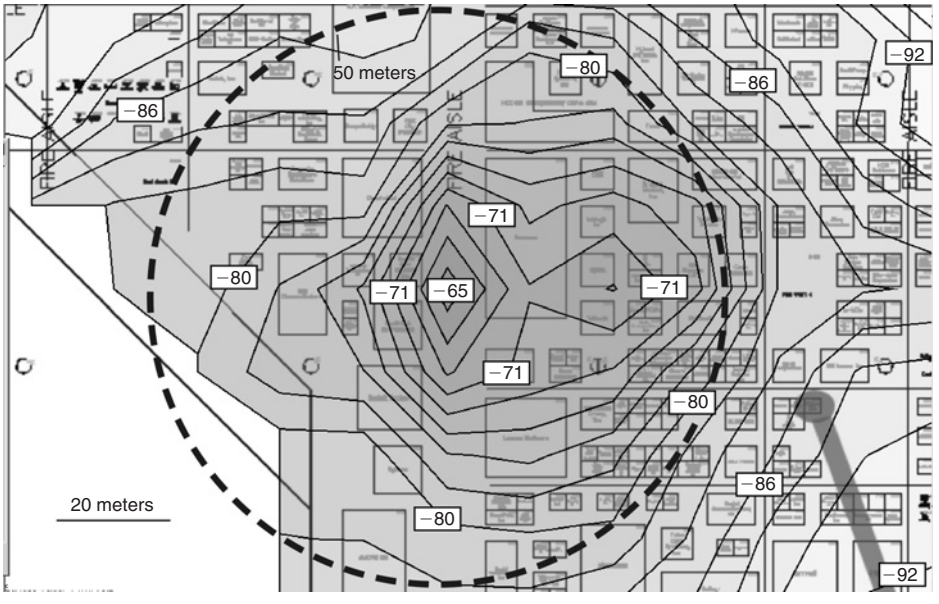


Figure 9.23: Survey of Signal Strength for a Single Access Point on the Trade Show Floor of a Convention Center; Estimated Signal Strength in dBm at Contour Edge Depicted by White Boxes (Image Courtesy of Jussi Kiviniemi, Ekahau Inc.)

more than 10 m. There are few significant obstacles to propagation in most cases within the open area, because display booths are typically formed mostly of plastic panels with plastic or metal rods for mechanical support, and in any case do not exceed a few meters in height. Note, however, that convention halls may be filled at floor level with mobile microwave absorbers—human beings! For good coverage, an access point ought to be placed several meters above floor level to allow the direct rays to clients to pass through only a few people or obstacles.

In this case, the exact location of the access point being monitored is not known and may be on a separate floor from that surveyed, so overly detailed examination of the data is inappropriate. However, it is clear that a region some tens of meters on a side is covered at a very reasonable signal strength of 75 dBm, and signal strength over much of the surveyed area exceeds typical low-rate sensitivity for 802.11 radios and will provide some sort of coverage. Indoor regions of more than a hundred meters in extent can be covered with a single access point.

Before we leave the subject of indoor propagation, it is worthwhile to add a few brief remarks on a subject we have heretofore rigorously ignored: *propagation exponents*. In free space we know that the received power will decrease as $(1/r^2)$, where r is the distance between transmitter and receiver. It would certainly be convenient if we could model indoor propagation using the Friis equation in exactly the same fashion, just using a different value for the exponent of r—that is, it would be nice to say indoors that received power is proportional to $(1/r^n)$ where n is not necessarily 2. A great deal of work in the literature has been devoted to deriving values of n that fit a particular data set. However, wide variations in the value are seen between different researchers and different data sets: values as small as 1.8 and as large as 5 have appeared in the literature.

It is the opinion of this author that such work resembles the old joke about looking for your wedding ring under the lamp post not because you lost it there but because the light is better. It would be nice if indoor data fit a modified exponential model with a substantially invariant value of n, but it doesn't. It's quite possible to see why: the dependence of signal on distance is inextricably tied up with what obstacles are encountered as the receiver moves away from the transmitter and thus uniquely dependent on the physical arrangement and construction of a given facility. Trying to force fit such particularized data sets into a modified distance dependence is like trying to evaluate the probability of checkmate in a chess position based on measuring the distance from your queen to your opponent's king with a ruler: the measurement is perfectly valid and quite irrelevant. Indoor propagation cannot be understood in the absence of knowledge about what is inside the doors. Although it might be possible to specify useful path loss exponents for specific types of room configurations (e.g., drywall partitions spaced 3 m apart with normal incidence), it seems to me much more sensible to start with free space and count walls and other obstacles that subtract from signal strength than to introduce a functional form that has no physical basis.

9.6 How Much Is Enough?

Now that we have examined how much signal power one might find available in various sorts of indoor environments, it behooves us to ask how much do we need? The basic answer is that the amount of signal power needed depends inversely on the data rate that is desired. Recall that the (S/N) needed increases as the number of bits per symbol increase and that the absolute noise level is proportional to the bandwidth and thus increases if we increase the rate at which symbols are transmitted. The lower limit on noise is the thermal noise entering the system, modified by the excess noise of the receiver as determined by its noise figure. For bandwidths on the order of 10 MHz, as used in WLAN systems, the thermal noise is around −104 dBm, and a typical receiver noise figure of 5–8 dB means that the effective noise floor is about −98 dBm for the radio chip. Adding a few decibels to account for board and switch losses, we might expect a noise floor of −95 dBm. To demodulate binary phase-shift keying reliably, we need a S/N of about 9 dB, but recall that for the lowest data rate of 802.11 classic (1 megabit per second [Mbps]), the receiver also benefits from around 10 dB of processing gain, because the true bandwidth of the signal is an order of magnitude less than the received bandwidth. Thus, it seems possible to achieve sensitivities in the mid-90s for 1 Mbps. To demodulate 64QAM, used in the highest rate modes of 802.11a/g, requires an (S/N) of around 26 dB and does not benefit from any spreading gain, so ignoring coding gain we might expect sensitivities of $(-95 + 26) \approx -70$ dBm at the highest rate of 54 Mbps.

Table 9.3 provides published sensitivities for a selection of commercial 802.11 radios, collected through early 2004. The results are in rather good agreement with the primitive estimates of the preceding paragraph. At low rates, the best sensitivity is −94 dBm and −91 dBm is a typical number, whereas at 54 Mbps sensitivity of −65 dBm is typical. Figure 9.24 shows actual measured data rate versus estimated received signal power for a number of commercial 802.11 radios. (The data rates reported in Figure 9.24 are those delivered from one TCP client to another across the wireless link. TCP, *transmission control protocol*, is the protocol used in Internet data transfers and delivers packets across a data link using *Internet protocol*, which in turn communicates with the Ethernet drivers that ship bits into the 802.11 clients. Thus, the reported rate includes delays due to overhead within the 802.11 packets and the Ethernet transport mechanism and reflects the true rate at which data can be moved across the link, very roughly 60% of the nominal peak rate for a given 802.11 transport mode.) The signal strength here is estimated from single-frequency measurements at corresponding positions and should be presumed accurate to no better than ±3 dB. The data rates behave more or less as one would expect from the sensitivities: orthogonal frequency-division multiplexing data rates begin to fall for signal power less than −65 dBm or so as slower modulations must be used, and rates extrapolate to zero at received powers in the mid-90s.

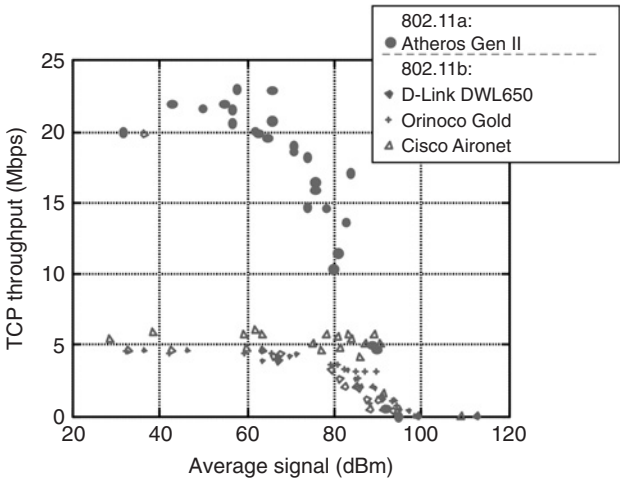


Figure 9.24: TCP Throughput vs. Signal Power (Estimated From CW Measurements) for Several Commercial 802.11 Radios (Slightly Edited Version of Data Published in "Correlating Link Loss With Data Throughput," Dobkin, High Frequency Electronics, January 2003, p. 22, by Permission of WJ Communications)

Reported data and measurements support the rule of thumb advanced by Jesse Frankel of AirMagnet that a received signal strength of about −75 dBm should be adequate to provide high throughput from most commercial 802.11b equipment, with sufficient margin (5–10 dB) to accommodate typical local fading and random variations in signal power. More aggressive guidelines can be used if control over the client population is available. By reference to Table 9.4 one finds that the reported sensitivities vary by as much as 7 dB, so the use of only high-quality clients and access points should allow perhaps another 5 dB of margin. However, system administrators wishing to follow such a path should recall that users are typically enamored of the client they have and would much prefer having it work adequately at their preferred locations than switching to a different client (which after all may create hardware and/or software problems for them). Using these guidelines, we can infer from Figures 9.21 to 9.23 that a single access point with typical output power around 14–16 dBm can be reasonably expected to cover a small residence and that a 100-mW (20 dBm) access point can cover a 25-m radius in an open industrial facility or perhaps 50 m in a large open room. Coverage is strongly affected by obstacles encountered by the direct ray, as noted in the previous section. In the sort of open-area cubicle environment often encountered in modern commercial facilities, the access point should be located high enough from the floor to avoid most obstacles and partitions for optimal coverage.

Note also that providing ample power at the corners of an open room with exterior walls using a high-powered centrally located access point will also probably result in significant leakage to the outside world (e.g., see Figure 9.21): if security is an important issue, the use

Table 9.4: Published Sensitivity for Various Commercial 802.11 Radios

Radio	Sensitivity at 54 Mbps (dBm)	Sensitivity at 11 Mbps (dBm)	Sensitivity at 1 Mbps (dBm)
Orinoco AP/Gold NIC card		−82	−94
Cisco Aironet 350		−85	−94
D-Link DWL900 AP		−79	−89
D-Link DWL650 NIC		−84	−90
D-Link DCF650 cf		−80	−88
D-Link DWLAG650	−73	−91	−94
Proxim 8550 AP		−83	−91
Surf'n'sip EL2511		−87	−95
Bewan USB11		−80	−88
Bewan 54 G	−65	−80	
Trendware TEW403	−65	−80	
Senao 2511		−83	−91
Eazix EZWFDU01		−85	−93
Eazix EZWFM5-02	−65	−80	−87
Summary	−67±4	−83±3	−91±3

of multiple low-power access points, preferably with inward-pointing directional antennas, is preferred. Such inward-looking access points mounted at the corners of a building, combined with a few centrally located access points, can achieve excellent high-rate coverage in the interior of a large building while keeping exterior leakage very small. Radiating cable can be used to provide supplementary coverage in corridors near exterior walls and for partitioned offices whose walls will cumulatively block coverage from the corner-mounted access points. (Even better security, as well as esthetic and environmental benefits, result from providing extensive evergreen foliage in the areas surrounding the building in question, so that would-be eavesdroppers in distant locations have no clear line of sight.

9.7 Indoor Interferers

9.7.1 Microwave Ovens

WLANs and wireless personal area networks mostly operate in the unlicensed bands around 2.4 and 5 GHz. In the United States, radios operating in these unlicensed bands are obligated under Federal Communications Commission (FCC) regulations to accept any interference they encounter, but the users don't have to be happy about the results. What are the likely sources of such interference?

High on everyone's list of suspects is the humble microwave oven. It is remarkably unlikely that microwave ovens will start to use other frequencies than 2.4 GHz, given the huge installed base, vast experience, and the cost-driven nature of consumer markets. Therefore, WLAN users who choose to operate at 2.4 GHz must live with them when they are present. Note that in the United States the FCC does not regulate microwave ovens; they are, however, subject to Food and Drug Administration restrictions that require less than 1 mW/cm^2 of equivalent radiated power be detected at 5 cm distance from any location on the oven surface.

Averaged spectra and time-dependent spectra for a representative microwave oven interferer are shown in Figure 9.25. Emission peaks at around 2440 MHz, but it is apparent that significant emission is present across about half of the ISM band. Emission is seen to be sporadic in time. The likely explanation is that the oven RF power is actually on for roughly half of each AC cycle of 16.7 msec (making for a simpler power supply). At a scan rate of 500 ms/scan for 500 MHz, the spectrum analyzer spends only 1 msec in each 1-MHz region, and thus on any given scan the analyzer may record peak RF output, reduced power, or nothing at all, even though the average oven peak power is constant from cycle to cycle.

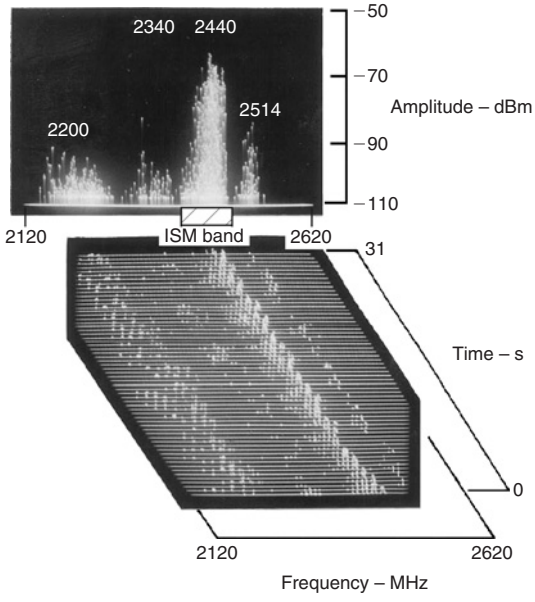


Figure 9.25: Emission Spectrum Due to a Microwave Oven Averaged Over Time (Top) and vs. Time (Bottom, 0.5-sec Scans) (From "Literature Search and Review of Radio Noise and Its Impact on Wireless Communications," Richard Adler, Wilbur Vincent, George Munsch, and David Middleton, US Naval Postgraduate School, For FCC TAC, 2002)

Figure 9.26 shows measured power emitted from a number of standard commercial microwave ovens at distances from 1 to 5.5 m, corrected to a standard distance of 5 m assuming free-space

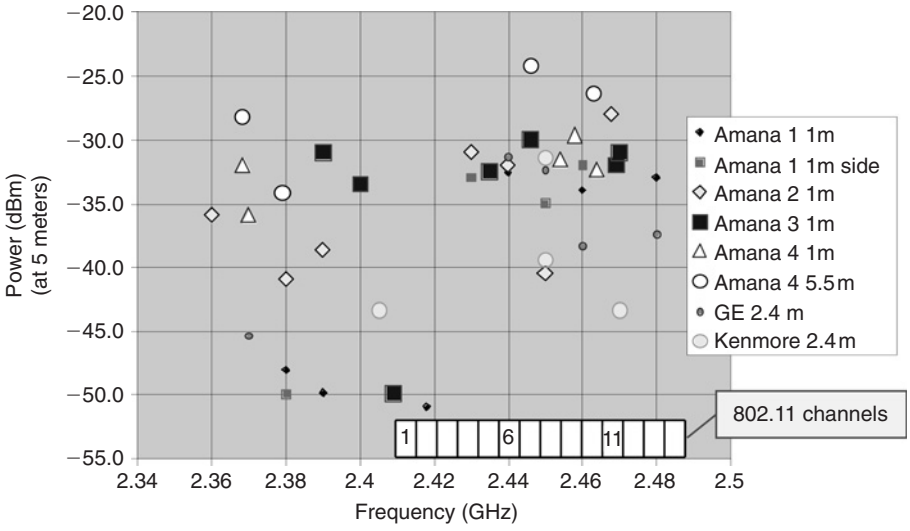


Figure 9.26: Measured Power at Selected Frequencies for Six Conventional Microwave Ovens; Measurements at 1–5.5 m, Using 2.4-GHz Quarter-Wave Monopole 1 m from Floor, Acquired with a Spectrum Analyzer in Peak-Hold Mode, Corrected to 5 m Assuming Free-Space Propagation

propagation. As can be seen in Figure 9.25, the emission spectrum of a microwave oven is complex and contains numerous peaks; the data in Figure 9.26 include only a few of the highest peaks for each oven. Note that different ovens do not have the same power spectrum, though all share the same general features of strong emission around 2.45 GHz. These data also show much stronger emission around 2.35 GHz than seen in Figure 9.25 for several ovens (though that is outside the nominal ISM band and hopefully of little import to 802.11 users).

Institute of Electrical and Electronic Engineers (IEEE) 802.11 WLANs divide the ISM band into 14 5-MHz channels centered from 2.412 to 2.484 GHz, though only channels 1, 6, and 11 are normally available as nonoverlapping channels in the United States. From Figure 9.26 it is clear that most ovens will have a modest effect on the low-numbered channels but will interfere significantly with channels 6 through 11. Fortunately, the sporadic nature of microwave oven interference suggests that only modest effects will be felt for short packets. This supposition was confirmed in the work of Chang and coworkers, who found that the same microwave oven interference that caused packet error rates of 30–40% in 400-byte packets for distances up to 20 m from the oven caused less than 5% packet errors in 100-, 200-, and 300-byte packets even when the receiver was only 4 m from the oven. Chang et al. describe their equipment as 802.11 direct-sequence spread spectrum (DSSS); if we assume they were using an 802.11 classic link at 1 Mbps, a 400-byte packet is roughly 3.3 msec (the exact value is unclear, because the authors have not recorded whether the 400-byte total covers only data

or includes headers). It is difficult to reconcile the measured 50% duty cycle of oven radiated power (e.g., see Unawong et al.) with the sudden threshold for bad packets they found above 300 bytes, but we can still qualitatively conclude that short packets have a good chance of finding their way around oven interference pulses. Because much higher data rates use shorter packets, it is reasonable to expect that their resilience to oven interference will be good. For example, at 11 Mbps, a maximum-length 1500-byte packet will be only a bit more than 1 msec in duration; at least half of the packets will encounter no interference. Note that when interference of this type is limiting performance, backing off to lower rates may actually increase packet error rate. Most control interfaces provide a mechanism for the user to adjust the threshold for packet fragmentation (sometimes known as interference robustness), which is the proper response to problems with oven interference.

Chang and coworkers report high bit error rates out to about 22 m from the microwave oven for 400-byte packets, with the radius of disturbance shrinking to about 17 m when a concrete wall was interposed between the oven and the receiver. Unfortunately, they provide no information on transmitter position or received signal strength, so we cannot generalize their results. As a rule of thumb, access points should be positioned so that locations of frequent usage are significantly closer to the nearest access point than to the nearest microwave oven, and ovens should be sequestered behind partition walls (concrete if possible) to further reduce their interference potential. Because the peak received power of 30 dBm or so at 5 m is quite comparable with that of an access point at a similar distance, interference will likely be serious if the oven is closer to the user than the desired access point and may still represent a limitation on high-rate data even when relatively distant. Finally, in view of the typical usage pattern for most microwave ovens in industrial locations, users stuck near an oven should consider breaking for lunch at lunch time.

9.7.2 Other WLAN Devices

Multiple access points within a single building or single floor of a large building are likely to provide overlapping regions of coverage. If all the access points are under the control of a single organization, their channels can be arranged so as to reduce interference. In the U.S. ISM band, as noted previously, only three nonoverlapping channels exist (1, 6, and 11): assuming good-quality radios, access points on distinct nonoverlapping channels should not interfere with each other. However, as previously noted, at least four channels are needed to provide a robust tiling of the plane with no adjacent interferers, and seven or more are preferred. The situation is worse if multiple-floor buildings with penetrable floors are present. To design a network that is truly interference free while providing complete coverage, it is desirable to operate in the UNII band where an adequate set of nonoverlapping channels is available.

The potential achievability of noninterfering channel allocations is somewhat irrelevant when overlapping access points are not under the control of a single organization, as is likely to be the case in multitenant facilities, particularly when the tenants are separated only by interior

partition walls. Fortunately, 802.11, like other Ethernet-style medium access controls (MACs), is robust to collisions and interference as long as loading is not too heavy. In general, both MACs will detect a packet failure and randomly back off for retransmission, resulting in a modest degradation in performance as long as the collisions are infrequent. However, Armour and coworkers showed that the HiperLAN MAC suffers more degradation in performance from an uncontrolled neighboring access point on the same frequency than does carrier-sense multiple access with collision avoidance (CSMA/CA), because HiperLAN does not expect unscheduled transmissions to overlap its assigned time slots.

Even if interference between neighboring access points is eliminated, the shared-medium nature of wireless links ensures the potential of cochannel interference between clients sharing the same access point. Interference between multiple clients is handled by the MAC layer of the protocol; in 802.11, the CSMA/CA MAC treats interference between two users as a dropped packet if it leads to a packet error and resends the packet after random backoff. In practice, this approach provides for up to five to seven simultaneous users, each receiving around 1 Mbps of throughput, in a single 802.11b cell. Because typical users are only occasionally transferring network data, up to 20 workstations or potential users can be accommodated in one access point. In HiperLAN and other centrally coordinated schemes, negligible interference between users in the same domain occurs because each is allocated transmission times.

As with microwave ovens, implementation of *request to send/clear to send* and packet fragmentation will increase robustness to interference from other clients and access points at the cost of reduced peak throughput. Another practical consequence of multiple access points at similar power levels is changes in association state of a client if it is set to associate with the strongest access point signal. Such "flipping" can be very irritating for the user but can be dealt with by requiring association to only the desired BSSID.

9.7.3 Bluetooth vs. Wi-Fi

Bluetooth devices transmit in 75 1-MHz channels, randomly selected 1600 times per second. Thus, in any given millisecond, collocated Wi-Fi or 802.11 g and Bluetooth devices can potentially interfere. The extent of the interference is likely to depend both on the ratio of the desired signal (Wi-Fi or Bluetooth depending on whose viewpoint one takes) to the interferer (BT or Wi-Fi): the *signal-to-interference ratio* (S/I). The (S/I) is in turn determined by two path losses: that from the wanted transmitter to the receiver and that from the interferer to the victim receiver. Interference is a bigger problem when the victim receiver is at the edge of its range. Finally, because Bluetooth is a frequency-hopping protocol and Wi-Fi is not, the frequency separation between channels is another variable; when the separation greatly exceeds the width of the Wi-Fi channel, little interference is likely in either direction.

Model results, reviewed by Shelhammer, show that the interference effect on Bluetooth devices of an interfering 802.11b device is mostly confined to the Bluetooth slots within about 6 MHz of

the Wi-Fi channel. Because there are 75 slots (in the United States) that can be used, this means that about (11/75) = 15% of slots are subject to interference. Bluetooth packets are generally (though not always) contained within a single frequency slot, so one can roughly say that the Bluetooth packet loss rate will not exceed 15% even when severe Wi-Fi interference is present. A 10–15% packet error rate has modest impact on data communications but will lead to noticeable degradation in voice quality. Proposed enhancements to the Bluetooth link standards, which provide simple retransmission capability, can nearly eliminate packet loss, but existing Bluetooth devices (of which tens of millions have been shipped) will not benefit from these improvements. Because of the robust Gaussian minimum-shift keying modulation used in Bluetooth, even the central frequency slots with maximum overlap with Wi-Fi will achieve bit error rates of less than 10^{-3} if (S/I) $>$ 5 dB; for a Bluetooth master and slave separated by 1 m, a Wi-Fi client device more than 4 m away from the victim receiver will have modest impact on the Bluetooth device. Note that if the Wi-Fi device in question is lightly used (as most are most of the time), effects will be proportionately smaller. These theoretical results were generally confirmed by the measurements of Karjalainen and coworkers. However, they found that much more severe effects could result if the WLAN and Bluetooth devices were placed together (as if on a single laptop device); in this case, the Bluetooth throughput could be degraded by as much as 80%, presumably due to the strong coupling of even the edges of the WLAN spectrum into the Bluetooth receiver.

Changing sides in the debate to become a Wi-Fi proponent, we find that the worst effects are again confined to Bluetooth slots within about 6 MHz of the Wi-Fi center frequency. At 1 Mbps, a Wi-Fi link can tolerate an (S/I) level of −3 dB (i.e., the Bluetooth signal is 3 dB *larger* than the Wi-Fi signal) before severe degradation of the bit error rate occurs because of the processing gain of the Barker code, even for the worst Bluetooth channel (which in this case is 1 MHz from the Wi-Fi center frequency). An 11-Mbps complementary code keying link, which has some coding gain but no processing gain, requires (S/I) $>$ +4 dB for the worst Bluetooth channel to ensure unaffected performance (in this case the 0-MHz channel, because the complementary code keying spectrum is a bit different from the Barker spectrum). A Bluetooth client more than about 4 m away from a Wi-Fi device with good (S/N) will have little effect on the latter, though the potential for interference increases when the Wi-Fi device nears the end of its range. Note that just as in the microwave oven case, a Wi-Fi device operating in a high-rate mode will suffer *less* from the interferer than at 1 Mbps, because the shorter packets are less likely to encounter an interfering Bluetooth slot. An 11-Mbps link will only suffer packet loss of around 4% even in the presence of an adjacent Bluetooth device, whereas the same situation at 1 Mbps produces packet loss of more than 50%.

Soltanian and coworkers showed in modeling that adaptive notch filters could be used to remove Bluetooth interference from a direct-sequence 802.11 signal. Orthogonal frequency-division multiplexing packets, used in 802.11 g, are potentially robust to narrow-channel interference, because it is in principle possible for the decoder to recognize and discard contaminated subcarriers. Doufexi and coworkers showed that with such adaptation, the impact of a Bluetooth

interferer is almost negligible even at (S/I) of −11 dB and quite modest (S/N) ratios. However, because "g" radios must be capable of using direct-sequence preambles audible to older radios, problems with interference would persist. In any case, these advanced capabilities have not yet been incorporated into commercial radios as far as the author can ascertain.

Wi-Fi–Bluetooth results are summarized in Table 9.5. The guard radius (the distance at which the interferer has little effect on the victim receiver) therein is only a guideline; for a specific physical arrangement, link loss for both transmitters and thus the (S/I) must be estimated.

Table 9.5: Wi-Fi and Bluetooth as Interferers

Inteferer	Victim	Minimum (S/I) for Negligible Interference	Worst-Case Packet Loss	Guard Radius for Negligible Interference
Wi-Fi	Bluetooth	+5 dB	15%	4 m
Bluetooth	Wi-Fi 1 Mbps	−3 dB	50%	2.5 m
Bluetooth	Wi-Fi 11 Mbps	+3 dB	4%	4 m
From Shelhammer, Figures 14-1–14-6.				

9.7.4 Cordless Phones

Cordless phones are an extremely popular convenience (though the *in*convenience of a misplaced ringing cordless phone sometimes seems to exceed the benefits!) in homes and (somewhat less frequently) in office environments. Most purchasers appear quite comfortable with obtaining handsets and base stations from the same manufacturer, and there is little commercial pressure for interoperability, so cordless phones are generally unique proprietary designs subject only to compliance with regulatory requirements. As a consequence, both frequency-hopping and direct-sequence designs are used. Output power and signal bandwidth also vary noticeably, though the signals are quite narrowband from the view of an 802.11-type network, appropriate to the modest data rate (roughly 60 Kbps if uncompressed, as little as 10 Kbps if compression is used) needed for voice. A few representative examples are summarized in Table 9.6. The output power of most of the phones is comparable with or larger than typical consumer (30–50 mW) or enterprise (100 mW) access points.

Table 9.6: A Few Representative Cordless Phone Schemes

Vendor	Scheme	Power (mW)	Bandwidth (MHz)
Uniden	DSSS	10	2.8
V-Tech	FHSS	60	1
Sharp	DSSS	100	1.8
Siemens	FHSS	200	0.8
Source: FCC web site.			

The emissions spectra of portable phones as interferers are shown in Figure 9.27. The received power from these nearby interferers is as large as 30 dBm. The time-dependent results show that the interference is sporadic, so that even if the direct-sequence phone lies on top of the wanted Wi-Fi channel, data transfer will still occur during idle times, but peak data rates will be significantly reduced when the phones are active. Phones are also likely to be used more extensively than microwave ovens in most environments (save perhaps fast-food restaurants),

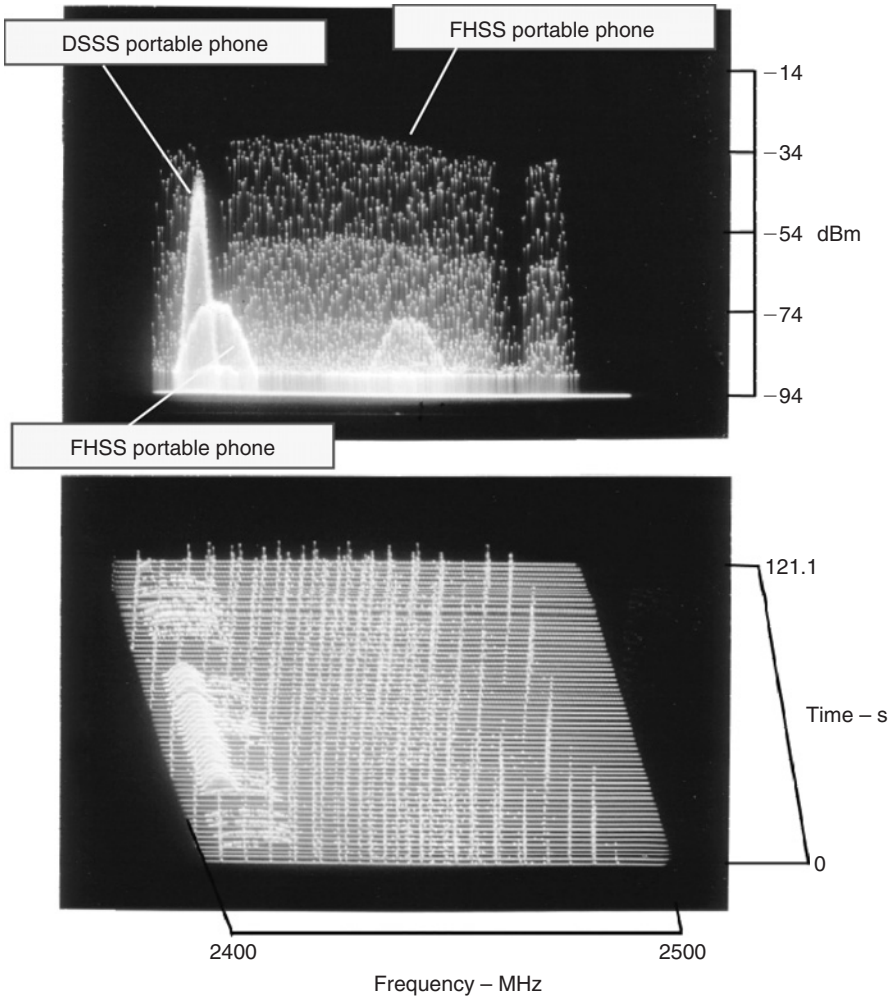


Figure 9.27: Spectra of a DSSS and Two FHSS Portable Phones Acting as Interferers: (Top) Averaged Power; (Bottom) Time-Dependent Power (Home Office, Davis, CA, 2002; Monopole Antenna ⇒ High Dynamic Range Preamplifier) (From *Literature Search and Review of Radio Noise and Its Impact on Wireless Communications*, Richard Adler, Wilbur Vincent, George Munsch, and David Middleton, U.S. Naval Postgraduate School, for FCC TAC, 2002)

so they constitute an important source of interference for Wi-Fi. However, the narrowness of the DSSS signals in frequency (so that only a few Bluetooth hops would be affected) and the minimal overlap of two random frequency-hopping spread spectrum (FHSS) patterns imply that cordless phones will not be significant interferers for Bluetooth links.

Because cordless phones are after all mobile, unlike most microwave ovens, it is generally impractical to sequester them behind absorbing walls or relegate them to locations more distant from the relevant access point. FHSS phone interferers will resemble Bluetooth interferers, sporadically causing packet errors as they hop onto the Wi-Fi channel, with the effect being less serious for higher data rate packets. The higher power of the phones makes the guard radius much larger than that cited in Table 9.5 for Bluetooth. In simple environments with only one or two access points and one or two DSSS phones (assuming they are under control of one person or organization), it is possible to adjust the Wi-Fi channels and the phone channels to avoid overlap. Phone operating manuals don't always provide the channel frequencies, but these can be found (with some labor) in the FCC compliance reports. For more complex environments with numerous phones and access points, DSSS phones can be generally set around Wi-Fi channels 3 or 4 and 8 or 9 to minimize overlap with Wi-Fi devices at 1, 6, or 11.

In the case of a complex system in which the intermingling of numerous cordless phones and WLAN devices is inevitable, two alternative courses of action still remain. Wi-Fi–based portable phones, using *voice-over-Internet-protocol*, are becoming widely available. Although these phones are Wi-Fi devices and thus will compete for the same medium as data, the CSMA/CA MAC ensures that coexistence will be more graceful than that of an uncoordinated proprietary interferer. The second approach is to relocate the data network to the 5-GHz band, where very few cordless phones operate today. The UNII band provides more bandwidth and channels and is not subject to interference from microwave ovens or Bluetooth devices. For mission-critical data applications in complex coverage networks subject to unavoidable interferers, 802.11a or dual-band networks should be considered.

9.8 Tools for Indoor Networks

9.8.1 Indoor Toolbox

Constructing and managing an indoor WLAN involves a number of tasks related to signals and propagation. Most networks use a number of fixed access points to provide service to a specific area. Planning the locations of the access points to ensure good coverage and minimize cost requires some thought be given to the unique propagation characteristics of the building or campus in question. A first estimate can be derived using building plans and the information presented previously in this chapter, but for complex installations, automated modeling may be helpful and surveying is essential. In multitenant environments, an examination of the interference environment, due to both other access points and clients

and non-LAN interferers, may be needed to understand what quality of service is achievable. Once the network is installed, ongoing monitoring may be helpful both to ensure that coverage remains acceptable despite the inevitable changes in the propagation environment as people, equipment, and partition walls move and change and to protect against new interferers and "rogue" access points. (Rogue access points are access points connected to the wired network without approval of the corporate or facility authority. Because of the Ethernet-based transparency of the 802.11 standard and the fact that most organizations assume physical security for their internal network and impose little additional access control on its ports, it is easy for an employee or occupant to simply plug a consumer access point into the internal Ethernet network, possibly without even activating Wired Equivalent Privacy encryption. Such an access point provides an open opportunity for any nearby person equipped with a client card to obtain Internet access through the wired network, possibly without the approval of the owners of the network. Many wired networks will also allow a client to see the list of servers available on the LAN with no authentication of any kind, an invitation for industrial espionage if the client is capable of sniffing passwords or other attacks. Thus, rogue access points represent significant security concerns.)

Commercial organizations and software change rapidly, particularly in developing fields like wireless networks, and any specific listings and advice provided in a book of this nature are likely to be out of date by the time the reader could take advantage of them. Therefore, with respect to such tools, it is the intention of this section merely to provide examples of equipment and software capabilities that are available at the time of this writing, so that the reader may be apprised of the sort of resources that might be helpful. Nothing herein should be taken as either a review or an endorsement of any particular hardware tool or software package.

9.8.2 Surveying

The purpose of surveying is to assess received signal strength as a function of location within (and perhaps beyond) the intended coverage area. Many standards, including the 802.11 standards, require that compliant radios provide some measure of the average power of each received packet (typically known as the received signal strength indication [RSSI]); in consequence, any compliant 802.11 radio can be used as a survey tool. The author has verified that RSSI results for at least a couple of commercial client cards are well correlated with received power of continuous-wave signals measured using a network analyzer; RSSI provides a useful semiquantitative indication of signal strength.

The simplest approach to surveying is thus to use any portable device equipped with a client card and software that displays RSSI. Many card vendors provide management utilities that offer access to received signal strength. There are also numerous public-domain and shareware utilities that provide lists of received access point and client signals, including

recent and average signal power: Netstumbler (Windows OS, www.netstumbler.org), Kismet (Linux, www.kismetwireless.net), and Kismac (Mac OS, www.binaervarianz.de/projekte/programmieren/kismac/) are some examples. A screen shot of a Kismac scan output is depicted in Figure 9.28. Simple surveying can be accomplished by manually recording received power on a facility map. For small facilities and simple installations, such an approach is quite adequate (and low in cost).

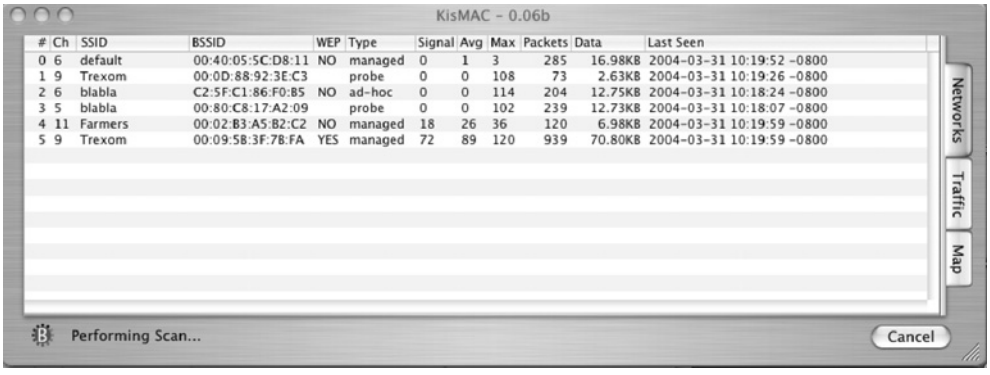


Figure 9.28: Example of Stumbler Application Scan Result, Showing Most Recent and Average Signal Strength Indicators for Each Access Point Received

For more complex installations, involving multiple access points and large multifloor buildings, manual surveys become impractically laborious. Fortunately, software tools are available to automate the process. Current tools allow the import of a graphic background (typically a building plan) on which the user can indicate a starting and stopping point by mouse clicks; data are then acquired at regular intervals as the user walks from the starting to the stopping point, and with the presumption that the walk was performed at a constant rate each data point may be assigned to a location. A thorough survey of a large facility can thus be acquired rapidly without sacrificing the flexibility of handheld portable devices for data acquisition. Surveying tools of this type are available from (at least) ABP Systems, Ekahau Inc., AirMagnet, and Berkeley Varitronics as of this writing. Berkeley Varitronics provides a custom personal digital assistant with calibrated Wi-Fi receiver, so that the absolute signal power is available, whereas the others run as software programs on a Wi-Fi–equipped portable device and are limited to RSSI inputs. Figure 9.29 shows a map of signal strength obtained using Ekahau Site Survey 2.0, superimposed on a building plan and record of the walking path used in acquiring the survey. Maps of other parameters of interest, such as data rate, coverage, S/N or S/I, and strongest access point in each location, are also available. Similar maps and features are available from the other vendors. These tools are much faster than manual surveying for a large facility!

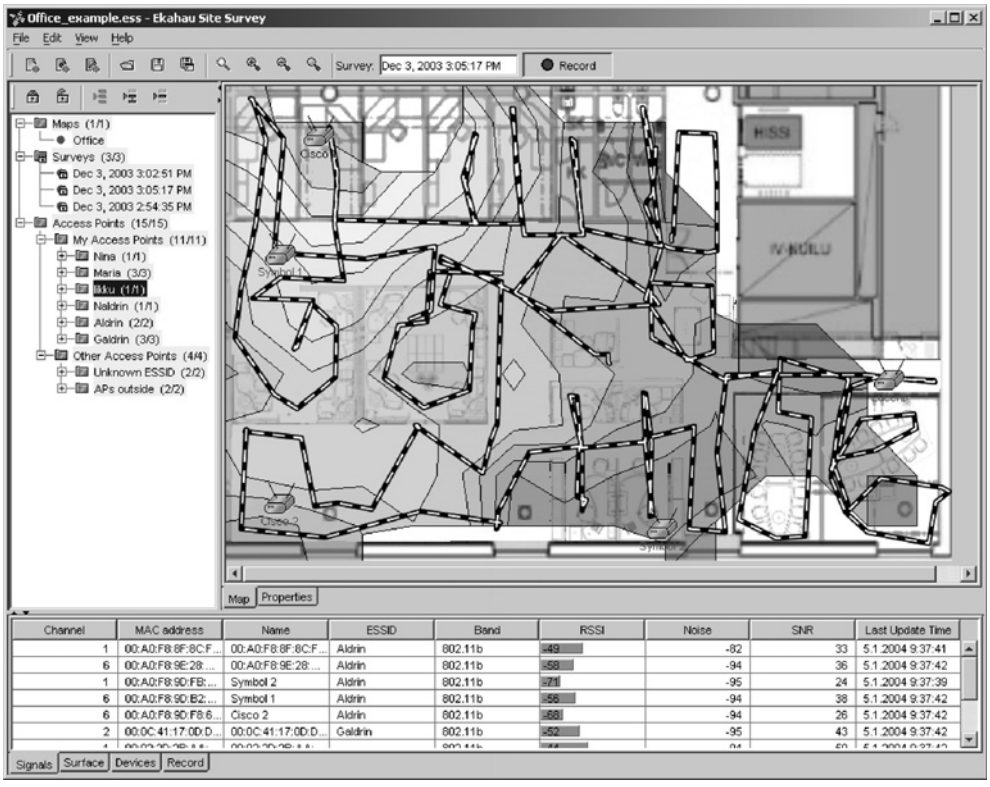


Figure 9.29: Survey Result for Signal Strength for a Specific Access Point Superimposed on Facility Map; Dashed Lines Indicate Walking Paths Used in Survey; File Management and Access Point List Windows Are Also Shown (Image Courtesy Ekahau Inc.)

In addition to the hardware and software needed on the receiving end, when surveying before installation the toolkit should include at least two portable easily installed access points so that candidate locations can be readily tested; two are recommended to allow for real-time examination of the overlap region between each pair. Another useful addition to the toolkit for large facilities is a tall portable support pole for raising a candidate access point close to the facility ceiling when installation there is likely.

9.8.3 Indoor Propagation Modeling

From the discussion presented in this chapter, the reader will have gathered that the first step in understanding propagation in a building is to examine the building layout and understand the obstacles to propagation that are present. For any given proposed access point location, a ruler placed over the map will allow a quick if rough evaluation of the likely signal strength, given that the approximate nature of the walls and fixed obstacles is known.

Automated modeling allows the planner to deal with larger more complex installations. Modeling uses ray tracing techniques; rays are launched at a variety of angles, and the obstacles the rays encounter result in the launch of new rays (e.g., due to reflection) and changes in the ray amplitude (due to absorption and partial transmission). The net phase accumulation along the ray can be tracked to attempt to reconstruct the actual received signal amplitude, though as we have noted it is very difficult in practice to know geometries to such precision as to make relative phase data meaningful. In most cases, the relative amplitude of the important rays is sufficient to establish average signal strength within a region. Phase accumulation is of course equivalent to propagation delay, so that power/delay profiles like Figure 9.19 can also be estimated to ensure that they do not exceed the tolerance of the relevant modulation for multipath delay.

The algorithmic art of a successful ray-tracing problem for complex environments is in the handling of the interaction of obstacles and rays to minimize the computational burden. The problems encountered are similar to those dealt with in rendering software used to produce computer-graphics images. Considerable ingenuity goes into establishing which obstacles and portions of obstacles are visible after a certain number of reflections and avoiding computation along paths that will never contribute to the received signal. A key choice in determining the size of the problem is the number of obstacle interactions to be evaluated; if one chooses (as we have for the manual approach) to model only absorption along the direct ray, the computational burden is much reduced, and qualitative features are still correctly reproduced in most cases. As with most computational modeling, the problem size increases drastically in progressing from two-dimensional to three-dimensional simulations, and many of the tricks used in other fields to reduce computational time cannot be applied here, because buildings generally do not have axes of symmetry.

Diffraction around obstacles is often treated in ray-tracing contexts using the *Uniform Theory of Diffraction* or the *Generalized Theory of Diffraction*. These methods attempt to account for the effect of an obstacle through the definition of effective rays emitted at the edges of the obstacle, where the amplitude and phase of these rays are determined by a fit to Fresnel integrals. Such look-up–based approaches are much more computationally efficient than a complex integration when a large number of irregularly shaped obstacles must be handled.

Commercial software tools specific to indoor propagation for WLAN systems are available from such vendors as Wireless Valley, Wibhu Technologies, ABP Systems, and Ekahau, among others. Most systems can take a building plan graphic as input but require some manual intervention to mark features as walls of a particular type, so that the appropriate RF properties can be assigned. Multifloor buildings can be simulated. Results generally include maps of signal strength and related properties such as signal to noise and interference. Some vendors have constructed databases of commercial access points and clients and thus can provide expected data rate and throughput for specified clients as a function of location.

9.9 Summary

Buildings in the modern world use a modest repertoire of structural materials, of which steel, concrete, wood, glass, gypsum, and bricks and masonry are most frequently encountered. Small buildings use load-bearing walls of concrete, reinforced masonry, or (for residences) plywood and non–load-bearing partitions of stud-reinforced gypsum. Large buildings rely on steel or reinforced concrete framing and do not use load-bearing walls. Floors may be made of wood or concrete on either wood or steel.

At microwave frequencies, typical thicknesses of gypsum, glass, and wood induce only modest absorption. Brick and masonry absorb noticeably. Concrete properties vary depending on mixture and cure state, but thick concrete walls will usually represent significant absorbers. Most common structural materials have refractive indices between 2 and 3 and so reflect significantly, though consequent transmission loss is modest except at glancing angles. Large metal obstacles, including both elements of structures such as columns or ducts and interior objects such as cabinets or server racks, cast shadows 10–20 dB deep lying within their geometric shadow. Floors of concrete on steel act as shields. With these observations in hand, one can construct a first approximation to the average received power at an indoor location by examining the absorption of the direct ray path from transmitter to receiver as it traverses walls, floors, and windows.

Real indoor structures show widely varying propagation characteristics that depend critically on what is in the way. Long RMS propagation delays, averaging around 100 nsec and reaching as high as 1000 nsec when exterior reflections are significant, occur because of reflections from interior structures, even when the direct-line distance from transmitter to receiver is small. Open areas have average signal strength close to that obtained in free space (though local fading occurs because of interactions with reflected rays); a different location at the same distance from the transmitter may have 40 dB less signal power due to the interposition of interior and exterior walls. For typical clients, average signal strength of -75 dBm at the receiver is sufficient to provide reliable high-rate service. If access points can be positioned to avoid most obstacles, interior areas approaching 100 m in span can be covered with a single access point; in the presence of multiple walls, the same access point can be limited to 10 or 15 m.

Common indoor interferers include microwave ovens, cordless phones, and rival Wi-Fi or Bluetooth devices. In each case, the time-dependent nature of the interference ensures that some traffic is likely to get through even in the worst cases, but physical configuration and channel planning can also help to achieve reliable high data rates in the presence of interferers.

Software and hardware tools are available for modeling and measuring propagation characteristics in indoor environments; use of these tools facilitates installation of complex networks and improves the utility of the result.

Further Reading

Building Construction Techniques

Building Construction Illustrated (3rd Edition), Francis Ching with Cassandra Adams, Wiley, 2000: Mr. Perry's favorite pictorial reference; a beautifully illustrated book.

Microwave Properties of Construction Materials

NISTIR 6055: NIST Construction Automation Program Report No. 3: "Electromagnetic Signal Attenuation in Construction Materials," William C. Stone, 1997; available from NIST or by web download

"Effect of Admixtures, Chlorides, and Moisture on Dielectric Properties of Portland Cement Concrete in the Low Microwave Frequency Range," K. Pokkuluri (thesis), Virginia Polytechnic Institute, October, 1998

"Different Kinds of Walls and Their Effect on the Attenuation of Radiowaves Indoors," P. Ali-Rantala, L. Ukkonen, L. Sydanheimo, M. Keskilammi, and M. Kivikoski, 2003 Antennas and Propagation Society International Symposium, vol. 3, p. 1020

"Measurement and Monitoring of Microwave Reflection and Transmission Properties of Cement-Based Specimens," Kharkovsky et al., IEEE Instrumentation and Meas Tech Conf, Budapest, Hungary, May 21–23, 2001 p. 513

"Measurement and Modeling of Propagation Losses in Brick and Concrete Walls for the 900 MHz band," D. Peña, R. Feick, H. Hristov, and W. Grote, IEEE Trans Antennas and Propagation 51, p. 31, 2003

"A Comparison of Theoretical and Empirical Reflection Coefficients for Typical Exterior Wall Surfaces in a Mobile Radio Environment," O. Landron, M. Feuerstein, and T. Rappaport, IEEE Trans Ant Prop 44, p. 341, 1996

"Reflection and Transmission Losses through Common Building Materials", Robert Wilson, available from Magis Networks, www.magisnetworks.com

The Chemistry of Silica, Ralph Iler, Wiley, 1979

Indoor Propagation Studies

"The Indoor Propagation Channel," H. Hashemi, Proceedings of the IEEE, vol. 81, no. 7, p. 943, 1993

"The Impact of Surrounding Buildings on Propagation for Wireless In-Building Personal Communications System Design," S. Seidel, T. Rappaport, M. Feuerstein, K. Blackard, and L. Grindstaff, 1992 IEEE Vehicular Technology Conference, p. 814

"Indoor Throughput and Range Improvements using Standard Compliant AP Antenna Diversity in IEEE 802.11a and ETSI HIPERLAN/2," M. Abdul Aziz, M. Butler, A. Doufexi, A. Nix, and P. Fletcher, 54th IEEE Vehicular Technology Conference, October 2001, vol. 4, p. 2294

"Outdoor/Indoor Propagation Modeling for Wireless Communications Systems," M. Iskander, Z. Yun, and Z. Zhang (U Utah), IEEE Antennas and Propagation Society, AP-S International Symposium (Digest), vol. 2, 2001, pp. 150–153

"Effective Models in Evaluating Radio Coverage in Single Floors of Multifloor Buildings," J. Tarng and T. Liu, IEEE Trans Vehicular Tech, vol. 48, no. 3, May 1999

"Correlating Link Loss with Data Throughput," D. Dobkin, High Frequency Electronics, January 2003, p. 22

Interference: General

"Literature Search and Review of Radio Noise and Its Impact on Wireless Communications," Richard Adler, Wilbur Vincent, George Munsch, and David Middleton, U.S. Naval Postgraduate School, for FCC TAC, 2002

Interference: Microwave Ovens

"A Novel Prediction Tool for Indoor Wireless LAN under the Microwave Oven Interference," W. Chang, Y. Lee, C. Ko, and C. Chen, available at http://www.cert.org/research/isw/isw2000/papers/2.pd

"Effects of Microwave Oven Interference on the Performance of ISM-Band DS/SS System," S. Unawong, S. Miyamoto, and N. Morinaga, 1998 IEEE International Symposium on Electromagnetic Compatibility, vol. 1, pp. 51–56

Interference: WLAN Self-interference

"The Impact of Power Limitations and Adjacent Residence Interference on the Performance of WLANs for Home Networking Applications," S. Armour, A. Doufexi, B. Lee, A. Nix, and D. Bull, IEEE Trans. Consumer Electronics, vol. 47, p. 502, 2001

Interference: Bluetooth and WLAN

"Coexistence of IEEE 802.11b WLAN and Bluetooth WPAN," Stephen Shelhammer, Chapter 14 in Wireless Local Area Networks, Bing (op. cit.), Wiley, 2002

"An Investigation of the Impact of Bluetooth Interference on the Performance of 802.11 g Wireless Local Area Networks," A. Doufexi, A. Arumugam, S. Armour, and A. Nix, 57th IEEE Vehicular Technology Conference, 2003, vol. 1, p. 680

"The Performance of Bluetooth System in the Presence of WLAN Interference in an Office Environment," O, Larjalainen, S. Rantala, and M. Kivikoski, 8th International Conference on Communication Systems (ICCS), 2002, vol. 2, p. 628

"Rejection of Bluetooth Interference in 802.11 WLANs," A. Soltanian, R. Van Dyck, and O. Rebala, Proceedings 56th IEEE Vehicular Technology Conference, 2002, vol. 2, p. 932

Modeling

"Propagation Modelling for Indoor Wireless Communications," W. Tam and V. Tran, Electronics and Communications Engineering Journal, October, 1995, p. 221

"Wideband Propagation Modeling for Indoor Environments and for Radio Transmission into Buildings," R. Hoppe, P. Wertz, G. Wolfle, and F. Landstorfer, 11th International Symposium on Personal Indoor and Mobile Radio Communications, vol. 1, p. 282

"Efficient Ray-Tracing Acceleration Techniques for Radio Propagation Modeling," F. Agelet et al., IEEE Trans. Vehicular Tech 49, no. 6 November, 2000

"Improving the Accuracy of Ray-Tracing Techniques for Indoor Propagation Modeling," K. Remley, H. Anderson, and A. Weisshaar, IEEE Trans Vehicular Tech, vol. 49, no. 6, p. 2350, 2000

Indoor Setup Guidelines

"Wireless LAN Design, Part 1: Fundamentals of the Physical Layer," Jesse Frankel, WiFi Planet, San Jose, CA, fall 2003

This page intentionally left blank

CHAPTER 10

Security in Wireless Local Area Networks

Praphul Chandra

10.1 Introduction

The 802.11 security architecture and protocol is called Wired Equivalent Privacy (WEP). It is responsible for providing authentication, confidentiality and data integrity in 802.11 networks. To understand the nomenclature, realize that 802.11 was designed as a "wireless Ethernet." The aim of the WEP designers was therefore to provide the same degree of security as is available in traditional wired (Ethernet) networks. Did they succeed in achieving this goal?

A few years back, asking that question in the wireless community was a sure-fire way of starting a huge debate. To understand the debate, realize that wired Ethernet[1] (the IEEE 802.3 standard) implements no security mechanism in hardware or software. However, wired Ethernet networks are inherently "secured" since the access to the medium (wires) which carry the data can be restricted or secured. On the other hand, in "wireless Ethernet" (the IEEE 802.11 standard) there is no provision to restrict access to the (wireless) media. So, the debate was over whether the security provided by WEP (the security mechanism specified by 802.11) was comparable to (as secure as) the security provided by restricting access to the physical medium in wired Ethernet. Since this comparison is subjective, it was difficult to answer this question. In the absence of quantitative data for comparison, the debate raged on. However, recent loopholes discovered in WEP have pretty much settled the debate, concluding that WEP fails to achieve its goals.

In this chapter, we look at WEP, why it fails and what is being done to close these loopholes. It is interesting to compare the security architecture in 802.11 with the security architecture in Traditional Wireless Networks (TWNs). Note that both TWNs and 802.11 use the wireless medium only in the access network; that is, the part of the network which connects the end-user to the network. This part of the network is also referred to as the last hop of the network. However, there are important architectural differences between TWNs and 802.11.

The aim of TWNs was to allow a wireless subscriber to communicate with any other wireless or wired subscriber anywhere in the world while supporting seamless roaming over large

[1] We use 802.3 as a standard of comparison since it is the most widely deployed LAN standard. The analogy holds true for most other LAN standards—more or less.

geographical areas. The scope of the TWNs, therefore, went beyond the wireless access network and well into the wired network.

On the other hand, the aim of 802.11 is only last-hop wireless connectivity. 802.11 does not deal with end-to-end connectivity. In fact, IP-based data networks (for which 802.11 was initially designed) do not have any concept of end-to-end connectivity and each packet is independently routed. Also, the geographical coverage of the wireless access network in 802.11 is significantly less than the geographical coverage of the wireless access network in TWNs. Finally, 802.11 has only limited support for roaming. For all these reasons, the scope of 802.11 is restricted to the wireless access network only. As we go along in this chapter, it would be helpful to keep these similarities and differences in mind.

10.2 Key Establishment in 802.11

The key establishment protocol of 802.11 is very simple to describe—there is none. 802.11 relies on "preshared" keys between the mobile nodes or stations (henceforth Stations (STAs)) and the Access Points (APs). It does not specify how the keys are established and assumes that this is achieved in some "out-of-band" fashion. In other words, key establishment is outside the scope of WEP.

10.2.1 What's Wrong?

Key establishment is one of the toughest problems in network security. By not specifying a key establishment protocol, it seems that the 802.11 designers were side-stepping the issue. To be fair to 802.11 designers, they did a pretty good job with the standard. The widespread acceptance of this technology is a testament to this. In retrospect, security was one of the issues where the standard did have many loopholes, but then again everyone has perfect vision in hindsight. Back to our issue, the absence of any key management protocol led to multiple problems as we discuss below.

1. In the absence of any key management protocol, real life deployment of 802.11 networks ended up using manual configuration of keys into all STAs and the AP that wish to form a Basic Service Set (BSS).
2. Manual intervention meant that this approach was open to manual error.
3. Most people cannot be expected to choose a "strong" key. In fact, most humans would probably choose a key which is easy to remember. A quick survey of the 802.11 networks that I had access to shows that people use keys like "abcd1234" or "12345678" or "22222222" and so on. These keys, being alphanumeric in nature, are easy to guess and do not exploit the whole key space.
4. There is no way for each STA to be assigned a unique key. Instead, all STAs and the AP are configured with the same key. As we will see in section 10.4.4, this means

that the AP has no way of uniquely identifying a STA in a secure fashion. Instead, the STAs are divided into two groups. Group One consists of stations that are allowed access to the network, and Group Two consists of all other stations (that is, STAs which are not allowed to access the network). Stations in Group One share a secret key which stations in Group Two don't know.

5. To be fair, 802.11 does allow each STA (and AP) in a BSS to be configured with four different keys. Each STA can use any one of the four keys when establishing a connection with the AP. This feature may therefore be used to divide STAs in a BSS into four groups if each group uses one of these keys. This allows the AP a little finer control over reliable STA recognition.

6. In practice, most real life deployments of 802.11 use the same key across BSSs over the whole extended service set (ESS).[2] This makes roaming easier and faster, since an ESS has many more STAs than a BSS. In terms of key usage, this means that the same key is shared by even more STAs. Besides being a security loophole to authentication (see Section 10.4.4), this higher exposure makes the key more susceptible to compromise.

10.3 Anonymity in 802.11

We saw that subscriber anonymity was a major concern in TWNs. Recall that TWNs evolved from the voice world (the PSTN). In data networks (a large percentage of which use IP as the underlying technology), subscriber anonymity is not such a major concern. To understand why this is so, we need to understand some of the underlying architectural differences between TWNs and IP-based data networks. TWNs use IMSI for call routing. The corresponding role in IP-based networks is fulfilled by the IP address. However, unlike the IMSI, the IP address is not permanently mapped to a subscriber. In other words, given the IMSI, it is trivial to determine the identity of the subscriber. However, given the IP address, it is extremely difficult to determine the identity of the subscriber. This difficulty arises because of two reasons. First, IP addresses are dynamically assigned using protocols like DHCP; in other words, the IP address assigned to a subscriber can change over time.

Second, the widespread use of Network Address Translation (NAT) adds another layer of identity protection. NAT was introduced to deal with the shortage of IP addresses.[3] It provides IP-level access between hosts at a site (local area network (LAN)) and the rest of the Internet without requiring each host at the site to have a globally unique IP address. NAT achieves this by requiring the site to have a single connection to the global Internet and at least one

[2] Recall that an ESS is a set of APs connected by a distribution system (like Ethernet).
[3] To be accurate, the shortage of IPv4 addresses. There are more than enough IPv6 addresses available but the deployment of IPv6 has not caught on as fast as its proponents would have liked.

globally valid IP address (hereafter referred to as GIP). The address GIP is assigned to the NAT translator (also known as NAT box), which is basically a router that connects the site to the Internet. All datagrams coming into and going out of the site must pass through the NAT box. The NAT box replaces the source address in each outgoing datagram with GIP and the destination address in each incoming datagram with the private address of the correct host. From the view of any host external to the site (LAN), all datagrams come from the same GIP (the one assigned to the NAT box). There is no way for an external host to determine which of the many hosts at a site a datagram came from. Thus, the usage of NAT adds another layer of identity protection in IP networks.

10.4 Authentication in 802.11

Before we start discussing the details of authentication in 802.11 networks, recall that the concept of authentication and access control are very closely linked. To be precise, one of the primary uses of authentication is to control access to the network. Now, think of what happens when a station wants to connect to a LAN. In the wired world, this is a simple operation. The station uses a cable to plug into an Ethernet jack, and it is connected to the network. Even if the network does not explicitly authenticate the station, obtaining physical access to the network provides at least some basic access control if we assume that access to the physical medium is protected. In the wireless world, this physical-access-authentication disappears.

For a station to "connect to" or associate with a wireless local area network (WLAN), the network-joining operation becomes much more complicated. First, the station must find out which networks it currently has access to. Then, the network must authenticate the station and the station must authenticate the network. Only after this authentication is complete can the station connect to or associate with the network (via the AP). Let us go over this process in detail.

Access points (APs) in an 802.11 network (Figure 10.1) periodically broadcast beacons. Beacons are management frames which announce the existence of a network. They are used by the APs to allow stations to find and identify a network. Each beacon contains a Service Set Identifier (SSID), also called the network name, which uniquely identifies an ESS. When an STA wants to access a network, it has two options: passive scan and active scan. In the former case, it can scan the channels (the frequency spectrum) trying to find beacon advertisements from APs in the area. In the latter case, the station sends probe-requests (either to a particular SSID or with the SSID set to 0) over all the channels one-by-one. A particular SSID indicates that the station is looking for a particular network. If the concerned AP receives the probe, it responds with a probe-response. A SSID of 0 indicates that the station is looking to join any network it can access. All APs which receive this probe-request and which want this particular station to join their network, reply back with a probe-response. In either case, a station finds out which network(s) it can join.

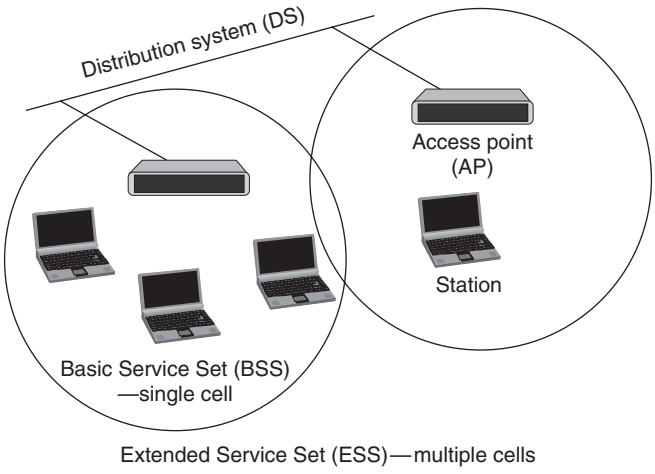
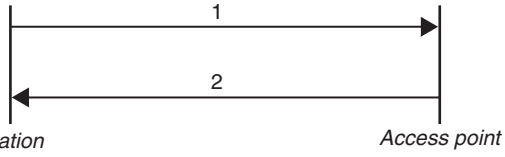


Figure 10.1: 802.11 System Overview

Next, the station has to choose a network it wishes to join. This decision can be left to the user or the software can make this decision based on signal strengths and other criteria. Once a station has decided that it wants to join a particular network, the authentication process starts. 802.11 provides for two forms of authentication: Open System Authentication (OSA) and Shared Key Authentication (SKA). Which authentication is to be used for a particular transaction needs to be agreed upon by both the STA and the network. The STA proposes the authentication scheme it wishes to use in its authentication request message. The network may then accept or reject this proposal in its authentication response message depending on how the network administrator has set up the security requirements of the network.

10.4.1 Open System Authentication

This is the default authentication algorithm used by 802.11 (Figure 10.2). Here is how it works. Any station which wants to join a network sends an authentication request to the appropriate AP. The authentication request contains the authentication algorithm that the station the wishes to use (0 in case of OSA). The AP replies back with an authentication



1. Authentication request : Auth alg = 0; Trans. num = 1.
2. Authentication resp. : Auth alg = 0; Trans. num = 2; Status = 0/*

Figure 10.2: 802.11 OSA

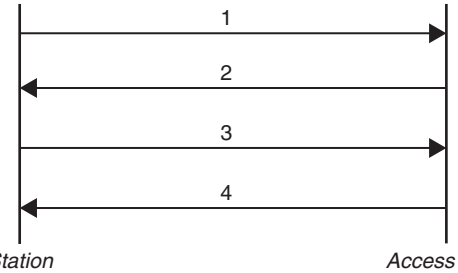
response thus authenticating the station to join the network[4] if it has been configured to accept OSA as a valid authentication scheme. In other words, the AP does not do any checks on the identity of the station and allows any and all stations to join the network. OSA is exactly what its name suggests: open system authentication. The AP (network) allows any station (that wishes to join) to join the network. Using OSA therefore means using no authentication at all.

It is important to note here that the AP can enforce the use of authentication. If a station sends an authentication request requesting to use OSA, the AP may deny the station access to the network if the AP is configured to enforce SKA on all stations.

10.4.2 Shared Key Authentication

SKA is based on the challenge-response system. SKA divides stations into two groups. Group One consists of stations that are allowed access to the network and Group Two consists of all other stations. Stations in Group One share a secret key which stations in Group Two don't know. By using SKA, we can ensure that only stations belonging to Group One are allowed to join the network.

Using SKA requires 1) that the station and the AP be capable of using WEP and 2) that the station and the AP have a preshared key. The second requirement means that a shared key must be distributed to all stations that are allowed to join the network before attempting authentication. How this is done is not specified in the 802.11 standard. Figure 10.3 explains how SKA works in detail.



1. Authentication request : Auth alg = 1; Trans. num = 1.
2. Authentication resp : Auth alg = 1; Trans. num = 2; Data = 128-byte random number.
3. Authentication resp : Auth alg = 1; Trans. num = 3; Data = encrypted (128-byte number rcvd in).
4. Authentication resp : Auth alg = 1; Trans. num = 4; Status = 0/*.

Figure 10.3: 802.11 SKA

When a station wants to join a network, it sends an authentication request to the appropriate AP which contains the authentication algorithm it wishes to use (1 in case of SKA). On receiving

[4] The authentication request from the station may be denied by the AP for reasons other than authentication failure, in which case the status field will be nonzero.

this request, the AP sends an authentication response back to the station. This authentication response contains a challenge-text. The challenge text is a 128-byte number generated by the pseudorandom-number-generator (also used in WEP) using the preshared secret key and a random Initialization Vector (IV). When the station receives this random number (the challenge), it encrypts the random number using WEP[5] and its own IV to generate a response to the challenge. Note that the IV that the station uses for encrypting the challenge is different from (and independent of) the IV that the AP used for generating the random number. After encrypting the challenge, the station sends the encrypted challenge and the IV it used for encryption back to the AP as the response to the challenge. On receiving the response, the AP decrypts the response using the preshared keys and the IV that it receives as part of the response. The AP compares the decrypted message with the challenge it sent to the station. If these are the same, the AP concludes that the station wishing to join the network is one of the stations which knows the secret key and therefore the AP authenticates the station to join the network.

The SKA mechanism allows an AP to verify that a station is one of a select group of stations. The AP verifies this by ensuring that the station knows a secret. This secret is the preshared key. If a station does not know the key, it will not be able to respond correctly to the challenge. Thus, the strength of SKA lies in keeping the shared key a secret.

10.4.3 Authentication and Handoffs

If a station is mobile while accessing the network, it may leave the range of one AP and enter into the range of another AP. In this section we see how authentication fits in with mobility (Figure 10.4).

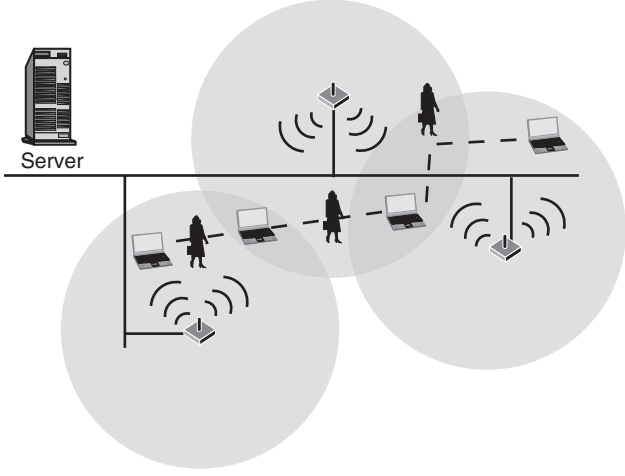


Figure 10.4: 802.11 Handoffs and Security

[5] WEP is described in section 10.5.

A STA may move inside a BSA (intra-BSA), between two BSAs (inter-BSA) or between two Extended Service Areas (ESAs) (inter-ESAs). In the intra-BSA case, the STA is static for all handoff purposes. Inter-ESA roaming requires support from higher layers (MobileIP for example) since ESAs communicate with each other at Layer 3.

It is the inter-BSA roaming that 802.11 deals with. A STA keeps track of the received signal strength (RSS) of the beacon with which it is associated. When this RSS value falls below a certain threshold, the STA starts to scan for stronger beacon signals available to it using either active or passive scanning. This procedure continues until the RSS of the current beacon returns above the threshold (in which case the STA stops scanning for alternate beacons) or until the RSS of the current beacon falls below the break-off threshold, in which case the STA decides to handoff to the strongest beacon available. When this situation is reached, the STA disconnects from its prior AP and connects to the new AP afresh (just as if had switched on in the BSA of the new AP). In fact, the association with the prior-AP is not "carried-over" or "handed-off" transparently to the new AP: the STA disconnects with the old AP and then connects with the new AP.

To connect to the new AP, the STA starts the connection procedure afresh. This means that the process of associating (and authenticating) to the new AP is the same as it is for a STA that has just powered on in this BSS. In other words, the prior-AP and the post-AP do not co-ordinate among themselves to achieve a handoff. [6] Analysis[7] has shown that authentication delays are the second biggest contributors to handoff times next only to channel scanning/probing time. This re-authentication delay becomes even more of a bottleneck for real time applications like voice. Although this is not exactly a security loophole, it is a "drawback" of using the security.

10.4.4 What's Wrong with 802.11 Authentication?

Authentication mechanisms suggested by 802.11 suffer from many drawbacks. As we saw, 802.11 specifies two modes of authentication—OSA and SKA. OSA provides no authentication and is irrelevant here.

SKA works on a challenge-response system as explained in section 10.4.2. The AP expects that the challenge it sends to the STA be encrypted using an IV and the pre-shared key. As described in section 10.2.1, there is no method specified in WEP for each STA to be assigned a unique key. Instead all STAs and the AP in a BSS are configured with the same key. This means that even when an AP authenticates a STA using the SKA mode, all it ensures is that the STA belongs to a group of STAs which know the preshared key. There is no way for the

[6] To be accurate, the IEEE 802.11 standard does not specify how the two APs should communicate with each other. There do exist proprietary solutions by various vendors which enable inter-AP communication to improve handoff performance.
[7] An Empirical Analysis of the IEEE 802.11 MAC layer Handoff Process—Mishra et al.

AP to reliably determine the exact identity of the STA that is trying to authenticate to the network and access it.[8]

To make matters worse, many 802.11 deployments share keys across APs. This increases the size of the group to which a STA can be traced. All STAs sharing a single preshared secret key also makes it very difficult to remove a STA from the allowed set of STAs, since this would involve changing (and redistributing) the shared secret key to all stations.

There is another issue with 802.11 authentication: it is one-way. Even though it provides a mechanism for the AP to authenticate the STA, it has no provision for the STA to be able to authenticate the network. This means that a rogue AP may be able to hijack the STA by establishing a session with it. This is a very plausible scenario given the plummeting cost of APs. Since the STA can never find out that it is communicating with a rogue AP, the rogue AP has access to virtually everything that the STA sends to it.

Finally, SKA is based on WEP, discussed in section 10.5. It therefore suffers from all the drawbacks that WEP suffers from too. These drawbacks are discussed in section 10.5.1.

10.4.5 Pseudo-Authentication Schemes

Networks unwilling to use SKA (or networks willing to enhance it) may rely on other authentication schemes. One such scheme allows only stations which know the network's SSID to join the network. This is achieved by having the AP responding to a probe-request from a STA only if the probe request message contains the SSID of the network. This in effect prevents connections from STAs looking for any wild carded SSIDs. From a security perspective, the secret here is the SSID of the network. If a station knows the SSID of the network, it is allowed to join the network. Even though this is a very weak authentication mechanism, it provides some form of protection against casual eavesdroppers from accessing the network. For any serious eavesdropper (hacker), this form of authentication poses minimal challenge since the SSID of the network is often transmitted in the clear (without encryption).

Yet another authentication scheme (sometimes referred to as address filtering) uses the MAC addresses as the secret. The AP maintains a list of MAC addresses of all the STAs that are allowed to connect to the network. This table is then used for admission control into the network. Only stations with the MAC addresses specified in the table are allowed to connect to the network. When a station tries to access the network via the AP, the AP verifies that the station has a MAC address which belongs to the above mentioned list. Again, even though this scheme provides some protection, it is not a very secure authentication scheme since most wireless access cards used by stations allow the user to change their MAC address via

[8] MAC addresses can be used for this purpose but they are not cryptographically protected in that it is easy to spoof a MAC address.

software. Any serious eavesdropper or hacker can find out the MAC address of one of the stations which is allowed access by listening in on the transmissions being carried out by the AP and then change their own MAC address to the determined address.

10.5 Confidentiality in 802.11

WEP uses a preestablished/preshared set of keys. Figure 10.5 shows how WEP is used to encrypt an 802.11 MAC Protocol Data Unit (MPDU). Note that Layer 3 (usually IP) hands over a MAC Service Data Unit (MSDU) to the 802.11 MAC layer. The 802.11 protocol may then fragment the MSDU into multiple MPDUs if so required to use the channel efficiently.

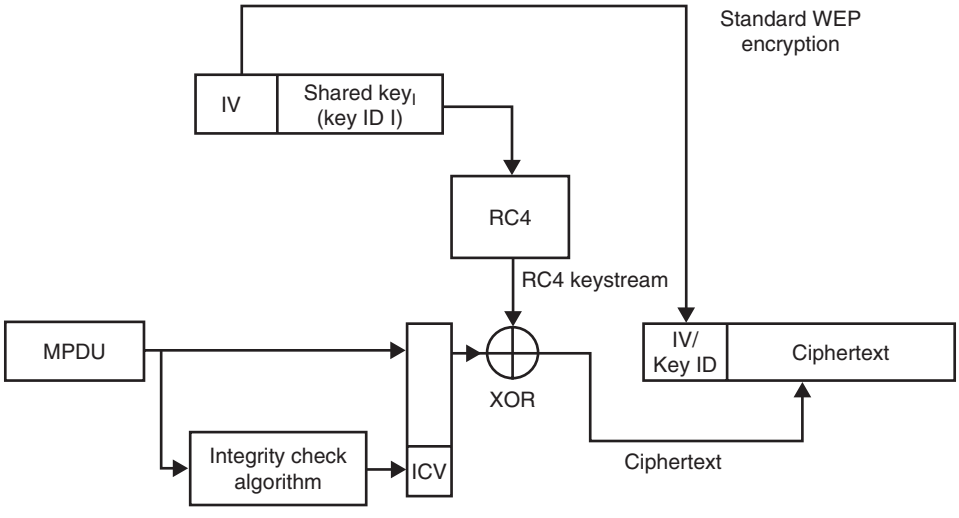


Figure 10.5: WEP

The WEP process can be broken down into the following steps.

Step 1: Calculate the Integrity Check Value (ICV) over the length of the MPDU and append this 4-byte value to the end of the MPDU. Note that ICV is another name for Message Integrity Check (MIC). We see how this ICV value is generated in section 10.6.

Step 2: Select a master key to be used from one of the four possible preshared secret keys. See section 10.2.1 for the explanation of the four possible preshared secret keys.

Step 3: Select an IV and concatenate it with the master key to obtain a key seed. WEP does not specify how to select the IV. The IV selection process is left to the implementation.

Step 4: The key seed generated in Step 3 is then fed to an RC4 key-generator. The resulting RC4 key stream is then XORed with the MPDU + ICV generated in Step 1 to generate the ciphertext.

Step 5: A 4-byte header is then appended to the encrypted packet. It contains the 3-byte IV value and a 1-byte key-id specifying which one of the four preshared secret keys is being used as the master key.

The WEP process is now completed. An 802.11 header is then appended to this packet and it is ready for transmission. The format of this packet is shown in Figure 10.6.

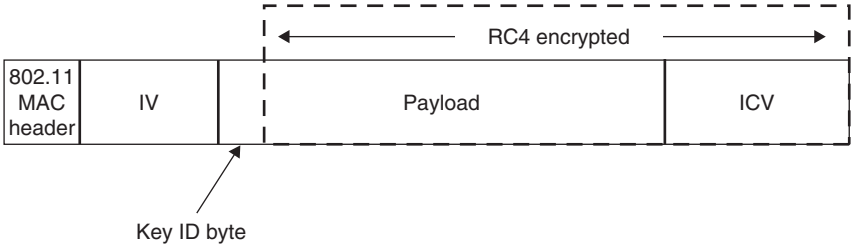


Figure 10.6: A WEP Packet

10.5.1 What's Wrong with WEP?

WEP uses RC4 (a stream cipher) in synchronous mode for encrypting data packets. Synchronous stream ciphers require that the key generators at the two communicating nodes must be kept synchronized by some external means because the loss of a single bit of a data stream encrypted under the cipher causes the loss of ALL data following the lost bit. In brief, this is so because data loss desynchronizes the key stream generators at the two endpoints. Since data loss is widespread in the wireless medium, a synchronous stream cipher is not the right choice. This is one of the most fundamental problems of WEP. It uses a cipher not suitable for the environment it operates in.

It is important to re-emphasize here that the problem here is not the RC4 algorithm.[9] The problem is that a stream cipher is not suitable for a wireless medium where packet loss is widespread. SSL uses RC4 at the application layer successfully because SSL (and therefore RC4) operates over TCP (a reliable data channel) that does not lose any data packets and can therefore guarantee perfect synchronization between the two end points.

The WEP designers were aware of the problem of using RC4 in a wireless environment. They realized that due to the widespread data loss in the wireless medium, using a synchronous stream cipher across 802.11 frame boundaries was not a viable option. As a solution, WEP attempted to solve the synchronization problem of stream ciphers by shifting synchronization requirement from a session to a packet. In other words, since the synchronization between the end-points is not perfect (and subject to packet loss), 802.11 changes keys for every packet.

[9] Though loopholes in the RC4 algorithm have been discovered too.

This way each packet can be encrypted or decrypted irrespective of the previous packet's loss. Compare this with SSL's use of RC4, which can afford to use a single key for a complete TCP session. In effect, since the wireless medium is prone to data loss, WEP has to use a single packet as the synchronization unit rather than a complete session. This means that WEP uses a unique key for each packet.

Using a separate key for each packet solves the synchronization problem but introduces problems of its own. Recall that to create a per-packet key, the IV is simply concatenated with the master key. As a general rule in cryptography, the more exposure a key gets, the more it is susceptible to be compromised. Most security architectures therefore try to minimize the exposure of the master key when deriving secondary (session) keys from it. In WEP however, the derivation of the secondary (per-packet) key from the master key is too trivial (a simple concatenation) to hide the master key.

Another aspect of WEP security is that the IV which is concatenated with the master key to create the per-packet key is transmitted in cleartext with the packet too. Since the 24-bit IV is transmitted in the clear with each packet, an eavesdropper already has access to the first three bytes of the per-packet key.

The above two weaknesses make WEP susceptible to an Fluhrer-Mantin-Shamir (FMS) attack which uses the fact that simply concatenating the IV (available in plain text) to the master key leads to the generation of a class of RC4 weak keys. The FMS attack exploits the fact that the WEP creates the per-packet key by simply concatenating the IV with the master-key. Since the first 24 bits of each per-packet key is the IV (which is available in plain text to an eavesdropper),[10] the probability of using weak keys[11] is very high. Note that the FMS attack is a weakness in the RC4 algorithm itself. However, it is the way that the per-packet keys are constructed in WEP that makes the FMS attack a much more effective attack in 802.11 networks.

The FMS attack relies on the ability of the attacker to collect multiple 802.11 packets which have been encrypted with weak keys. Limited key space (leading to key reuse) and availability of IV in plaintext which forms the first 3 bytes of the key makes the FMS attack a very real threat in WEP. This attack is made even more potent in 802.11 networks by the fact that the first 8 bytes of the encrypted data in every packet are known to be the Sub-Network Access Protocol (SNAP) header. This means that simply XORing the first 2 bytes of the encrypted pay-load with the well known SNAP header yields the first 2 bytes of the generated key-stream. In the FMS attack, if the first 2 bytes of enough key-streams are known then the RC4 key can be recovered. Thus, WEP is an ideal candidate for an FMS attack.

[10] Remember that each WEP packet carries the IV in plaintext format prepended to the encrypted packet.
[11] Use of certain key values leads to a situation where the first few bytes of the output are not all that random. Such keys are known as weak keys. The simplest example is a key value of 0.

The FMS attack is a very effective attack but is by no means the only attack which can exploit WEP weaknesses. Another such attack stems from the fact that one of the most important requirements of a synchronous stream cipher (like RC4) is that the same key should not be reused EVER. Why is it so important to avoid key reuse in RC4? Reusing the same key means that different packets use a common key stream to produce the respective ciphertext. Consider two packets of plaintext (P1 and P2) which use the same RC4 key stream for encryption.

$$\text{Since } C1 = P1 \oplus RC4(key)$$

$$\text{And } C2 = P2 \oplus RC4(key)$$

$$\text{Therefore } C1 \oplus C2 = P1 \oplus P2$$

Obtaining the XOR of the two plaintexts may not seem like an incentive for an attack but when used with frequency analysis techniques it is often enough to get lots of information about the two plaintexts. More importantly, as shown above, key reuse effectively leads to the effect of the key stream canceling out! An implication of this effect is that if one of the plaintexts (say P1) is known, P2 can be calculated easily since $P2 = (P1 \oplus P2) \oplus P1$. Another implication of this effect is that if an attacker (say, Eve) gets access to the <P1, C1> pair,[12] simply XORing the two produces the key stream K. Once Eve has access to K, she can decrypt C2 to obtain P2. Realize how the basis of this attack is the reuse of the key stream, K.

Now that we know why key reuse is prohibited in RC4, we look at what 802.11 needs to achieve this. Since we need a new key for every single packet to make the network really secure, 802.11 needs a very large key space, or rather a large number of unique keys. The number of unique keys available is a function of the key length. What is the key length used in WEP? Theoretically it is 64 bits. The devil, however, is in the details. How is the 64-bit key constructed? 24 bits come from the IV and 40 bits come from the base-key. Since the 40-bit master key never changes in most 802.11 deployments,[13] we must ensure that we use different IVs for each packet in order to avoid key reuse. Since the master key is fixed in length and the IV is only 24 bits long, the effective key length of WEP is 24 bits. Therefore, the key space for the RC4 is 2^N where N is the length of the IV. 802.11 specified the IV length as 24.

To put things in perspective, realize that if we have a 24 bit IV ($\rightarrow 2^{24}$ keys in the key-space), a busy base station which is sending 1500 byte-packets at 11 Mbps will exhaust all keys in the key space in $(1500*8)/(11*10^6*2^{24})$ seconds or about five hours. On the other hand, RC4 in SSL would use the same key space for 2^{24} ($=10^7$) sessions. Even if the application has 10,000 sessions per day, the key space would last for three years. In other words, an 802.11 BS using RC4 has to reuse the same key in about five hours whereas an application using SSL RC4 can avoid key reuse for about three years. This shows clearly that the fault lies not in the

[12] This is not as difficult as it sounds.
[13] This weakness stems from the lack of a key-establishment or key-distribution protocol in WEP.

cipher but in the way it is being used. Going beyond an example, analysis of WEP has shown that there is a 50% chance of key reuse after 4823 packets, and there is 99% chance of collision after 12,430 packets. These are dangerous numbers for a cryptographic algorithm.

Believe it or not, it gets worse. 802.11 specifies no rules for IV selection. This in turn means that changing the IV with each packet is optional. This effectively means that 802.11 implementations may use the same key to encrypt all packets without violating the 802.11 specifications. Most implementations, however, vary from randomly generating the IV on a per-packet basis to using a counter for IV generation. WEP does specify that the IV be changed "frequently." Since this is vague, it means that an implementation which generates per-packet keys (more precisely the per-MPDU key) is 802.11-compliant and so is an implementation which re-uses the same key across MPDUs.

10.6 Data Integrity in 802.11

To ensure that a packet has not been modified in transit, 802.11 uses an Integrity Check Value (ICV) field in the packet. ICV is another name for message integrity check (MIC). The idea behind the ICV/MIC is that the receiver should be able to detect data modifications or forgeries by calculating the ICV over the received data and comparing it with the ICV attached in the message. Figure 10.7 shows the complete picture of how WEP and CRC32 work together to create the MPDU for transmission.

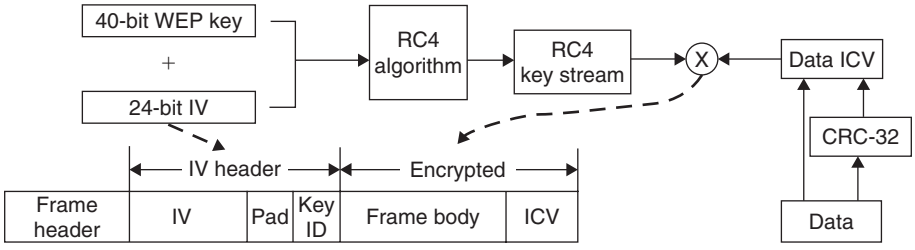


Figure 10.7: Data Integrity in WEP

The underlying assumption is that if Eve modifies the data in transit, she should not be able to modify the ICV appropriately to force the receiver into accepting the packet. In WEP, ICV is implemented as a Cyclic Redundancy Check-32 bits (CRC-32) checksum which breaks this assumption. The reason for this is that CRC-32 is linear and is not cryptographically computed, i.e., the calculation of the CRC-32 checksum does not use a key/shared secret. Also, this means that the CRC32 has the following interesting property:

$$CRC(X \oplus Y) = CRC(X) \oplus CRC(Y)$$

Now, if X represents the payload of the 802.11 packet over which the ICV is calculated, the ICV is CRC(X) which is appended to the packet. Consider an intruder who wishes to change the value of X to Z. To do this, they calculate $Y = X \oplus Z$. Then she captures the packet from the air-interface, XORs X with Y and the XORs the ICV with CRC(Y). Therefore, the packet changes from {X, CRC(X)} to $\{X \oplus Y, CRC(X) \oplus CRC(Y)\}$ or simply $\{X \oplus Y, CRC(X \oplus Y)\}$. If the intruder now retransmits the packets to the receiver, the receiver would have no way of telling that the packet was modified in transit. This means that we can change bits in the payload of the packet while preserving the integrity of the packet if we also change the corresponding bits in the ICV of the packet.

Note that an attack like the one described above works because flipping bit x in the message results in a deterministic set of bits in the CRC that must be flipped to produce the correct checksum of the modified message. This property stems from the linearity of the CRC32 algorithm.

Realize that even though the ICV is encrypted (cryptographically protected) along with the rest of the payload in the packet, it is not cryptographically computed; that is, calculating the ICV does not involve keys and cryptographic operations. Simply encrypting the ICV does not prevent an attack like the one discussed above. This is so because the flipping of a bit in the ciphertext carries through after the RC4 decryption into the plaintext because $RC4(k, X \oplus Y) = RC4(k, X) \oplus Y$ and therefore:

$$RC4(k, CRC(X \oplus Y)) = RC4(k, CRC(X)) \oplus CRC(Y)$$

The problem with the message integrity mechanism specified in 802.11 is not only that it uses a linear integrity check algorithm (CRC32) but also the fact that the ICV does not protect all the information that needs to be protected from modification. Recall from Section 10.5 that the ICV is calculated over the MPDU data; in other words, the 802.11 header is not protected by the ICV. This opens the door to redirection attacks as explained below.

Consider an 802.11 BSS where an 802.11 STA (Alice) is communicating with a wired station (Bob). Since the wireless link between Alice and the access point (AP) is protected by WEP and the wired link between Bob and access point is not,[14] it is the responsibility of the AP to decrypt the WEP packets and forward them to Bob. Now, Eve captures the packets being sent from Alice to Bob over the wireless link. She then modifies the destination address to another node, say C (Charlie), in the 802.11 header and retransmits them to the AP. Since the AP does not know any better, it decrypts the packet and forwards it to Charlie. Eve, therefore, has the AP decrypt the packets and forward them to a destination address of choice.

The simplicity of this attack makes it extremely attractive. All Eve needs is a wired station connected to the AP and she can eavesdrop on the communication between Alice and Bob

[14] WEP is an 802.11 standard used only on the wireless link.

without needing to decrypt any packets herself. In effect, Eve uses the infrastructure itself to decrypt any packets sent from an 802.11 STA via an AP. Note that this attack does not necessarily require that one of the communicating stations be a wired station. Either Bob or Charlie (or both) could as easily be other 802.11 STAs which do not use WEP. The attack would still hold since the responsibility of decryption would still be with the AP. The bottom line is that the redirection attack is possible because the ICV is not calculated over the 802.11 header. There is an interesting security lesson here. A system can't have confidentiality without integrity, since an attacker can use the redirection attack and exploit the infrastructure to decrypt the encrypted traffic.

Another problem which stems from the weak integrity protection in WEP is the threat of a replay attack. A replay attack works by capturing 802.11 packets transmitted over the wireless interface and then replaying (retransmitting) the captured packet(s) later on with (or without) modification such that the receiving station has no way to tell that the packet it is receiving is an old (replayed) packet. To see how this attack can be exploited, consider a hypothetical scenario where Alice is an account holder, Bob is a bank and Eve is another account holder in the bank. Suppose Alice and Eve do some business and Alice needs to pay Eve $500. So, Alice connects to Bob over the network and transfers $500 from her account to Eve. Eve, however, is greedy. She knows Alice is going to transfer money. So, she captures all data going from Alice to Bob. Even though Eve does not know what the messages say, she has a pretty good guess that these messages instruct Bob to transfer $500 from Alice's account to Eve's. So, Eve waits a couple of days and replays these captured messages to Bob. This may have the effect of transferring another $500 from Alice's account to Eve's account unless Bob has some mechanism for determining that he is being replayed the messages from a previous session.

Replay attacks are usually prevented by linking the integrity protection mechanism to either timestamps and/or session sequence numbers. However, WEP does not provide for any such protection.

10.7 Loopholes in 802.11 Security

To summarize, here is the list of things that are wrong with 802.11 security:

1. 802.11 does not provide any mechanism for key establishment over an unsecure medium. This means key sharing among STAs in a BSS and sometimes across BSSs.
2. WEP uses a synchronous stream cipher over a medium, where it is difficult to ensure synchronization during a complete session.
3. To solve the previous problem, WEP uses a per-packet key by concatenating the IV directly to the preshared key to produce a key for RC4. This exposes the base key or master key to attacks like FMS.

4. Since the master key is usually manually configured and static and since the IV used in 802.11 is just 24 bits long, this results in a very limited key-space.
5. 802.11 specifies that changing the IV with each packet is optional, thus making key reuse highly probable.
6. The CRC-32 used for message integrity is linear.
7. The ICV does not protect the integrity of the 802.11 header, thus opening the door to redirection attacks.
8. There is no protection against replay attacks.
9. There is no support for a STA to authenticate the network.

Note that the limited size of the IV figures much lower in the list than one would expect. This emphasizes the fact that simply increasing the IV size would not improve WEP's security considerably. The deficiency of the WEP encapsulation design arises from attempts to adapt RC4 to an environment for which it is poorly suited.

10.8 WPA

When the loopholes in WEP, the original 802.11 security standard, had been exposed, IEEE formed a Task Group: 802.11i with the aim of improving upon the security of 802.11 networks. This group came up with the proposal of a Robust Security Network (RSN). A RSN is an 802.11 network which implements the security proposals specified by the 802.11i group and allows only RSN-capable devices to join the network, thus allowing no "holes." The term hole is used to refer to a non-802.11i compliant STA which by virtue of not following the 802.11i security standard could make the whole network susceptible to a variety of attacks.

Since making a transition from an existing 802.11 network to a RSN cannot always be a single-step process (we will see why in a moment), 802.11i allows for a Transitional Security Network (TSN) which allows for the existence of both RSN and WEP nodes in an 802.11 network. As the name suggests, this kind of a network is specified only as a transition point and all 802.11 networks are finally expected to move to a RSN. The terms RSN and 802.11i are sometimes used interchangeably to refer to this security specification.

The security proposal specified by the Task Group-i uses the Advanced Encryption Standard (AES) in its default mode. One obstacle in using AES is that it is not backward compatible with existing WEP hardware. This is so because AES requires the existence of a new more powerful hardware engine. This means that there is also a need for a security solution which can operate on existing hardware. This was a pressing need for vendors of 802.11 equipment. This is where the Wi-Fi alliance came into the picture.

The Wi-Fi alliance is an alliance of major 802.11 vendors formed with the aim of ensuring product interoperability. To improve the security of 802.11 networks without requiring a hardware upgrade, the Wi-Fi alliance adopted Temporal Key Integrity Protocol (TKIP) as the security standard that needs to be deployed for Wi-Fi certification. This form of security has therefore come to be known as Wi-Fi Protected Access (WPA). WPA is basically a prestandard subset of 802.11i which includes the key management and the authentication architecture (802.1X) specified in 802.11i. The biggest difference between WPA and 802l.11i (which has also come to be known as WPA2) is that instead of using AES for providing confidentiality and integrity, WPA uses TKIP and MICHAEL respectively. We look at TKIP/WPA in this section and the 802.11i/WPA2 using AES in the next section.

TKIP stands for Temporal Key Integrity Protocol. It was designed to fix WEP loopholes while operating within the constraints of existing 802.11 equipment (APs, WLAN cards and so on). To understand what we mean by the "constraints of existing 802.11 hardware," we need to dig a little deeper. Most 802.11 equipment consists of some sort of a WLAN Network Interface Card (NIC) (also known as WLAN adapter) which enables access to an 802.11 network. A WLAN NIC usually consists of a small microprocessor, some firmware, a small amount of memory and a special-purpose hardware engine. This hardware engine is dedicated to WEP implementation since software implementations of WEP are too slow. To be precise, the WEP encryption process is implemented in hardware. The hardware encryption takes the IV, the base (master) key and the plaintext data as the input and produces the encrypted output (ciphertext). One of the most severe constraints for TKIP designers was that the hardware engine cannot be changed. We see in this section how WEP loopholes were closed given these constraints.

10.8.1 Key Establishment

One of the biggest WEP loopholes is that it specifies no key-establishment protocol and relies on the concept of preshared secret keys which should be established using some out-of-band mechanism. Realize that this is a system architecture problem. In other words, solving this problem requires support from multiple components (the AP, the STA and usually also a backend authentication server) in the architecture.

One of the important realizations of the IEEE 802.11i task group was that 802.11 networks were being used in two distinct environments: the home network and the enterprise network. These two environments had distinct security requirements and different infrastructure capacities to provide security. Therefore, 802.11i specified two distinct security architectures. For the enterprise network, 802.11i specifies the use of IEEE 802.1X for key establishment and authentication. As we will see in our discussion in the next section, 802.1X requires the use of a backend authentication server. Deploying a back end authentication server is

not usually feasible in a home environment. Therefore, for home deployments of 802.11, 802.11i allows the use of the "out-of-band mechanism" (read manual configuration) for key establishment.

We look at the 802.1X architecture in the next section and see how it results in the establishment of a Master Key (MK). In this section, we assume that the two communicating end-points (the STA and the AP) already share a MK which has either been configured manually at the two end-points (WEP architecture) or has been established using the authentication process (802.1X architecture). This section looks at how this MK is used in WPA.

Recall that a major loophole in WEP was the manner[15] in which this master key was used which made it vulnerable to compromise. WPA solves this problem by reducing the exposure of the master key, thus making it difficult for an attacker to discover the master key. To achieve this, WPA adds an additional layer to the key hierarchy used in WEP. Recall from Section 6.4 that WEP uses the master key for authentication and to calculate the per-packet key. In effect there is a two-tier key hierarchy in WEP: the master (preshared secret) key and the per-packet key.

WPA extends the two-tier key-hierarchy of WEP to a multitier hierarchy (See Figure 10.8). At the top level is still the master key, referred to as the Pair-wise Master Key (PMK) in WPA. The next level in the key hierarchy is the PTK which is derived from the PMK. The final level is the per-packet keys which are generated by feeding the PTK to a key-mixing function. Compared with the two-tier WEP key hierarchy, the three-tier key hierarchy of WPA avoids exposing the PMK in each packet by introducing the concept of PTK.

As we saw, WPA is flexible about how the master key (PMK in WPA) is established. The PMK, therefore, may be a preshared[16] secret key (WEP-design) or a key derived from an authentication process like 802.1X.[17] WPA does require that the PMK be 256 bits (or 32 bytes) long. Since a 32-byte key is too long for humans to remember, 802.11 deployments using preshared keys may allow the user to enter a shorter password which may then be used as a seed to generate the 32-byte key.

The next level in the key hierarchy after the PMK are the PTK. WPA uses the PMK for deriving the Pair-wise Transient Keys (PTK) which are basically session keys. The term PTK is used to refer to a set of session keys which consists of four keys, each of which is 128 bits long. These four keys are as follows: an encryption key for data, an integrity key

[15] The per-packet key is obtained by simply concatenating the IV with the preshared secret key. Therefore, a compromised per-packet key exposes the preshared secret key.
[16] As we saw, this usually means that the keys are manually configured.
[17] It is expected that most enterprise deployments of 802.11 would use 802.1X while the preshared secret key method (read manual configuration) would be used by residential users.

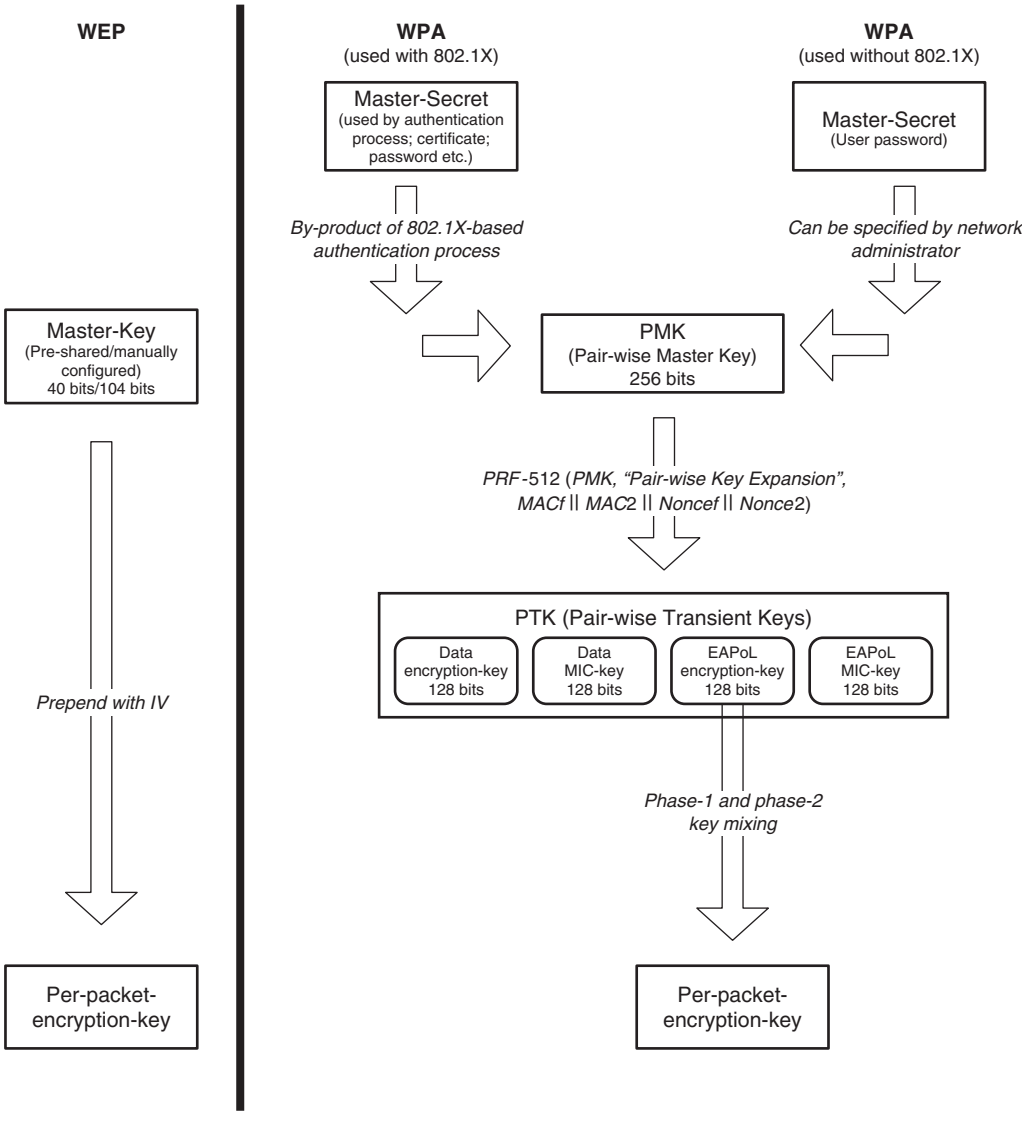


Figure 10.8: Key Hierachy in 802.11

for data, an encryption key for EAPoL messages and an integration key for EAPoL messages. Note that the term *session* here refers to the association between a STA and an AP. Every time a STA associates with an AP, it is the beginning of a new session and this results in the generation of a new PTK (set of keys) from the PMK. Since the session keys are valid only for a certain period of time, they are also referred to as temporal keys and the set of four session keys together is referred to as the Pair-wise Transient Keys (PTK). The PTK are derived from the PMK using a Pseudorandom Function (PRF). The PRFs used for derivation

of PTKs (and nonces) are explicitly specified by WPA and are based on the HMAC-SHA algorithm.

PTK = PRF-512(PMK, "Pair-wise key expansion", AP_MAC || STA_MAC || ANonce || SNonce)

Realize that to obtain the PTK from the PMK we need five input values: the PMK, the MAC addresses of the two endpoints involved in the session and one nonce each from the two endpoints. The use of the MAC addresses in the derivation of the PTK ensures that the keys are bound to sessions between the two endpoints and increases the effective key space of the overall system.

Realize that since we want to generate a different set of session keys from the same PMK for each new session,[18] we need to add another input into the key generation mechanism which changes with each session. This input is the nonce. The concept of nonce is best understood by realizing that it is short for Number-Once. The value of nonce is thus arbitrary except that a nonce value is never used again.[19] Basically it is a number which is used only once. In our context, a nonce is a unique number (generated randomly) which can distinguish between two sessions established between a given STA and an AP at different points in time. The two nonces involved in PTK generation are generated, one each, by the two end points involved in the session; i.e., the STA (SNonce) and the AP (ANonce). WPA specifies that a nonce should be generated as follows:

ANonce = PRF-256(Random Number, "Init Counter", AP_MAC || Time)

SNonce = PRF-256(Random Number, "Init Counter", STA_MAC || Time)

The important thing to note is that the PTKs are effectively shared between the STA and the AP and are used by both the STA and the AP to protect the data/EAPoLmessages they transmit. It is therefore important that the input values required for derivation of PTK from the PMK come from *both* the STA and the AP. Note also that the key derivation process can be executed in parallel at both endpoints of the session (the STA and the AP) once the Nonces and the MAC addresses have been exchanged. Thus, both the STA and the AP can derive the same PTK from the PMK simultaneously.

The next step in the key hierarchy tree is to derive per-packet keys from the PTK. WPA improves also upon this process significantly. Recall from Section 10.5 that the per-packet key was obtained by simply concatenating the IV with the master key in WEP. Instead of

[18] If a STA disconnects from the AP and connects back with an AP at a later time, these are considered two different sessions.

[19] To be completely accurate, nonce values are generated such that the probability of the same value being generated twice is very low.

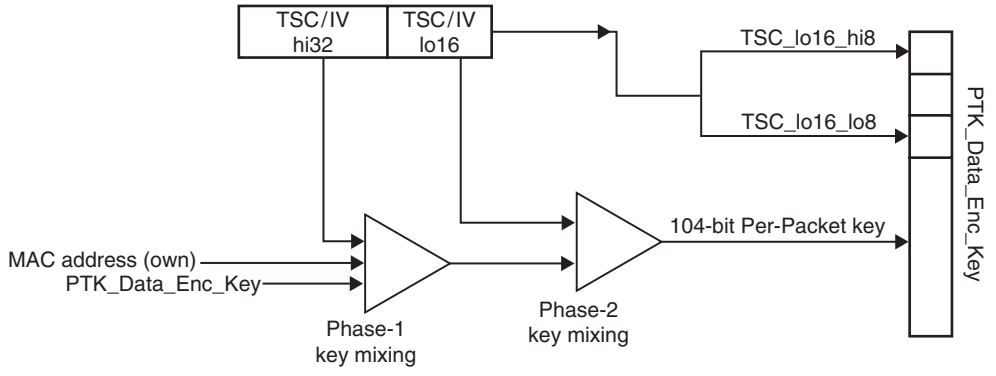


Figure 10.9: TKIP Encryption

simply concatenating the IV with the master key, WPA uses the process shown in Figure 10.9 to obtain the per packet key. This process is known as per-packet key mixing and is shown in Figure 10.9.

In phase one, the session data encryption key is "combined" with the high order 32 bits of the IV and the MAC address. The output from this phase is "combined" with the lower order 16 bits of the IV and fed to phase two, which generates the 104-bit per-packet key. There are many important features to note in this process:

1. It assumes the use of a 48-bit IV (more of this in section 10.8.2).
2. The size of the encryption key is still 104 bits, thus making it compatible with existing WEP hardware accelerators.
3. Since generating a per-packet key involves a hash operation which is computation intensive for the small MAC processor in existing WEP hardware, the process is split into two phases. The processing intensive part is done in phase one whereas phase two is much less computation intensive.
4. Since phase one involves the high order 32 bits of the IV, it needs to be done only when one of these bits change; that is, once in every 65,536 packets.
5. The key-mixing function makes it very hard for an eavesdropper to correlate the IV and the per-packet key used to encrypt the packet.

10.8.2 Authentication

As we said in the previous section, 802.11i specified two distinct security architectures. For the home network, 802.11i allows the manual configuration of keys just like WEP. For the enterprise network however, 802.11i specifies the use of IEEE 802.1X for key establishment and authentication. We just summarize the 802.1X architecture in this section.

802.1X is closely architected along the lines of EAPoL (EAP over LAN). Figure 10.10a shows the conceptual architecture of EAPoL and Figure 10.10b shows the overall system architecture of EAPoL. The controlled port is open only when the device connected to the authenticator has been authorized by 802.1x. On the other hand, the uncontrolled port provides a path for extensible authentication protocol over LAN (EAPoL) traffic ONLY. Figure 10.10a shows how access to even the uncontrolled port may be limited using MAC filtering.[20] This scheme is sometimes used to deter DoS attacks.

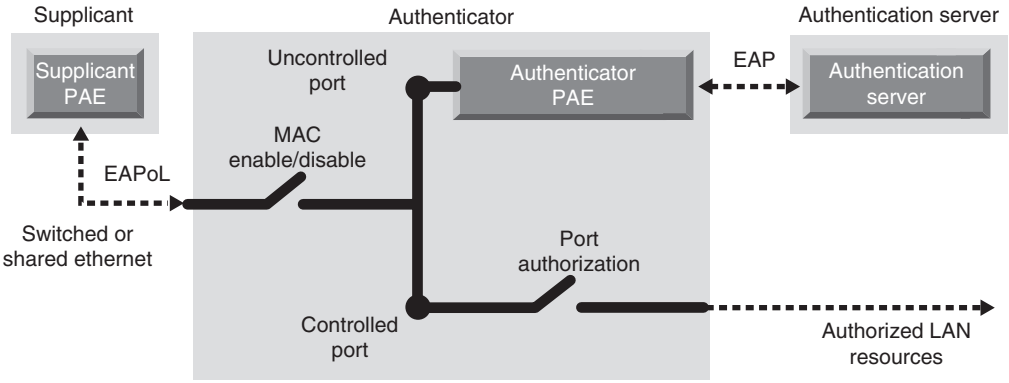


Figure 10.10a: 802.1X/EAP Port Model

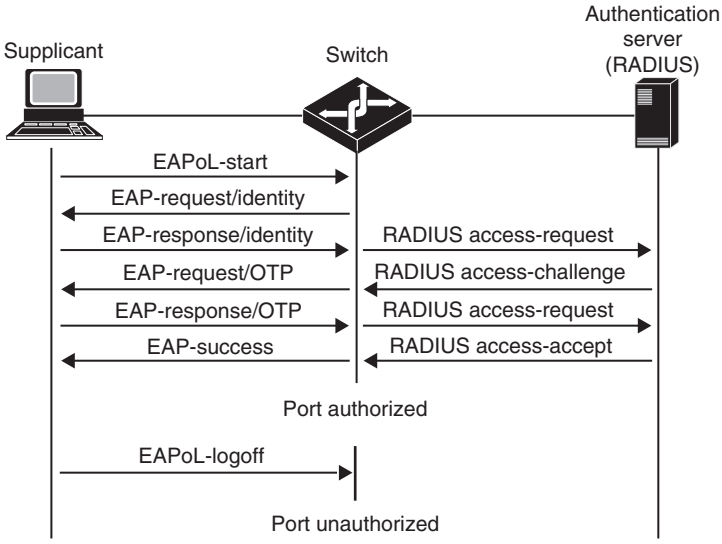


Figure 10.10b: EAPoL

[20] Allowing only STAs with have a MAC address which is "registered" or "known" to the network.

EAP specifies three network elements: the supplicant, the authenticator and the authentication server. For EAPoverLAN, the end user is the supplicant, the Layer 2 (usually Ethernet) switch is the authenticator controlling access to the network using logical ports, and the access decisions are taken by the backend authentication server after carrying out the authentication process. Which authentication process to use (MD5, TLS and so on) is for the network administrator to decide.

EAPoL can be easily adapted to be used in the 802.11 environment as shown in Figure 10.10c. The STA is the supplicant, the AP is the authenticator controlling access to the network, and there is a backend authentication server. The analogy is all the more striking if you consider that an AP is in fact just a Layer 2 switch, with a wireless and a wired interface.

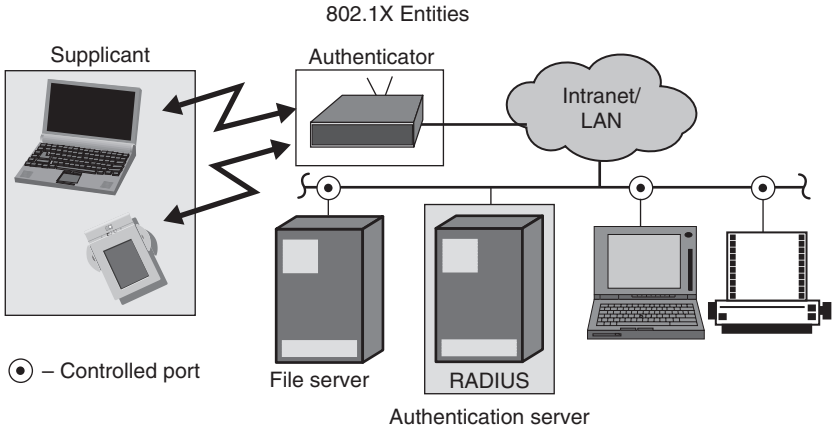


Figure 10.10c: EAP Over WLAN

There is, however, one interesting piece of detail that needs attention. The 802.1X architecture carries the authentication process between the supplicant (STA) and the backend authentication server.[21] This means that the master key (resulting from an authentication process like TLS) is established between the STA and backend server. However, confidentiality and integrity mechanisms in the 802.11 security architecture are implemented between the AP and the STA. This means that the session (PTK) and per packet keys (which are derived from the PMK) are needed at the STA and the AP. The STA already has the PMK and can derive the PTK and the per-packet keys. However, the AP does not yet have the PMK. Therefore, what is needed is a mechanism to get the PMK from the authentication server to the AP securely.

Recall that in the 802.1X architecture, the result of the authentication process is conveyed by the authentication server to the AP so that the AP may allow or disallow the STA access to

[21] With the AP controlling access to the network using logical ports.

the network. The communication protocol between the AP and the authentication server is not specified by 802.11i but is specified by WPA to be RADIUS. Most deployments of 802.11 would probably end up using RADIUS. The RADIUS protocol does allow for distributing the key securely from the authentication server to the AP and this is how the PMK gets to the AP.

Note that 802.1X is a framework for authentication. It does not specify the authentication protocol to be used. Therefore, it is up to the network administrator to choose the authentication protocol they want to plug in to the 802.1X architecture. One of the most often discussed authentication protocols to be used with 802.1X is TLS. Section 3.3.3 discusses how the TLS protocol is used with EAPoL. Figure 10.10d summarizes how TLS can be used as an authentication protocol in a EAP over WLAN environment. The EAP-TLS protocol is well documented. It has been analyzed extensively and no significant weaknesses have been found in the protocol itself. This makes it an attractive option for security use in 802.1X. However, there is a deployment issue with this scheme.

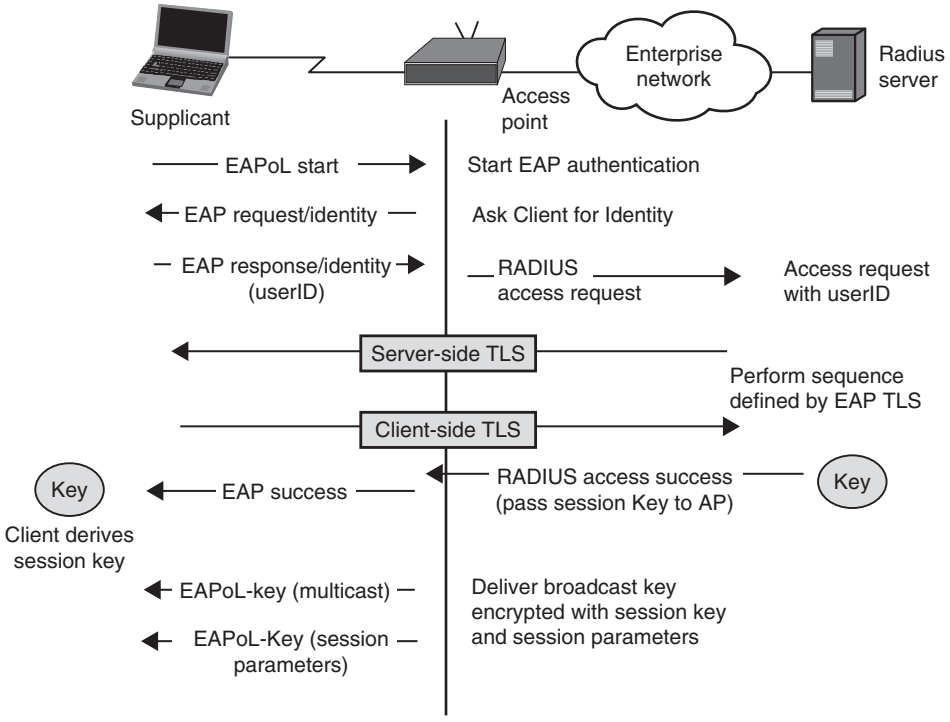


Figure 10.10d: 802.1X Network Architecture

Note that EAP-TLS relies on certificates to authenticate the network to the clients and the clients to the networks. Requiring the network (the servers) to have certificates is a common theme in most security architectures. However, the requirement that each client be issued

a certificate leads to the requirement of the wide spread deployment of PKI. Since this is sometimes not a cost effective option, a few alternative protocols have been proposed: EAP-TTLS (tunneled TLS) and PEAP. Both of these protocols use certificates to authenticate the network (the server) to the client but do not use certificates to authenticate the client to the server. This means that a client no longer needs a certificate to authenticate itself to the server: instead the clients can use password-based schemes (CHAP, PAP and so on) to authenticate themselves. Both protocols divide the authentication process in two phases. In phase 1, we authenticate the network (the server) to the client using a certificate and establish a TLS tunnel between the server and the client. This secure[22] TLS channel is then used to carry out a password-based authentication protocol to authenticate the client to the network (server).

10.8.3 Confidentiality

Recall from Section 10.5.1 that the fundamental WEP loophole stems from using a stream cipher in an environment susceptible to packet loss. To work around this problem, WEP designers changed the encryption key for each packet. To generate the per-packet encryption key, the IV was concatenated with the preshared key. Since the preshared key is fixed, it is the IV which is used to make each per-packet key unique. There were multiple problems with this approach.

First, the IV size at 24 bits was too short. At 24 bits there were only 16,777,216 values before a duplicate IV value was used. Second, WEP did not specify how to select an IV for each packet.[23] Third, WEP did not even make it mandatory to vary the IV on a per-packet basis—realize that this meant WEP explicitly allowed reuse of per-packet keys. Fourth, there was no mechanism to ensure that the IV was unique on a per station basis. This made the IV collision space shared between stations, thus making a collision even more likely. Finally, simply concatenating the IV with the preshared key to obtain a per-packet key is cryptographically unsecure, making WEP vulnerable to the FMS attack. The FMS attack exploits the fact that the WEP creates the per-packet key by simply concatenating the IV with the master-key. Since the first 24 bits of each per-packet key is the IV (which is available in plain text to an eavesdropper),[24] the probability of using weak keys[25] is very high.

First off, TKIP doubles the IV size from 24 bits to 48 bits. This results in increasing the time to key collision from a few hours to a few hundred years. Actually, the IV is increased from 24 bits to 56 bits by requiring the insertion of 32 bits between the existing WEP IV and the start of the encrypted data in the WEP packet format. However, only 48 bits of the IV are used since eight bits are reserved for discarding some known (and some yet to be discovered) weak keys.

[22] Secure since it protects the identity of the client during the authentication process.
[23] Implementations vary from a sequential increase starting from zero to generating a random IV for each packet.
[24] Remember that each WEP packet carries tha IV in plain text format prepended to he encrypted packet.
[25] Use of certain key values leads to a situation where the first few bytes of the output are not all that random. Such keys are known as weak keys. The simplest example is a key value of 0.

Simply increasing the IV length will, however, not work with the existing WEP hardware accelerators. Remember that existing WEP hardware accelerators expect a 24-bit IV as an input to concatenate with a preshared key (40/104-bit) in order to generate the per-packet key (64/128-bit). This hardware cannot be upgraded to deal with a 48-bit IV and generate an 88/156-bit key. The approach, therefore, is to use per-packet key mixing as explained in Section 10.8.1. Using the per-packet key mixing function (much more complicated) instead of simply concatenating the IV to the master key to generate the per-packet key increases the effective IV size (and hence improves on WEP security) while still being compatible with existing WEP hardware.

10.8.4 Integrity

WEP used CRC-32 as an integrity check. The problem with this protocol was that it was linear. As we saw in Section 10.6, this is not a cryptographically secure integrity protocol. It does however have the merit that it is not computation intensive. What TKIP aims to do is to specify an integrity protocol which is cryptographically secure and yet not computation intensive so that it can be used on existing WEP hardware which has very little computation power. The problem is that most well known protocols used for calculating a message integrity check (MIC) have lots of multiplication operations and multiplication operations are computation intensive. Therefore, TKIP uses a new MIC protocol—MICHAEL—which uses no multiplication operations and relies instead on shift and add operations. Since these operations require much less computation, they can be implemented on existing 802.11 hardware equipment without affecting performance.

Note that the MIC value is added to the MPDU in addition to the ICV which results from the CRC32. It is also important to realize that MICHAEL is a compromise. It does well to improve upon the linear CRC-32 integrity protocol proposed in WEP while still operating within the constraints of the limited computation power. However, it is in no way as cryptographically secure as the other standardized MIC protocols like MD5 or SHA-1. The TKIP designers knew this and hence built in countermeasures to handle cases where MICHAEL might be compromised. If a TKIP implementation detects two failed forgeries (two packets where the calculated MIC does not match the attached MIC) in one second, the STA assumes that it is under attack and as a countermeasure deletes its keys, disassociates, waits for a minute and then re-associates. Even though this may sound a little harsh, since it disrupts communication, it does avoid forgery attacks.

Another enhancement that TKIP makes in IV selection and use is to use the IV as a sequence counter. Recall that WEP did not specify how to generate a per-packet IV.[26] TKIP explicitly requires that each STA start using an IV with a value of 0 and increment the value by one for each packet that it transmits during its session[27] lifetime. This is the reason the IV can also be

[26] In fact, WEP did not even specify that the IV had to be changed on a per-packet basis.
[27] An 802.11 session refers to the association between a STA and an AP.

used as a TKIP Sequence Counter (TSC). The advantage of using the IV as a TSC is to avoid the replay attack to which WEP was susceptible.

TKIP achieves replay protection by using a unique IV with each packet that it transmits during a session. This means that in a session, each new packet coming from a certain MAC address would have a unique number.[28] If each packet from Alice had a unique number, Bob could tell when Eve was replaying old messages. WEP does not have replay protection since it cannot use the IV as a counter. Why? Because WEP does not specify how to change IV from one packet to another and as we saw earlier, it does not even specify that you need to.

10.8.5 The Overall Picture: Confidentiality + Integrity

The overall picture of providing confidentiality and message integrity in TKIP is shown in Figure 10.10e.

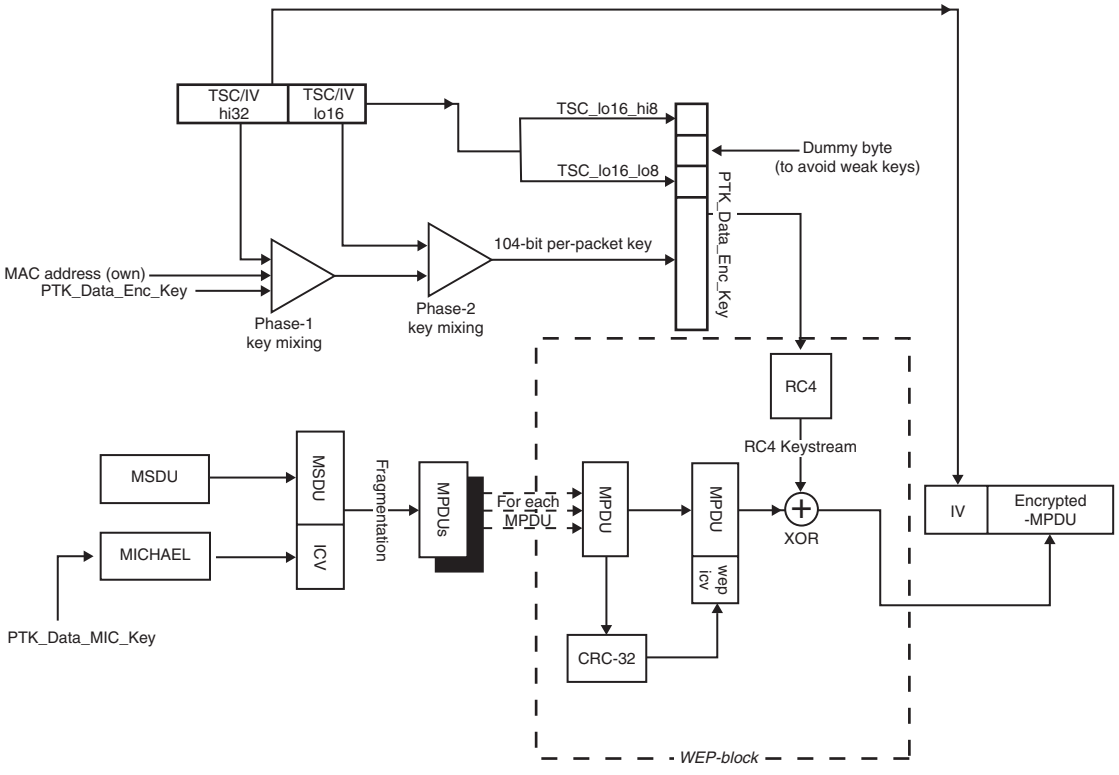


Figure 10.10e: TKIP—The Complete Picture

[28] At least for 900 years—that's when the IV rolls over.

10.8.6 How Does WPA Fix WEP Loopholes?

In section 10.7 we summarized the loopholes of WEP. At the beginning of section 10.8 we said that WPA/TKIP was designed to close these loopholes while still being able to work with existing WEP hardware. In this section, we summarize what WPA/TKIP achieves and how.

WEP	WPA
Relies on preshared (out-of-band) key establishment mechanisms. Usually leads to manual configuration of keys and to key sharing among STAs in a BSS (often ESS).	Recommends 802.1X for authentication and key-establishment in enterprise deployments. Also supports preshared key establishment like WEP.
Uses a synchronous stream cipher which is unsuitable for the wireless medium.	Same as WEP.
Generates per-packet key by concatenating the IV directly to the master/preshared key thus exposing the base-key/master-key to attacks like FMS.	Solves this problem by (a) introducing the concept of PTK in the key hierarchy and (b) by using a key mixing function instead of simple concatenation to generate per-packet keys. This reduces the exposure of the master key.
Static master key + Small size of IV + Method of per-packet key generation → Extremely limited key space.	Increases the IV size to 56 bits and uses only 48 of these bits reserving 8-bits to discard weak keys. Also, use of PTK which are generated afresh for each new session increases the effective key space.
Changing the IV with each packet is optional → key-reuse highly probable.	Explicitly specifies that both the transmitter and the receiver initialize the IV to zero whenever a new set of PTK is established[29] and then increment it by one for each packet it sends.
Linear algorithm (CRC-32) used for message integrity → Weak integrity protection.	Replaces the integrity check algorithm to use MICHAEL which is nonlinear. Also, specifies countermeasures for the case where MICHAEL may be violated.
ICV does not protect the integrity of the 802.11 header → Susceptible to Redirection Attacks.	Extends the ICV computation to include the MAC source and destination address to protect against Redirection attacks.
No protection against replay attacks.	The use of IV as a sequence number provides replay protection.
No support for a STA to authenticate the network.	Use of 802.1X in enterprise deployments allows for this.

[29] This usually happens every time the STA associates with an AP.

10.9 WPA2 (802.11i)

Recall from section 10.8 that Wi-Fi protected access (WPA) was specified by the Wi-Fi alliance with the primary aim of enhancing the security of existing 802.11 networks by designing a solution which could be deployed with a simple software (firmware) upgrade and without the need for a hardware upgrade. In other words, WPA was a stepping stone to the final solution which was being designed by the IEEE 802.11i task group. This security proposal was referred to as the Robust Security Network (RSN) and also came to be known as the 802.11i security solution. The Wi-Fi alliance integrated this solution in their proposal and called it WPA2. We look at this security proposal in this section.

10.9.1 Key Establishment

WPA was a prestandard subset of IEEE 802.11i. It adopted the key-establishment, key hierarchy and authentication recommendations of 802.11i almost completely. Since WPA2 and 802.11i standard are the same, the key-establishment process and the key hierarchy architecture in WPA and WPA2 are almost identical. There is one significant difference though. In WPA2, the same key can be used for the encryption and integrity protection of data. Therefore, there is one less key needed in WPA2. For a detailed explanation of how the key hierarchy is established see section 10.8.1.

10.9.2 Authentication

Just like key establishment and key hierarchy, WPA had also adopted the authentication architecture specified in 802.11i completely. Therefore, the authentication architecture in WPA and WPA2 is identical. For a detailed explanation of how the authentication architecture see section 10.8.2.

10.9.3 Confidentiality

In this section we look at the confidentiality mechanism of WPA2 (802.11i). Recall that the encryption algorithm used in WEP was RC4, a stream cipher. Some of the primary weaknesses in WEP stemmed from using a stream cipher in an environment where it was difficult to provide lossless synchronous transmission. It was for this reason that Task Group i specified the use of a block encryption algorithm when redesigning 802.11 security. Since AES was (and still is) considered the most secure block cipher, it was an obvious choice. This was a major security enhancement since the encryption algorithm lies at the heart of providing confidentiality.

In addition to specifying an encryption algorithm for providing system security, what is also needed is to specify a mode of operation. To provide confidentiality in 802.11i, AES is used in the counter mode. Counter mode actually uses a block cipher as a stream cipher, thus

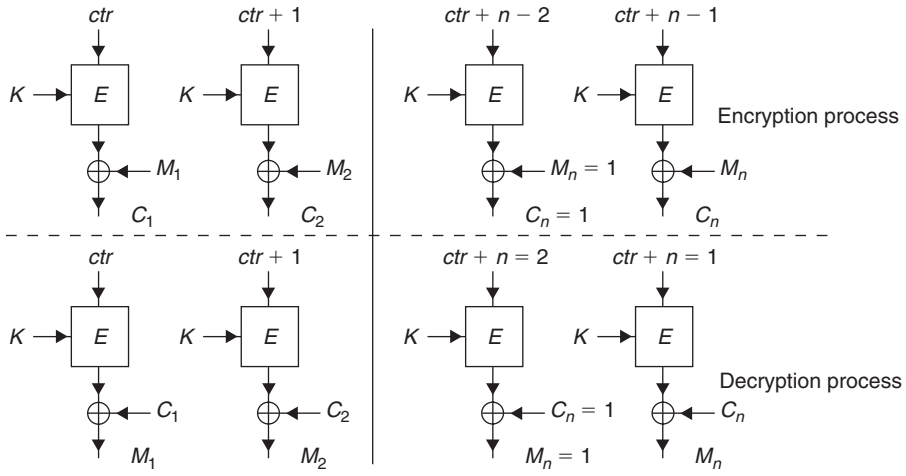


Figure 10.11: AES Counter Mode

combining the security of a block cipher with the ease of use of a stream cipher. Figure 10.11 shows how AES counter mode works.

Using the counter mode requires a counter. The counter starts at an arbitrary but predetermined value and is incremented in a specified fashion. The simplest counter operation, for example, would start the counter with an initial value of 1 and increment it sequentially by 1 for each block. Most implementations however, derive the initial value of the counter from a nonce value that changes for each successive message. The AES cipher is then used to encrypt the counter to produce a key stream. When the original message arrives, it is broken up into 128-bit blocks and each block is XORed with the corresponding 128 bits of the generated key stream to produce the ciphertext.

Mathematically, the encryption process can be represented as $C_i = M_i\,(+)\,E_k(i)$ where i is the counter. The security of the system lies in the counter. As long as the counter value is never repeated with the same key, the system is secure. In WPA2, this is achieved by using a fresh key for every session (See section 10.8.1.).

To summarize, the salient features of AES in counter mode are as follows:

1. It allows a block cipher to be operated as a stream cipher.
2. The use of counter mode makes the generated key stream independent of the message, thus allowing the key stream to be generated before the message arrives.
3. Since the protocol by itself does not create any interdependency between the encryption of the various blocks in a message, the various blocks of the message can be encrypted in parallel if the hardware has a bank of AES encryption engines.

4. Since the decryption process is exactly the same as encryption,[30] each device only needs to implement the AES encryption block.
5. Since the counter mode does not require that the message be broken up into an exact number of blocks, the length of the encrypted text can be exactly the same as the length of the plain text message.

Note that the AES counter mode provides only for the confidentiality of the message and not the message integrity. We see how AES is used for providing the message integrity in the next section. Also, since the encryption and integrity protection processes are very closely tied together in WPA2/802.11i, we look at the overall picture after we have discussed the integrity process.

10.9.4 Integrity

To achieve message integrity, Task Group i extended the counter mode to include a Cipher Block Chaining (CBC)-MAC operation. This is what explains the name of the protocol: AES-CCMP where CCMP stands for Counter-mode CBC-MAC protocol. The CBC-MAC protocol (also known as CBC-residue) is reproduced here in Figure 10.12 where the black boxes represent the encryption protocol (AES in our case).

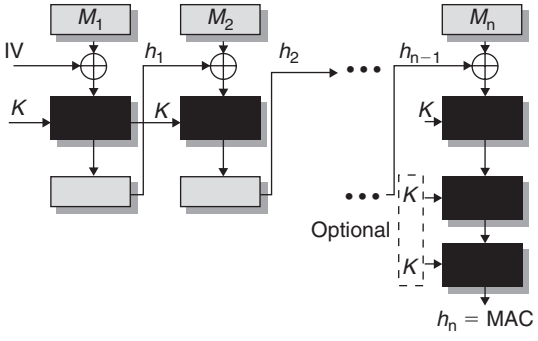


Figure 10.12: AES CBC-MAC

As shown in the figure, CBC-MAC XORs a plaintext block with the previous cipher block before encrypting it. This ensures that any change made to any cipher text (for example by a malicious intruder) block changes the decrypted output of the last block and hence changes the residue. CBC-MAC is an established technique for message integrity. What Task Group i did was to combine the counter mode of operation with the CBC-MAC integrity protocol to create the CCMP.

[30] XORing the same value twice leads back to the original value.

10.9.5 The Overall Picture: Confidentiality + Integrity

Since a single process is used to achieve integrity and confidentiality, the same key can be used for the encryption and integrity protection of data. It is for this reason that there is one less key needed in WPA2. The complete process which combines the counter mode encryption and CBC-MAC integrity works as follows.

In WPA2, the PTK is 384 bits long. Of this, the most significant 256 bits form the EAPoL MIC key and EAPoL encryption key. The least significant 128 bits form the data key. This data key is used for both encryption and integrity protection of the data. Before the integrity protection or the encryption process starts, a CCMP header is added to the 802.11 packet before transmission. The CCMP header is eight bytes in size. Of these eight bytes, six bytes are used for carrying the Packet Number (PN) which is needed for the other (remote) end to decrypt the packet and to verify the integrity of the packet. One byte is reserved for future use and the remaining byte contains the key ID. Note that the CCMP header is prepended to the payload of the packet and is not encrypted since the remote end needs to know the PN before it starts the decryption or the verification process. The PN is a per-packet sequence number which is incremented for each packet processed

The integrity protection starts with the generation of an Initialization Vector (IV) for the CBC-MAC process. This IV is created by the concatenation of the following entities: flag, priority, source MAC address, a PN and DLen as shown in Figure 10.13.

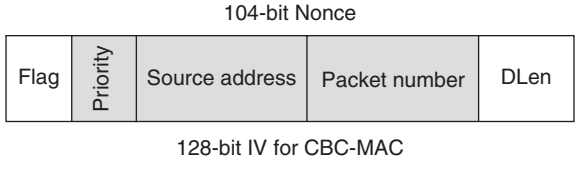


Figure 10.13: IV for AES CBC-MAC

The flag field has a fixed value of 01011001. The priority field is reserved for future use. The source MAC address is self explanatory and the packet number (PN) is as we discussed above. Finally, the last entity DLen indicates the data length of the plaintext. Note that the total length of the IV is 128 bits and the priority, source address and the packet number fields together also form the 104-bit nonce (shaded portion of Figure 10.13) which is required in the encryption process. The 128-bit IV forms the first block which is needed to start the CBC-MAC process described in Section 10.9.4. The CBC-MAC computation is done over the 802.11 header and the MPDU payload. This means that this integrity protection scheme also protects the source and the destination MAC address, the quality of service (QoS) traffic class and the data length. Integrity protecting the header along with the MPDU

payload protects against replay attacks. Note that the CBC-MAC process requires an exact number of blocks to operate on. If the length of the plaintext data cannot be divided into an exact number of blocks, the plaintext data needs to be padded for the purposes of MIC computation.

Once the MAC has been calculated and appended to the MPDU, it is now ready for encryption. It is important to re-emphasize that only the data-part and the MAC part of the packet are encrypted whereas the 802.11 header and the CCMP header are not encrypted. From section 10.9.3, we know that the AES-counter mode encryption process requires a key and a counter. The key is derived from the PTK as we discussed.

The counter is created by the concatenation of the following entities: Flag, Priority, Source MAC address, a packet number (PN) and Ctr as shown in Figure 10.14.

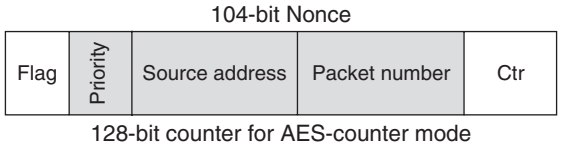


Figure 10.14: Counter for AES Counter Mode

Comparing Figure 10.14 with Figure 10.13, we see that the IV for the integrity process and the counter for the encryption process are identical except for the last sixteen bits. Whereas the IV has the last sixteen bits as the length of the plaintext, the counter has the last sixteen bits as Ctr. It is this Ctr which makes the counter a real "counter." The value of Ctr starts at one and counts up as the counter mode proceeds. Since the Ctr value is sixteen bits, this allows for up to 2^{16} (65,536) blocks of data in a MPDU. Given that AES uses 128-bit blocks, this means that an MPDU can be as long as 2^{23}, which is much more than what 802.11 allows, so the encryption process does not impose any additional restrictions on the length of the MPDU.

Even though CCMP succeeds in combining the encryption and integrity protocol in one process, it does so at some cost. First, the encryption of the various message blocks can no longer be carried out in parallel since CBC-MAC requires the output of the previous block to calculate the MAC for the current block. This slows down the protocol. Second, CBC-MAC requires the message to be broken into an exact number of blocks. This means that if the message cannot be broken into an exact number of blocks, we need to add padding bytes to it to do so. The padding technique has raised some security concerns among some cryptographers but no concrete deficiencies/attacks have been found against this protocol.

The details of the overall CCMP are shown in Figure 10.15 and finally the following table compares the WEP, WPA and WPA2 security architectures:

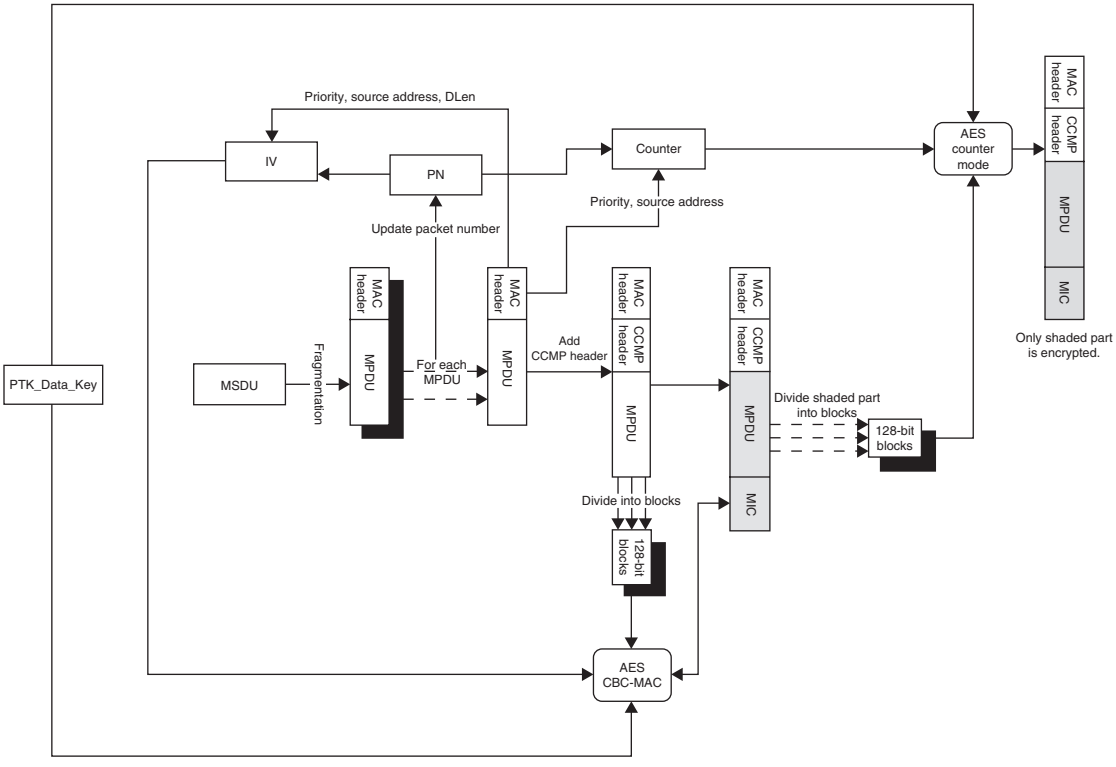


Figure 10.15: WPA2-The Complete Picture

WEP	WPA	WPA2
Relies on preshared a.k.a. out-of band key establishment mechanisms. Usually leads to manual configuration of keys and to key sharing among STAs in a BSS (often ESS).	Recommends 802.1X for authentication and key-establishment in enterprise deployments. Also supports preshared key establishment like WEP.	Same as WPA.
Uses a synchronous stream cipher which is unsuitable for the wireless medium.	Same as WEP.	Replaces a stream cipher (RC4) with a strong block cipher (AES).
Generates per-packet key by concatenating the IV directly to the master/preshared key thus exposing the base-key/master-key to attacks like FMS.	Solves this problem (a) by introducing the concept of PTK in the key hierarchy and (b) by using a key mixing function instead of simple concatenation to generate per-packet keys. This reduces the exposure of the master key.	Same as WPA.

WEP	WPA	WPA2
Static master key + Small size of IV + Method of per-packet key generation → Extremely limited key space.	Increases the IV size to 56 bits and uses only 48 of these bits reserving 8-bits to discard weak keys. Also, use of PTK which are generated afresh for each new session increases the effective key space.	Same as WPA.
Changing the IV with each packet is optional → key-reuse highly probable.	Explicitly specifies that both the transmitter and the receiver initialize the IV to zero whenever a new set of PTK is established[31] and then increment it by one for each packet it sends.	Same as WPA.
Linear algorithm (CRC-32) used for message integrity → Weak integrity protection.	Replaces the integrity check algorithm to use MICHAEL which is nonlinear. Also, specifies countermeasures for the case where MICHAEL may be violated.	Provides for stronger integrity protection using AES-based CCMP.
ICV does not protect the integrity of the 802.11 header → Susceptible to Redirection Attacks.	Extends the ICV computation to include the MAC source and destination address to protect against Redirection attacks.	Same as WPA.
No protection against replay attacks.	The use of IV as a sequence number provides replay protection.	Same as WPA.
No support for a STA to authenticate the network.	No explicit attempt to solve this problem but the recommended use of 802.1X could be used by the STA to authenticate the network.	Same as WPA.

[31] This usually happens every time the STA associates with an AP.

CHAPTER 11

Voice Over Wi-Fi and Other Wireless Technologies

Praphul Chandra
David A. Lide

11.1 Introduction

So far we have been discussing voice over Wi-Fi in the context of pure VoIP technology, running over 802.11a/b/g-based networks. In this chapter we will look at voice over Wi-Fi in conjunction with other wireless technologies, including proposed 802.11 extensions, cellular, WiMax (and other wireless broadband technologies), Bluetooth, and conventional cordless phone systems.

The field of telecommunications is in flux. Probably the most often-used word today in telecommunications is *convergence*. VoIP has been a big step towards the convergence of the voice (PSTN) and data (IP).

Till very recently, the C-word was limited to the discussion of wired networks and [cellular] wireless networks, and the discussions were dominated exclusively by voice-oriented wireless networks where IP had not make any significant inroads. With the emergence of Wi-Fi, IP has now gone wireless. Furthermore, the emergence of VoWi-Fi means that industry pundits are talking about the convergence of voice (GSM, 3G) and data (Wi-Fi, WiMax) in the wireless domain too. But what does all this mean? What will the wired and wireless networks of tomorrow look like? Which technology will "win"? We look at these issues in this chapter.

11.2 Ongoing 802.11 Standard Work

One characteristic of the 802.11 standard is the ever-present enhancement process. Some of these enhancements are just now coming into the market, but there are others still under specification. This section summarizes the ongoing work and the possible impact on a voice application.

Table 11.1 outlines the various 802.11x projects as a reference.

11.2.1 802.11n

The 802.11n project is working on techniques to (a) increase the user throughput of 802.11 to over 100 Mbps and (b) increase the range of communication. This will be accomplished

Table 11.1: Summary of 802.11 Projects

802.11 Project	Description	Status (as of January 2006)	Impact to Voice
.a	Up to 54 Mbps in the 5-GHz range. Introduces the OFDM modulation technique.	Ratified; products in the market.	802.11a is important because it adds more available channels and thus increases 802.11 voice-carrying capacity. Some impact to scanning and roaming algorithms.
.b	DSSS transmission rates in the 2-GHZ ISM band. This is the base standard.	Ratified.	
.c	Bridging operation. Not a standalone standard; work merged in to 802.1d.		N/A
.d	Global harmonization.	Ratified	
.e	Quality of Service	Ratified (Wi-Fi variants WMM and WMM-SA).	WMM provides two important features for the voice application: • Prioritization of voice over other traffic • Power save method that is useful for voice traffic
.f	Inter-AP protocols	Ratified, but never really accepted in industry.	N/A
.g	Enhancement of transmission rates to 54 Mbps.	Ratified.	Higher capacity means more voice calls per AP.
.h	Regulatory enhancements and spectrum management for .a.	Ratified.	Some impact to the scanning algorithm.
.i	Security enhancements.	Ratified. Wi-Fi variants: WPA and WPA2.	See Chapter 10.
.j	Japanese Regulatory—defines frequency requirements for 4.9-GHz and 5-GHz deployment in Japan.	Ratified.	No real impact.
.k	Measurement enhancements.	Under development.	Resource Management.
.l			No project.
.m	Editorial cleanup of specifications.	Ongoing.	N/A

(*Continued*)

Table 11.1: (continued)

802.11 Project	Description	Status (as of January 2006)	Impact to Voice
.n	Enhancement of transmission rates to wide bands throughputs.	Under development.	See text below.
.o			Not a project.
.p	Wireless access in a vehicular environment.	Under development.	See text below.
.q			Not a project.
.r	Enhanced roaming.	Under development.	
.s	Mesh networking.	Under development.	See text below.
.t	Wireless performance prediction and test.	Under development.	See text below.
.u	Internetworking with external networks.	Under development.	See text below.
.v	Wireless network management.	Under development.	This project is looking at ways for APs to actively control station radiosettings; for example, through network management protocols such as SNMP. It is closely tied with the 802.11k project.
.w	Security for management. frames	Under development.	This project is looking at extending 802.11i to protect management frames.
.x			Not a project (.x is used to refer to the entire 802.11 protocol umbrella).
.y	802.11 in the 3.65–3.7-GHz wavelengths (opened up in the US in July 2005).	Under development.	New spectrum means, of course, more potential capacity for voice traffic. There will also be a roaming/ scanning impact as per mixed 802.11a/b/g solution today. 802.11y will also include a standard mechanism to avoid spectrum interference with other users of the spectrum. This, in theory, will simplify opening up additional frequency bands in the future.
.z			No project.

through the combined use of several technologies. One of the main technologies is the use of multiple input, multiple output (MIMO) antennas for send and receive. With MIMO, the transmitter and receiver can take advantage of the inherent reflections (multipath) that occur during radio transmission instead of being negatively impacted. The sender and receiver essentially generate/receive multiple data streams (up to four, two typically) that are spatially separated. Using sophisticated mathematics, the receiver is able to recover the data streams. The advantage of this approach is that the data rate can be increased and the range enhanced because the receiver can recover the signal better under noisy conditions.

802.11n also includes an improved OFDM modulation technique that can handle wider bandwidth and coding rate.

Another technology is the use of wider channels: instead of 20-MHz channels for 802.11 a/b/g, 802.11n can make use of 40-MHz channels. This is an optional feature.

Packet aggregation and bursting techniques along with a more sophisticated block acknowledgment scheme and a smaller interframe spacing (RIFS) are also used to minimize the overhead per data packet transmitted. Aggregation is especially important in mixed mode, where 802.11n devices must coexist with legacy 802.11b/g devices. 802.11n also defines an optional mode (Greenfield) where legacy support is not required. It is in this mode where the maximum data rates will be achievable (in mixed mode, for example, an 802.11n device will at a minimum need to bracket each high throughput with a legacy transmission such as RTS/CTS to clear the medium).

The standard will also include enhanced power-management techniques in addition to the U-APSD method we discussed in the previous chapter. One important, mandatory technique, referred to as MIMO power save, allows the more power-intensive MIMO mode to be disabled for transmissions that do not require the high throughput—i.e., voice. Without this, MIMO operations, at least for initial chip sets, will be more costly in terms of battery life than mature 802.11b/g solutions.

A second, optional technique, power save multi poll (PSMP), uses a microscheduling technique to manage the channel efficiently. PSMP comes in two flavors, scheduled and unscheduled. Scheduled PSPM works in conjunction with scheduled access 802.11e. At a high level, an 802.11n Wi-Fi phone will create a scheduled PSMP traffic stream, via a TSPEC, at the start of a call. A PSMP frame will be sent by the AP according to the service period (i.e., packetization period used in the call), and will be synchronized to the scheduled power save (S-APSD) interval that is being used. PSMP frames are management frames that are sent to a broadcast destination. Each PSMP frame contains multiple station-information fields, where a particular field gives a station ID (AID) and time offsets when the station is allowed to send and receive and for how long. In effect, these frames are minischedules for the medium (up to 8 ms at a time). The Wi-Fi phone will wake up to receive the PSMP frame and then know when to send/receive its voice packets.

With unscheduled PSMP, the 802.11n Wi-Fi phone will use U-APSD. The U-APSD configuration may or may not have been set up via a TSPEC. In either case, the receipt of a trigger frame will cause the AP to issue a PSMP frame to schedule the transmission of queued frames from its delivery-enabled access category queues. The PSMP will also schedule when the Wi-Fi phone will issue acknowledgments.

To further improve both performance and power saving during PSMP, a new acknowledgment scheme known as Multiple TID Block ACK (MTBA) is used. This ACK scheme allows ACKs (a) to be delayed, and (b) combined together so that all packets received in the PSMP period can be acknowledged with one acknowledgment.

The impact to the voice-over-Wi-Fi application of 802.11n is obvious. Greater data throughput means more calls can be carried per access point and voice/data can be more easily mixed on the same network. One issue with 802.11n handsets will be reduced battery life due to the higher power requirements for 802.11n functions. The MIMO power-save mode will hopefully alleviate some of this and make the power consumption for 802.11n VoWi-Fi devices comparable to today's 802.11b/g solutions.

11.2.2 802.11p

This IEEE project is looking at the changes necessary for 802.11 to operate in the 5.9-GHz licensed intelligent transportation system band (ITS). The focus is on vehicular applications such as toll collection, safety, and commerce. One aspect of the project that is relevant to voice is the requirement for fast (i.e., vehicle speed) handoffs. The approaches being defi ned in this project may impact the fast roaming work done in 802.11r and vice versa.

11.2.3 802.11s

The purpose of this project is to standardize the protocols for an 802.11-based wireless distribution system (WDS). The proposed architecture for this system is a meshed network that is wirelessly connected using 802.11 protocols. Nodes on the network will dynamically discover neighbors, and enhanced routing protocols will be present to facilitate efficient packet routing. 802.11s compliant networks will be self-organizing.

A standardized wireless distribution system for 802.11 will increase the coverage capability of 802.11 networks, thus making the applicability of 802.11 voice services even more compelling. Instead of being restricted to disjoint islands of Wi-Fi hot spots, one can envision a 802.11s-based mesh network covering large (cell-phone scale) spaces. However, the 802.11s protocols will need to ensure that the quality of service and security concerns for voice that we have discussed earlier are addressed. Furthermore, the per-[voice] packet latency introduced by a mesh topology will have an impact on the overall quality of service that is achievable.

11.2.4 802.11t

The 802.11t project is concerned with performance measurement. This project will result in recommendations (not standards) in the areas of measurement and test techniques and metrics to be tracked. This project and a related Wi-Fi voice test project are interesting in that they recognize that voice-over-802.11 performance test requirements are very different from traditional data performance. Data-performance testing is mostly concerned with throughput, and the metrics of interest are maximum bits per second that can be transmitted or received under various conditions (such as data packet size).

Voice performance, however, is more concerned with the packet loss, latency and jitter that are present under overall BSSS loading. Artificial test and measurement is a difficult problem because to really assess the impact of all three performance areas on a voice call will require incorporating the techniques used in VoIP to mitigate network impacts. For example, VoIP terminals will use a jitter buffer to handle voice packet interarrival variance, and will utilize packet-loss concealment techniques to recover from lost packets. These will need to be factored into the measurement process.

Power management and battery life are important 802.11 phone metrics. These are impacted heavily by access-point performance such as beacon interval stability, and ps-poll/null-frame response times. Test methodologies need to be developed for these and other power-related metrics.

A final area of voice-unique testing is in the area of roaming. A standardized test methodology to measuring roaming and handoff performance is a highly desirable product of this project.

11.2.5 802.11u

The 802.11u project is aimed at adding features to 802.11 that facilitate interoperation in multiple 802.11 network environments. These features include network enrollment, network selection and service advertisement. In general, the results of this project should facilitate the Type C roaming. 802.11u will have some overlap with another IEEE project, 802.21. We will discuss this further below.

One important feature being addressed in the 802.11u project is the handling of emergency calling. One proposal being considered is to use a spare bit in the existing TSPEC element to indicate that the requested resources are to be used for an emergency call. A phone could then issue an ADDTS message to the AP with this "emergency service" bit set when a "911" call was placed.

11.3 Wi-Fi and Cellular Networks

Wi-Fi/802.11 is, at its basic level, a radio technology. This section will examine how voice over Wi-Fi will interact with the current "reigning" champion of voice/radio technology: today's cellular networks.

Wi-Fi and cellular are for the most part complementary technologies. They individually provide solutions to a subset of the wireless space. In particular, Wi-Fi can be used to provide coverage in areas where cellular is less effective: indoors, hospitals, airports, and urban canyons.

The inclusion of a voice-over-Wi-Fi capability into a cell-phone handset can be viewed as the ultimate goal for voice over Wi-Fi. There are several reasons for this goal. An obvious one is economy of scale. Economically, a voice-over-Wi-Fi implementation (chip set/software solution) will reap immense benefits from deployment in the 500-million plus cellular-handset market. Such volumes allow for the research and development necessary to create new classes of system on a chip specifically tailored to meet conventional cellular and Wi-Fi requirements. We should expect chip sets in future that include Wi-Fi radios, MAC/Baseband integrated with cellular modems and processors. With this integration comes a reduction in cost that will surely spill over into pure voice-over-Wi-Fi space as well.

Secondly, like its parent technology VoIP, voice over Wi-Fi will piggyback on the ongoing improvements to the data networks that are being upgraded to provide enhanced data services to the basic cell phone. Examples of this include the 3G data networks that have been coming online in the past five years.

A third factor is the usefulness of Wi-Fi technology to augment areas where cellular technology is lacking. Specifically, areas where cellular coverage is problematic (in doors, airports, hospitals, urban canyons) can be handled by overlapping Wi-Fi networks. The introduction of Wi-Fi-based mesh networks, driven by the maturation of 802.11s, will contribute to this trend. For example, vendor studies have shown that a city-wide, Wi-Fi mesh network can be a more cost-effective approach to providing wireless coverage than deploying 3G (1xEV-DO). A Wi-Fi mesh network could be, for example, situated in street lamp posts (which can be rented cheaply from the city government) as opposed to a cell tower that would require space from an office building or home.

A fourth factor is bandwidth. The cellular network providers would love to have the ability to move customers off their precious spectrum wherever possible. For example, when in your broadband-enabled home, why not let the cell phone make calls over the IP-based broadband network, via your in-home Wi-Fi infrastructure? A side effect of this is that cellular providers, riding on top of broadband access, now have a means to get customer phone minutes when he is at home. By providing an integrated access point and VoIP gateway equipment that allows the customer's conventional home telephony (i.e., POTs phones) to place VoIP calls back to the cellular base network, just like with his dual-mode cell phone, the cellular providers can cut into the traditional home voice-service monopoly held by the local telephone companies. This is one of the key drivers of fixed/mobile convergence, with the goal being to get customers to sign up to an all-inclusive home and mobile phone service from the cellular providers.

A final factor is the cellular world trend to move to using an SIP-based call-signaling protocol known as IP Multimedia Subsystems, or IMS. We will discuss IMS further below. The use of SIP to control cell-phone call signaling as well as voice-over-Wi-Fi signaling makes it easy (relatively speaking) to architect a unified phone with seamless handoffs between the cell world and the Wi-Fi world.

11.3.1 Dual-Mode Issues

There are several issues, however, to overcome before the dual-mode, cellular and Wi-Fi phones become a reality. These include:

- Handoffs between the two networks. This is especially a problem if the existing cellular signaling mechanisms are used while on the cellular network and VoIP signaling is used when in Wi-Fi mode. One approach is to simply not allow switchover while a call is in progress. Another is to use the same signaling protocol for both networks. We will look at this case further below.
- Billing. Two networks means two billing systems, assuming that the Wi-Fi portion is not free.
- Phone integration. Integration of a dual-mode phone is a nontrivial exercise. One area of difficulty is the reuse of key hardware and software components. For example, today's cell phone utilizes highly optimized systems on a chip, including possibly accelerator hardware for audio codecs, echo cancellation and other number-crunching algorithms that are executed on the voice samples. These may be highly integrated with the cellular voice processing, so reusing them when in voice-over-Wi-Fi mode may be difficult.
- Power management. Today's cell phones achieve their battery life levels through a combination of power-efficient hardware and power-aware protocols. For example, the cell-phone voice-sample processing subsystem (codec, echo canceller, etc.) is closely tied to the cellular network "timeslot" so that the entire phone can wake up out of a low-power state only when needed to process, send and receive samples. A Wi-Fi phone, in contrast, does not have such a close coupling between the voice-processing subsystem and the actual Wi-Fi network. Also, a dual-mode phone will require Wi-Fi channel scanning (and its associated power requirements.
- Codec usage. In cellular telephony, the use of codec is very closely tied to the cellular network. For example the GSM-AMR codec rates match the cellular network transmission "time slots" exactly. Thus, an existing voice subsystem for a cellular phone may not be able to easily accommodate "standard" VoIP audio codecs such as G729ab and G723. While RTP profiles for the cellular codecs (GSM, EVRC) are defined for VoIP, not all VoIP devices will implement them due to their complexity and processing requirements. Thus, when in VoIP mode, a dual-mode handset may fail to negotiate a codec for a VoIP call, or may require a transcoder somewhere in the network.

11.3.2 Convergence Strategies

There are two basic strategies for Wi-Fi voice/cellular convergence with several variations being proposed or prototyped. The first basic strategy, an example of which is being proposed for GSM networks, is an approach where the lower-layer cellular protocols are replaced with IP. The higher-layer cellular network protocols are then tunneled over the IP network. A gateway function at the border between the IP network and cellular backbone is provided to terminate the tunnel. In the case of GSM, this approach is known as Unlicensed Mobile Access (UMA).

The second strategy, being proposed first for CDMA networks but also applicable to GSM networks, is using IMS as a common signaling protocol for both the pure cellular network signaling and the voice-over-Wi-Fi network.

11.3.2.1 UMA

Unlicensed mobile access, as defined in UMA Architecture (Stage 2) R1.0.43 (2005-425-298), is "an extension of GSM/GPRS mobile services into the customer's premises that is achieved by tunneling certain GSM/GPRS protocols between the customer's premises and the Core Network over a broadband IP network, and relaying them through an unlicensed radio link inside the customer's premises."

Under UMA, a dual-mode GSM/Wi-Fi handset uses the GSM network when it is available and no Wi-Fi network is present. When a suitable Wi-Fi network comes into range, however, the phone will switch over to using voice over Wi-Fi. As defined, UMA is not Wi-Fi-specific and is designed, in theory, to use any unlicensed access technology. Wi-Fi and Bluetooth are called out as initial candidates in the UMA Stage 2 specification.

In the case of a UMA/Wi-Fi-capable handset, GSM call signaling and voice compression will be used in both the GSM cell network and the Wi-Fi network. When using the GSM spectrum (referred to as GERAN/UTRAN mode), the phone operates as a pure GSM phone with some additional functionality present to enable GSM to do Wi-Fi roaming. We will discuss this additional component a little later. Once a switch to a Wi-Fi network has taken place, the phone will use Wi-Fi to access the Internet and will set up a secure tunnel back to the GSM core network (this is referred to as UMAN mode in the specification). Over this tunnel, pure GSM signaling messages and RTP encapsulated media packets will be sent and received. This is illustrated in Figure 11.1.

This tunnel is terminated at a secure gateway (SGW) component of a special network device known as the UMA network controller, or UNC. The UNC/SGW sits at the border between the GSM core network and the Internet, and acts as a gateway between the VoIP and GSM world. On the VoIP side, the UNC looks like the endpoint of a TCP/IP connection. On the GSM side, the UNC looks like a GSM base station. The UNC routes data (call-signaling messages or media packets) between the IP network and GSM networks. There are two kinds

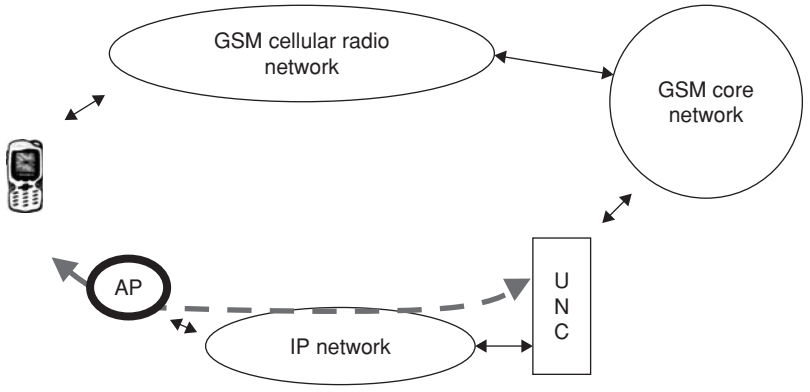


Figure 11.1: UMA Overview

of UNCs: provisioning UNCs and serving UNCs. The provisioning UNC is used for initial phone bring-up, and will typically redirect the phone to a serving UNC. The FDQN of the provisioning UNC and its associated secure gateway will be typically provisioned into the dual-mode phone.

The tunnel between the dual-mode phone and the UNC will be secured using the standard internet security protocol, IPsec. The specification calls for the use of the IPsec Encapsulating Security Protocol (ESP) in tunnel mode with AES encryption (Cipher Block Chaining Mode), and SHA1 authentication. Figure 11.2 illustrates the protocol layering of UMA for voice and signaling, respectively.

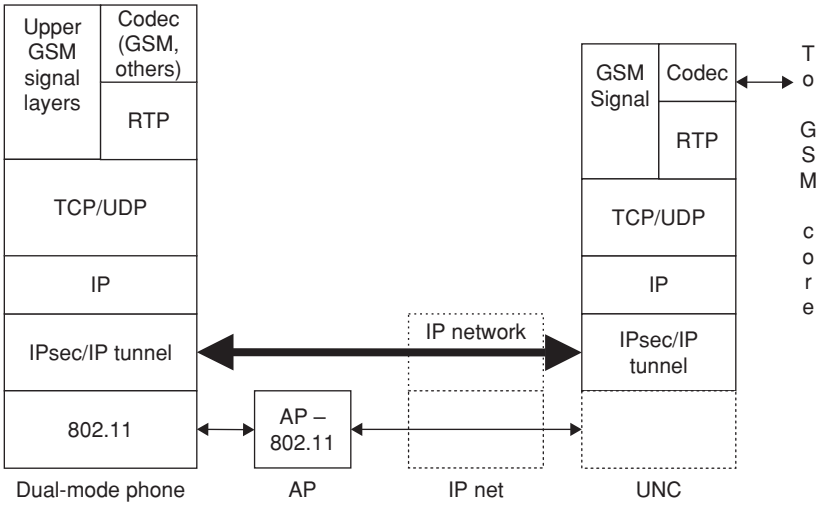


Figure 11.2: UMA Signaling Secure Tunneling

The IPsec security association is set up via the IKE key-management protocol (v2). UMA defines two IKE profiles, one using EAP-SIM and one using EAP-AKA (authentication and key agreement). Both these profiles allow for fast reauthentication. This is useful to reduce the workload due to full IKE v2 handshaking and to speed up the registration process, especially if the dual-mode, UMA phone has roamed onto a new Wi-Fi network (where its assigned IP address has been changed).

It is important to note that this secure tunnel is used for all data between the dual-mode phone and the UNC, including voice traffic. We will have more to say about the bandwidth efficiency of this scheme a little later.

The first operation after setting up the tunnel is to register. This may take several steps and additional tunnels, especially if this is the first time that the phone has booted because of the serving UNC discovery procedure. The procedure consists of the following steps:

- Discovery of the default serving UNC. This will be provided by the provisioning UNC.
- Registration with the default serving UNC.
- The default serving UNC may accept the registration, or may redirect the phone to use a different serving UNC.
- In the latter case, the registration will be repeated to the new serving UNC. This will involve setting up another secure tunnel.
- The registration is completed by the dual-mode phone sending a Register Accept message.

The FDQN of the serving UNC/SGW can be saved for subsequent reboots along with the AP BSSID. On subsequent reboots, the phone can attempt to register directly with the serving UNC that it had previously used when connected to this AP. Note that the phone can be redirected to a different serving UNC so that the discover procedure can take place at any time. Also the discovery procedure may be necessary if the phone roams to a new Wi-Fi network where it has not previously operated.

Once the secure tunnel is in place, and the dual-mode phone has registered with the UNC, the dual-mode phone can use the tunnel to transport the higher layers of the GSM call-signaling protocols.

The UNC gateway function translates between the GSM core network and the secure tunnel. The tunnel is also used for the media traffic flow. While VoIP allows for a variety of codecs to be used, UMA uses the cellular standard GSM AMR or WB-AMR codecs (RFC 3267). This is preferred since a call might have roamed from or may in future roam to the cellular network.

The codec samples are RTP encapsulated before being transmitted through the secure tunnel. One comment on this is that use of IPsec and secure tunneling introduces substantial overhead to each voice packet. Given that a 20-ms GSM voice frame (at the full AMR rate of 12.2 kbps) will contain 244 bits of speech payload plus 12 bits of frame overhead for a total of 32 bytes, we can compute that a UMA-secure/tunneled RTP packet will effectively be 126 bytes plus layer 2 headers (see Table 11.2).

Table 11.2: Effective GSM RTP Packet Size in UMA Tunnel (with AES Encryption in CBC Mode, SHA1 Authentication)

Packet Element	Size (bits)
Frame payload (12.2 kbps rate)	244
CRM	4
Table of contents	6
RTP padding	2
RTP header	96
UDP/IP	224
2nd IP header (tunnel)	160
IPsec ESP header	32
IV	128
Padding	0
Trailer	16
Authentication	96
Total	**1008**
% Overhead	**321%**

A dual-mode, UMA phone can be set up in one of four preferences:

1. GSM only (i.e., never use the UMAN mode of operation).
2. GSM preferred (i.e., use GERAN/UTRAN mode where possible, switching to UMAN mode only when the GSM network is not available).
3. UMAN preferred (i.e., use voice over Wi-Fi where possible).
4. UMAN only (i.e., switch to UMAN mode immediately after the phone starts up and registers on the GSM network).

The procedure to switch from GERAN/UTRAN mode to UMAN mode (GSM to Wi-Fi handover) is referred to as "rove in" and the reverse procedure is referred to as "rove out." Unlike the inter-802.11 roaming, these two roaming procedures are of the "make before break" type, meaning that the new mode must be fully established before the old mode is

disconnected. This is important because, as we have seen before, there are multiple protocol steps in the secure tunnel and registration procedures before voice can actually be delivered to/from the handset. UMA has the goal of seamless switchover between the two modes. It will be interesting to see if this achievable in practice. One problem area will be the delay introduced by the RTP jitter buffer when in UTRAN mode. When in GERAN mode, the phone will be operating without any jitter buffer (this is not required due to the TDMA protocol used in the GSM network). As soon as the UTRAN mode is enabled, the initial RTP packets from the core network will need to be delayed in the handset so that the handset jitter-buffer can be primed. Thus, the user will potentially hear a gap in the conversation equal to the jitter buffer nominal setting. One way around this, potentially, is to begin with a shallow jitter buffer and let it adapt aggressively if network conditions require a deeper buffer. The corresponding jitter buffer in the UNC will need the same kind of work-around.

The UMA specifications include recommendations for the Wi-Fi network that a dual-mode, UMA phone will utilize. These recommendations include:

- Use of 802.11 security, WEP or WPA PSK.
- QoS: Use of WMM is recommended. The specification suggests "simulating" WMM if the AP does not support the feature, by using a "non-standard," smaller backoff window and interframe delay. The specification also calls for the dual-mode phone to use link level 802.1D priority markings and/or the IP layer TOS/DCSP value of received packets for outgoing packets.
- Power save: The specification calls for the use of 802.11 power save when not in an active call but, interestingly, states that the 802.11 power save should not be used during a call. Presumably this was written before the voice-friendly U-APSD power-save mode was defined by 802.11/Wi-Fi.
- Roaming and scanning: The specification calls for background scanning at an un-defined interval, "depending on power conservation strategies." The specification recommends the use of RSSI as the key metric to determine when to roam. The specification also states that Wi-Fi roaming is to be isolated from the upper layer protocols, unless the IP address has changed as a result of the roam. For inter-BSS roaming, UMA suggests a target of 100 ms for the device to switch to a new AP.
- From a hardware capability point of view, the specification requires the physical characteristics shown in Table 11.3.
- Finally, the specification recommends the use of "intelligent" packet loss concealment algorithms to mitigate packet loss.

On the AP side, the specification recommends a beacon period of 100 ms.

Table 11.3: Wi-Fi Physical Characteristics Required for UMA

Characteristic	Specification
Transmit power (at antenna)	+17 dBm
Receive sensitivity	−87 dBm @ 1 Mbps
Antenna gain	−0 dBi

UMA, when operating in UMAN mode, has provisions for some amount of RTP session negotiation. UMA allows the following call parameters to be "negotiated," via information elements in the tunneled signaling packets:

- RTP port (UDP)
- RTCP port (UDP)
- Sample size (VoIP packetization period)
- Redundancy/mode tables (see below)
- Initial GSM codec mode to use
- RTP dynamic payload type to use for the audio codec

These parameters can also be changed mid-call through tunneled signaling messages.

The GSM codec has a built-in redundancy mode that is applicable to transport over Wi-Fi and over IP in general. UMA has provisions to take advantage of this feature. The feature works as follows: GSM inherently supports various modes of operation (or bit rates), ranging from 4.72 kbps to 12 kbps for the narrowband AMR codec (additional rates are available in the wideband, WB-AMR codec). Any of these rates can be used in a call, and the RTP packing format for GSM AMR includes a field (Codec Mode Request or CMR) with which a receiver can signal the other side that he desires a codec rate change. Furthermore, an RTP GSM AMR packet may contain redundancy in the form of copies of previously transmitted GSM frames. This is referred to as forward error correction. UMA allows a redundancy/mode table to be exchanged via information elements in the tunneled GSM/UMA call-signaling packets. This table gives, for each rate, the desired redundancy level (UMA restricts the options to none, one level or two-level). A separate table can also be present that defines, for each mode, the frame-loss rate threshold and a hysteresis level to control when a receiver should try to switch rates. Armed with these tables, a UMA handset can monitor the frame loss it is seeing and, when configured loss thresholds are hit, the handset can use the CMR field to request a rate/redundancy change. Similarly, the receiver in the UNC can do the same. As a note, it is unlikely that a rate change alone will accomplish much when in UMAN mode because of the packet protocol and security overhead mentioned above. However, the switch to a lower bit rate with redundancy has the effect of protecting for packet loss without increasing the overall RTP packet size. This is an important consideration for Wi-Fi QoS networks with admission control and also has a slight impact on power consumption when operating in the Wi-Fi network.

The GSM codec also has a feature known as unequal bit error detection and protection. This is accomplished by organizing the codec payload bits into three classes: A, B and C. Class A bits are the most important bits, Class B are next, and Class C are the least important. For example, in the GSM AMR 12.2-kbps rate, 81 bits out of the total 244 bits in a 20-ms frame are deemed Class A. Thus, in theory, a received packet with corruption in the Class B or C area of the payload could still be used (and not completely dropped). This scheme is, of course, very useful in GSM cellular networks where the packet integrity checks are adjusted to reflect the payload bit classes. Unfortunately, when operating in UMAN mode over a Wi-Fi/IP network, this codec feature is not applicable (although it would be beneficial). The problem is that, first, a UDP checksum covers the entire UDP packet so that UDP checksums would need to be completely disabled for the scheme to work. Even more damaging is the use of the IPsec-protected tunnel. IPsec performs a message authentication check across the entire payload. Thus, bit errors in Class B and C areas would result in the packet being dropped due to authentication failures. Finally, and most damaging, the 802.11 link-level security authentication checks (if enabled) would also fail for the same reason. Using this feature would require "application" knowledge to be propagated down to all layers of the protocol stacks; clearly this is not a feasible approach, at least for the near future.

A final area of interest with UMA is its handling of 911 emergency calls. A UMA dual-mode phone can first be configured as part of the registration process as to which network is preferred to make emergency calls. Secondly, the UMA call-setup message includes an information element to indicate the type of call. "Emergency" is an option in this IE. Finally, UMA has several options for managing location information:

- UMA handsets can send the identifier of their attached AP. The UNC can then use this to "look up" the APs location when an emergency call is placed. This, of course, is only 100% accurate if every possible AP (that UMA allows—UMA has provisions to restrict the APs which the dual-mode phones can use to obtain UMA service) has been registered so that its location is in a back-end database.
- UMA handsets can send their own location if they know this from other means—e.g., GPS.

11.3.2.2 IMS

The second approach to dual-mode telephony is the use of VoIP in both the cellular and Wi-Fi modes, via IMS. The IP Multimedia Subsystem is a key element of the third generation (3G) architecture. 3G is, briefly, is a collaboration of various standards bodies to define the next (third) generation cellular networks. 3G is a unification of the cellular world and the Internet, and IMS is the mechanism that enables IP-level services.

IMS is based on SIP. As we have discussed earlier, SIP is now the main VoIP call-signaling protocol. The reasons for changing from conventional cellular signaling to a system based on

SIP are beyond the scope of this book. The driving factor is unification of services, with the idea being that IMS-based signaling can facilitate the deployment of voice, video, presence and other services. However, the use of SIP/IMS and VoIP over the existing cellular networks has some interesting technical challenges that, if solved, will play into a pure voice-over-Wi-Fi scenario as well.

One issue is bandwidth. Today's cellular networks are constrained as to the amount of bandwidth available for a voice call. SIP-based VoIP call signaling utilizes a relatively inefficient, text-based protocol to communicate call signaling information. Furthermore, if you look at a VoIP media packet, a good portion of this packet will be composed of packet header information. Thus, IMS signaling and media will require more bandwidth than the current cellular protocols.

In the case of SIP messages, one approach is to use compression techniques such as those defined in RFC 3320. With RFC 3320, a layer that performs lossless compression (e.g., gzip) can be inserted between the application (voice-signaling SIP stack) and the network protocol stack (TCP/IP). On transmission, this layer can run a native implementation of a compression algorithm. However, on reception, this layer makes use of the Universal Decompressor Virtual Machine (UDVM), which is essentially a JAVA-like virtual machine tailored specifically for decompression operations. The instructions or "byte-codes" to be executed are provided by the sender. The advantage of this approach is that the actual compression algorithm can be controlled entirely by the transmit side; it can pick any algorithm desired, perform the compression on a message, and send it along with the UDVM instructions on how to decompress it (the bytes codes would only need to be sent with the first message, assuming the same algorithm is used throughout the signaling session).

To tackle the problem with the media packet protocol overhead introduced by the RTP/UDP/IP headers, an approach is to use a technique called robust header compression (RHOC- RFC 4362). RHOC can work across a link layer (e.g., in 802.11 between the phone and AP). The basic idea behind RHOC and its related header-compression protocols is to define at the start of a packet flow (e.g., at call setup), which fields in the packet headers are static, which fields update according to simple rules (e.g., the RTP timestamp), and which fields need to be sent along with each packet. Once these are set up, the static fields and those that change in a simple way can be stripped before packet transmission. The receiver will then reconstruct the complete header before forwarding the packet. The protocols include mechanisms to recover from delivery problems—for example, if a burst of packets is lost for some reason, the preset header information may need to be changed.

11.4 WiMax

WiMax is a new wireless technology, defined by IEEE 802.16x standardts. The core standard, 802.16, defines protocols for a broadband wireless infrastructure, operating in the 10–66 GHz

frequency range. The basic topology defined in the specification is point-to-multipoint. The targeted data throughput range was 70 Mbits, with a peak rate of 268 Mbps and a typical cell radius of 1–3 miles. This base standard was subsequently enhanced with a suite of amendments known as 802.16a. These added considerations for more spectrum bands (licensed and unlicensed), support for non-line-of-sight architectures, new physical-layer specifications and enhancements to the MAC layer. These later changes included consideration for quality of service and different types of traffic, including voice.

A second version of WiMax is currently being defined. This version, based on the 802.16e specification, is addressing mobility and roaming considerations. It will include support for hard and soft handoffs and improved power-saving techniques. It introduces a new PHY layer optimized for mobility.

Like 802.11, the 802.16 specifications include multiple physical layers. 802.16a defines three protocols:

- A single-carrier modulation format.
- Orthogonal Frequency Division Multiplexing (OFDM), with a 256-point transform. This is the same modulation technique used in 802.11g.
- Multiuser OFDM (OFDMA), with a 2048-point transform.

802.16 adds a new variant of OFDMA, referred to as SOFDMA (the "S" stands for scalable). The variant provides better performance for multiple users under varying conditions.

The media access layer for 802.16 is quite a bit different than for 802.11. It is based closely on the data-over-cable specification (DOCSIS). The relationship between WiMax and Wi-Fi is still to be defined. The conventional school of thought is that WiMax will become a "lastmile" technology, providing an alternative for the currently deployed broadband technology (i.e., DSL, cable, fiber-to-the-home, etc.). A WiMax CPE device, for example, could contain an 802.11 subsystem as well, just like today's cable and DSL broadband routers. With this architecture, a voice-over-Wi-Fi phone would be another subscriber to the WiMax backbone.

802.16e, with its support for mobility, muddies the waters. Conceivably, 802.16e could be used as a replacement for Wi-Fi in some environments.

11.5 VoWi-Fi and Bluetooth

Bluetooth (BT) is radio technology geared at the 2.4–2.5835-GHz ISM unlicensed frequency band just like Wi-Fi (802.11 b/g). Table 11.4 summarizes 802.11/Wi-Fi and Bluetooth technology.

There are three classes of BT devices, each with a different maximum output/power and corresponding range profile. These are summarized in Table 11.5.

Table 11.4: Wi-Fi (b/g) / Bluetooth Comparison

Wi-Fi	Bluetooth
Direct Sequence Spread Spectrum (DSSS).	Frequency Hop Spread Spectrum (FHSS).
Use only 22 MHz × 3 (channel) = 66 MHz.	Use 1 MHz × 79 (channel) = 79 MHz. The hop rate is 1600 hops/sec.
Power: 1 W (30 dBm).	Power: 1 to 100 mW.
Data rate: up to 56 Mbps at close ranges, 5.9 Mbps < 175 ft, 5.5 Mbps at 250 ft.	Max data rate 550 kbps at 250 ft.
Range up to 100 meters, depends on power and environment.	Power range: 100 meters (class 1), 10 meters (class 2), 10 cm (class 3).
Each Wi-Fi network uses 1 channel, max 3 nonoverlapping networks.	FCC requires BT devices to hop ⩾ 75 channels up to max 79 channels.
Defined only layer-2 protocols, a common way to access Internet through the AP.	Define different layers of protocols. Profiles allow for different voice and data application interworking.
Security is in the layer 2 (WEP, WAP, WAP2, etc.).	Security is in layer 2 (LMP) and security architecture for different layers.
Allow one AP and many stations to bind.	Allow one pair of AG (Audio Gateway) and handset or hands-free to pair.

Table 11.5: BT Device Power Classes

Power Class	Max Output Power (mW)	Maximum Range (meters)
Class 1	100 mW	100 m
Class 2	2.5 mW	10 m
Class 3	1 mW	10 cm

The current uses of BT (at least in class 2 and 3) make it a complementary technology to Wi-Fi. In the context of a VoWi-Fi phone, a BT subsystem might be present to provide low-rate, close-proximity wireless access to peripherals. The most likely scenario is where a VoWi-Fi phone would have a BT subsystem to allow the use of a BT handset or handsfree device as the end audio transducer.

This leads us to the main issue with BT and Wi-Fi: coexistence. In a nutshell, the BT physical layer utilizes a frequency-hopping technique that unfortunately can cause interference in an 802.11 b/g network (and vice versa). As described earlier, Wi-Fi/802.11 b/g standards in North America divide the ISM band into 11 overlapping channels (In Europe and Japan additional channels may be present). Only three channels—1, 6, 11—are nonoverlapping. Each channel utilizes 22 MHz of the ISM band; thus 3 × 22 = 66 MHz out of the 88.35-MHz ISM band will be occupied by a fully loaded Wi-Fi deployment.

Bluetooth, on the other hand, uses a frequency-hopping technique across almost the entire ISM band. Each hop frequency is 1 MHz and up to 79 channels are allowed. Furthermore, BT specifies a hop rate of 1600 hops/sec. This means that transmission from a BT device will definitely overlap with Wi-Fi transmissions if it is in range and the Wi-Fi transmission is long enough. This is shown graphically in Figure 11.3.

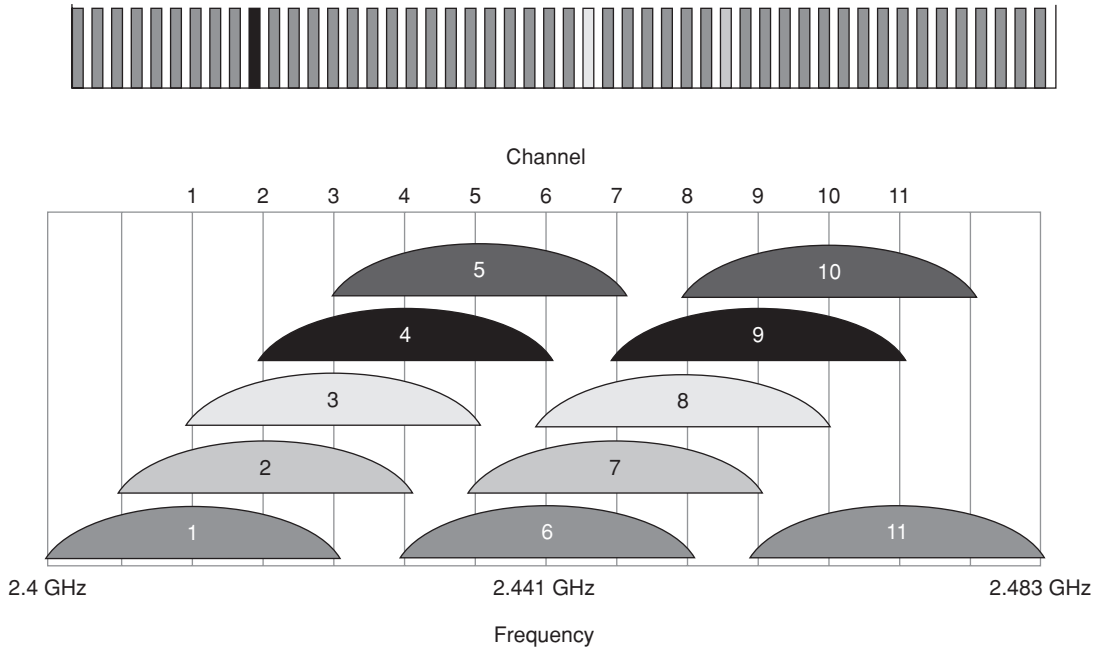


Figure 11.3: 802.11 b/g Frequency Bands

The impact of the interference from BT devices on Wi-Fi equipment is to effectively raise the Wi-Fi channel bit-error rate. A BT device that wants to transmit will be unaware of Wi-Fi activity and will not delay its transmission. If it is in range and its frequency-hopping scheme happens to overlap the Wi-Fi transmission, the Wi-Fi receiver will see a degraded signal and can either miss the packet or detect a CRC error. In either case, the Wi-Fi transmitter will need to resend.

Packet retransmission in Wi-Fi typically will lead to a reduction of transmission rate, as the Wi-Fi devices attempt to react to what they perceive as a noisy environment. Thus, data that would be normally transmitted at 54 Mbps may eventually be transmitted at a rate of 11 Mbps. This reduces the overall throughput of the Wi-Fi network and has other side effects for voice such as increased latency and power consumption.

Figure 11.4 conceptually illustrates the effect of a BT transmitter on Wi-Fi throughput. The figure plots Wi-Fi throughput versus received signal strength for the cases where the BT

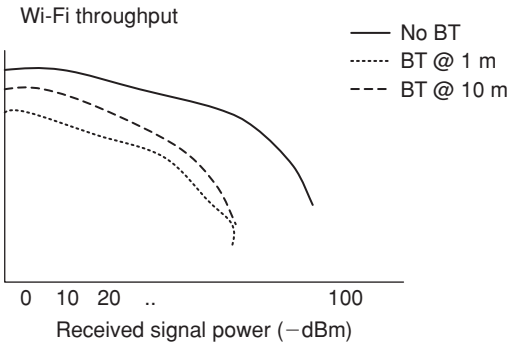


Figure 11.4: Wi-Fi Throughput vs. Received Signal Strength (AP-STA distance)

device is transmitting or not. Received signal strength here is used as a generalization of distance between the Wi-Fi device and its access point. The net effect of a BT transmitter in close proximity is to sharply degrade throughput, even when the Wi-Fi device is close to the AP. The degree of impact of the BT transmitter is correlated to the BT device location to the Wi-Fi device.

As an unpleasant side effect, Wi-Fi packets sent at the lower Wi-Fi rates will stay on the air longer and are hence even more likely to experience BT interference. Note that, because of the frequency-hopping technique used in BT, a Wi-Fi transmitter may not be able to detect that a BT transmission is in progress and back off. This in contrast to the case of Wi-Fi channel overlap; in this situation a Wi-Fi device will be more likely to detect the 802.11 energy and can then back off.

Before discussing ways for BT and Wi-Fi to coexist, we need to discuss the types of BT connections that can be used. BT has two types of link-level protocols: asynchronous connectionless (also known as ACL), and synchronous connection-oriented (SCO). ACL BT protocols are typically low rate and allow for packet retransmission. SCO connections are higher speed and (in BT 1.0 devices) do not allow for packet retransmission. SCO connections are used, for example, to communicate to BT headsets and hands-free devices. Different coexistence techniques are required for each of these, depending on the type of Wi-Fi traffic. For a VoWi-Fi phone with an adjunct BT handset or hands-free device, we are interested in Wi-Fi voice coexistence with BT voice over an SCO connection.

Several coexistence schemes for BT and Wi-Fi are possible. We will discuss some of these below. It is important to recognize that a combination of schemes will be required for a full robust solution.

One technique, utilized by Wi-Fi and BT chipset providers such as Texas Instruments, is to provide silicon-level interfaces so that the two chipsets can collaborate to minimize interference. In the Texas Instruments solution, for example, its two chip sets share a coexistence interface

over which information on when transmission is taking place can be exchanged. If a BT transmission is going on, the Wi-Fi transmission can be delayed and vice versa. This approach works best for Wi-Fi data and BT data (i.e., ACL) coexistence, with a couple of limitations. It is not enough to solve Wi-Fi voice and BT voice (SCO) coexistence problems, however.

A second set of techniques comes from the 1.2 version of the BT standard. This update has taken steps to address the coexistence issue by incorporating two new features: adaptive frequency hopping, and an enhance SCO link protocol (ESCO).

The BT 1.2 adaptive frequency-hopping scheme allows a BT device that has knowledge of the 802.11 device that it is co-located with to adjust its frequency-hopping scheme accordingly.

For example, if a BT device knows that its 802.11 counterpart is operating on Channel 1, it can select frequencies out of the 802.11 channel 1 subband for its hopping sequence. Thus, the BT device would use 57 out of the possible 79 channels. Using this technique in practice for the case of VoWi-Fi and a BT headset/hands-free device has several issues to overcome:

- The BT 1.2 specification does not define how the BT device learns the channel use. Typically this would require a software interface between the Wi-Fi and BT chipset/device driver.
- The adaptive frequency-hopping scheme does not help as much in cases where all multiple 802.11 channels are in use, such as would be the case in an enterprise environment. The 1.2 compatible BT devices can skip around the 802.11 channel that the colocated Wi-Fi is actively using, but still may interfere with other Wi-Fi devices as the user roams.
- Furthermore, in a multiple AP environment the BT device will need to change its hopping sequence whenever the Wi-Fi device decides to roam. This will most likely result in an interruption in the BT data stream.
- There will be impact on the scanning techniques. For example, the use of unicast probes to discover new APs on other channels will be problematic.
- A final issue is that a combined Wi-Fi/Bluetooth device will have cost pressures to share a single antenna. The above techniques are appropriate if each subsystem has a dedicated antenna and there is a minimal degree of RF separation between the two. When the antenna is shared, it is unlikely that the frequency-hopping adjustment approach will be effective.

ESCO allows for higher speed and retransmission on the SCO links. This will improve the quality of the BT transmissions (e.g., voice to/from a BT handsfree device).

In short, coexistence for BT voice and VoWi-Fi is still an open technical challenge.

11.6 VoWi-Fi and DECT

We have left this topic near the end as it is perhaps the most controversial, especially to DECT proponents. Digital Enhanced Cordless Telecommunications is a popular wireless standard, mostly in Europe, that—it can be argued—provides a complete wireless telephony solution today. DECT works in the 1.9-GHz band (some versions are available in the 2.4-GHz band) and utilizes a time-division multiplexing approach to bandwidth allocation. It was primarily geared for cordless telephony. An overview of DECT is given in Table 11.6.

Table 11.6: DECT Summary

Characteristic	DECT
Frequency Band	1.9 GHz
Access Method	TDMA
Data Rate	2 Mbps – being expanded to 20 Mbps for data services
Range	50 meters indoors, 300 meters outdoors
Modulation	Gaussian Minimum Shift Keying (GMSK)
Voice Codecs	G726 (ADPCM) [32 kbps]
Voice Signaling	ISDN based
Handovers/Roaming	Built into protocol
Security	GSM based
Cost	Approx ½ compatible 802.11 solution
Battery Life	~12 hrs talk, 100+ standby

In this regard, DECT can be considered a competing technology to VoWi-Fi. Let's identify the advantages touted by DECT adherents (many of these are based on original, 802.11b voice-over-Wi-Fi implementations)

- Cost: DECT handsets are cheap! This is based partly on the maturity of the technology, so that highly integrated hardware solutions are available.
- Power: DECT was designed upfront to be power efficient. It utilizes a TDMA-based method. DECT receivers can shut off their radios until their time slot occurs.
- Handset performance: Again, because of the maturity of the DECT handset market, industrial-strength (shock, temperature, dust, etc.) equipment is available today.
- Handoffs: DECT was designed with handoffs in mind.
- Range: DECT has inherently better range than 802.11. 802.11 can extend range by adding repeaters or access points, but this adds to cost and has limitations based on the ISM channel bandwidth.

- Quality of Service: The original DECT objections to Wi-Fi QoS (or lack thereof) were based on 802.11b deployments.
- Security: Most DECT objections to Wi-Fi security are based on the WEP implementations. As we have seen, WPA and WPA2 have addressed these concerns. However, as we also have seen, the use of WPA and WPA2 authentication and key-distribution methods makes fast handovers more complex.

The bottom line is that, while DECT does have advantages over VoWLAN for pure telephony, these are due primarily to its inherent limitation of being primarily a telephony protocol. As a "telephony-first" protocol, DECT will naturally win in a phone-only environment. But if we add data to the mix, voice over Wi-Fi will be a more attractive solution. Wi-Fi is the clear winner for providing wireless data service. Voice over Wi-Fi, as it runs on top of the data network, will succeed just as pure voice over IP is.

The other issue with DECT is how it plays into a VoIP backbone. DECT uses a ADPCM codec between the handset and the base station. This requires ADPCM transcoding if another low bit-rate codec is to be used for the network portion call. Furthermore, DECT uses ISDN-based signaling between the base station and the phone. This will need to be translated into SIP VoIP signaling.

We can also look at DECT and Wi-Fi in another light, that of convergence. There are various projects underway to merge DECT and Wi-Fi together. One approach has been to integrate the upper layers of DECT with the 802.11 MAC and PHY.

11.7 VoWi-Fi and Other Ongoing 802.x Wireless Projects

In this section we will take a quick look at three ongoing IEEE wireless standards and their potential relationship with VoWi-Fi.

11.7.1 802.20

The mission of the 802.20 project (also referred to as Mobile Broadband Wireless Access or MBWA) is to "develop the specification for an efficient packet based air interface that is optimized for the transport of IP based service." There is a special emphasis on mobility in this project, with goals of handling subscribers moving at speeds of up to 155 miles per hour (e.g., for high-speed train service). By this definition, there is overlap somewhat with the goals of 802.16e. However, the scope of 802.20 is limited to below the 3.5-GHz band, while 802.16e covers additional spectrum. Also, 802.20 is targeting a much lower data rate (around 1 Mbps) than 802.16. As an IP-based service, 802.20 must deal with similar issues as 802.11 when used to carry VoIP. There is a lot of debate on how 802.20 and WiMax (802.16e) will evolve, since there is a great deal of overlap. It is possible that 802.20 will be restricted to the high-speed domain only.

11.7.2 802.21

The 802.21 is an interesting project with special relevance to the voice application. Its goal is to "develop standards to enable handover and interoperability between heterogeneous network types including both 802 and non 802 networks." In other words, the project is involved with standardizing the type C and D roaming. This is also referred to as Media Independent Handoff, or MIH. Among the topics that 802.21 is investigating is the definition of a common interface between various layer 2 (802.11, 802.16, etc.) and layer 3 to facilitate the roaming and handoff process.

The draft standard discusses the concept of link-level "triggers." These are link-level events that can be passed to layer 3 to provide link state information to the roaming decision process. Link-level triggers include such events as link up/down, link quality above or below a defined threshold, link QoS state, perceived link range or throughput, and even network cost. In a pure 802.11 network, the link-level triggers correspond to the VoWi-Fi device's roaming triggers, such as the RSSI, beacon miss rate, and retransmit rate. However, in 802.11, these triggers were not explicitly called out as such and their use is up to the device manufacturer. The 802.21 project attempts to define these and to provide a framework for their configuration and reporting.

One difference in the 802.21 framework is that it includes the idea that triggers can come from the remote side of the connection, as opposed to being generated solely by the local side. In an 802.21-enabled 802.11 network, for example, the AP would be able to use a layer-2 message to send a "suggestion" that the station roam.

Another aspect of 802.21 is support for the "make before break" concept. 802.11 today requires that a station disconnect from one AP before connecting to another ("break" before "make"). This approach has some drawbacks, especially for voice, because there will be a period of outage between the "break" and the subsequent "make." We can mitigate some of this latency in a pure BSS roaming situation through such techniques as WPA2 preauthentication. However, with dissimilar network roaming, the latency in security setup, IP address provisioning, etc. will be too long for the goal of seamless mobility.

802.21 also introduces the idea of a mobility-management service. This is a network-based service that mobile devices can register with to obtain information about other networks that could be roaming candidates. This is somewhat analogous to the proposed 802.11k AP list that we discussed in Chapter 8, but it covers not just APs, but also cellular base stations, WiMax head ends, and so forth.

11.7.3 802.22

The 802.22 project is working on how to use portions of the RF spectrum, currently allocated to television broadcasting, for carrying wireless data services. In particular, the UHF/VHF TV bands between 54 and 862 MHz are being targeted, both specific TV channels as well as

guard bands (white space). This standard is still under development, but it looks to be based on 802.11 protocols, possibly adding an additional PHY layer and enhancing the MAC layer to deal with longer range.

One interesting aspect of proposed 802.22 networks, also referred to as wireless regional area networks (or WRANs), is that they will utilize a new technology known as *cognitive radio*. The proposed ITU definition of this technology is: "a radio or system that senses and is aware of its operational environment and can dynamically, autonomously, and intelligently adapt its radio operating parameters." The basic idea is to allow the wireless nodes to manage the spectrum in a distributed fashion, by observing the environment.

11.8 Conclusion

This chapter has taken a look at the future of voice over Wi-Fi, and how it may evolve, coexist and interact with other wireless technologies.

References

[11.1] Camarillo, G. and M. Garcia-Martin, *The 3G IP Multimedia Subsystem*, John Wiley, 2004.

[11.2] Fixed, nomadic, portable and mobile applications for 802.16-2004 and 802.16e WiMAX networks, November 2005, prepared by Senza Fili Consulting on behalf of the WiMAX Forum.

[11.3] IEEE Standard 802.16: *A Technical Overview of the Wireless MAN™ Air Interface for Broadband Wireless Access.*

[11.4] Unlicensed Mobile Access (UMA) Protocols (Stage 3), R 1.0.4, 5/2/2005.

[11.5] "Global, Interoperable Broadband Wireless Networks: Extending WiMAX Technology to Mobility," *Intel Technology Journal*, August 20, 2004.

[11.6] "Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN," *Intel Technology Journal*, August 20, 2004.

[11.7] A Generalized model for Link Level Triggers, V. Gupta, et al., [802.21 Contribution] http://www.comsoc.org/oeb/Past_Presentations/CityWiFiMesh_Apr04.pdf

[11.8] RFC 3267 Real-Time Transport Protocol (RTP) Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs, J. Sjoberg et al., June 2002.

[11.9] RFC 4362 RObust Header Compression (ROHC): A Link-Layer Assisted Profile for IP/UDP/RTP, L.E. Jonsson et al., December 2005.

[11.10] RFC 3320—Signaling Compression (sigcomp), R. Price et al., January 2003.

This page intentionally left blank

CHAPTER 12

Mobile Ad Hoc Networks

Farid Dowla
Asis Nasipuri

A mobile ad hoc network, such as the one shown in Figure 12.1, is a collection of digital data terminals equipped with wireless transceivers that can communicate with one another without using any fixed networking infrastructure. Communication is maintained by the transmission of data packets over a common wireless channel. The absence of any fixed infrastructure, such as an array of base stations, make ad hoc networks radically different from other wireless LANs. Whereas communication from a mobile terminal in an "infrastructured" network, such as a cellular network, is always maintained with a fixed base station, a mobile terminal (node) in an ad hoc network can communicate directly with another node that is located within its radio transmission range. In order to transmit to a node that is located outside its radio range, data packets are relayed over a sequence of intermediate nodes using a store-and-forward "multihop" transmission principle. All nodes in an ad hoc network are required to relay packets on behalf of other nodes. Hence, a mobile ad hoc network is sometimes also called a multihop wireless network.

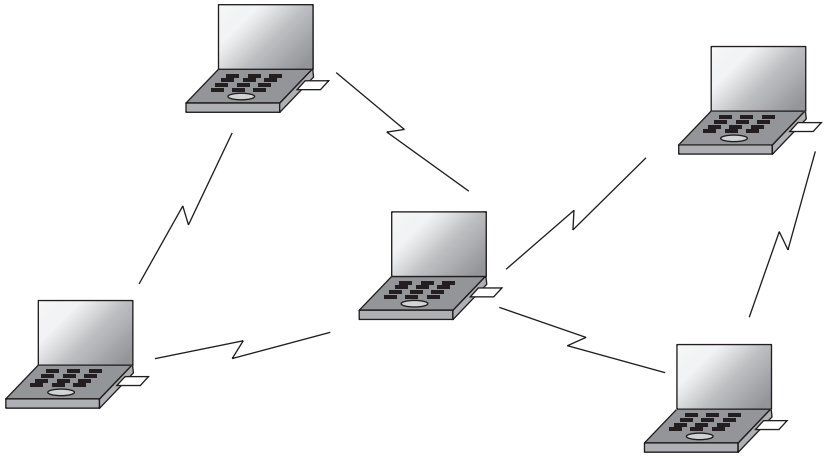


Figure 12.1: A Mobile Ad Hoc Network

Since no base stations are required, ad hoc networks can be deployed quickly, without having to perform any advance planning or construction of expensive network infrastructure.

Hence, such networks are ideally suited for applications where such infrastructure is either unavailable or unreliable. Typical applications include military communication networks in battlefields, emergency rescue operations, undersea operations, environmental monitoring, and space exploration. Because of their "onthe-fly" deployment quality and relatively low cost of implementation, ad hoc networks are also used in places where they are cheaper than their infrastructured counterparts. Examples of these applications consist of a network of laptop computers in conference rooms, network of digital electronic equipment and appliances (e.g., VCR, television, computer, printer, remote control) to form a home area network, networks of mobile robots, and wireless toys [Refs. 12.43, 12.14, 12.49]. Recently, there has been a growing interest of using ad hoc networks of wireless sensors to perform unmanned distributed surveillance and tracking operations [Ref. 12.47].

The design of ad hoc networks faces many unique challenges. Most of these arise due to two principal reasons. The first is that all nodes in an ad hoc network, including the source nodes, the corresponding destinations, as well as the routing nodes forwarding traffic between them, may be mobile. As the wireless transmission range is limited, the wireless link between a pair of neighboring nodes break as soon as they move out of range. Hence, the network topology that is defined by the set of physical communication links in the network (wireless links between all pairs of nodes that can directly communicate with each other) can change frequently and unpredictably. This implies that the multihop path for any given pair of source and destination nodes also changes with time. Mobility also causes unpredictability in the *quality* of an existing wireless link between neighbors. A second reason that makes the design of ad hoc networks complicated is the absence of centralized control. All networking functions, such as determining the network topology, multiple access, and routing of data over the most appropriate multihop paths, must be performed in a distributed way. These tasks are particularly challenging due to the limited communication bandwidth available in the wireless channel.

These challenges must be addressed in all levels of the network design. The physical layer must tackle the path loss, fading, and multi-user interference to maintain stable communication links between peers. The data link layer (DLL) must make the physical link reliable and resolve contention among unsynchronized users transmitting packets on a shared channel. The latter task is performed by the medium access control (MAC) sublayer in the DLL. The network layer must track changes in the network topology and appropriately determine the best route to any desired destination. The transport layer must match the delay and packet loss characteristics specific to such a dynamic wireless network. Even the application layer needs to handle frequent disconnections.

Although this area has received a lot of attention in the past few years, the idea of ad hoc networking started in the 1970s when the U.S. Defense Advanced Research Projects Agency (DARPA), sponsored the PRNET (Packet Radio Network) project in 1972 [Ref. 12.26]. This was followed by the SURAN (Survivable Adaptive Radio Network) project in the 1980s

[Ref. 12.52]. These projects supported research on the development of automatic call setup and maintenance in packet radio networks with moderate mobility. However, interest in this area grew rapidly in the 1990s due to the popularity of a large number of portable digital devices such as laptop and palmtop computers, and the common availability of wireless communication devices. The rising popularity of the Internet added to the interest to develop internetworking protocols for mobile ad hoc networks operating in license-free radio frequency bands (such as the Industrial-Scientific-Military or ISM bands in the United States). In the interest of developing IP-based protocols for ad hoc networking, a working group for Mobile Ad Hoc Networking (MANET) was formed within the Internet Engineering Task Force (IETF) [Ref. 12.20]. The DoD (Department of Defense) also renewed its support on similar research objectives by starting the GloMo (Global Mobile Information Systems) and the NTDR (Near-Term Digital Radio) projects. Spurred by the growing interest in ad hoc networking, a number of commercial standards were developed in the late 1990s. These included the IEEE 802.11 physical layer and MAC protocol in 1995 [Ref. 12.10], which have since then evolved into more updated versions. Today, one can build an ad hoc network by simply plugging in 802.11 PCMCIA cards into laptop computers. Bluetooth [Ref. 12.13] and Hiperlan [Ref. 12.53] are some other examples of related existing products. In this chapter we discuss some of the key challenges, protocols, and future directions of mobile ad hoc networks.

12.1 Physical Layer and MAC

The main aspects of designing the physical transmission system are dependent on the characteristics of the radio propagation channel such as path loss, interference (co-channel), and fading. In addition, since mobile terminals usually have limited power resources, the transceiver must be power efficient. These aspects are taken into account while designing the modulation, coding, and power control features in the radio equipment. In principle, the radio equipment in the nodes forming a mobile ad hoc network can use any technology as long as it provides reliable links between neighboring mobile terminals on a common channel. Candidate physical layers that have gained prominence are infrared and spread-spectrum radio.

The MAC plays the key role in determining the channel usage efficiency by resolving contention amongst a number of unsupervised terminals sharing the common channel. An efficient MAC protocol would allow the transmissions from independent nodes to be separated in time and space, thereby maximizing the probability of successful transmissions and maintaining fairness among all users. Though research on medium access schemes for wired local area networks (LANs) have been done for many years, the same concepts cannot be directly applied to wireless LANs. In a wired medium, a transmitted signal is received with the same signal strength at all terminals connected to the same shared medium. Hence a terminal in a LAN can avoid contention by sensing the presence of a carrier to determine if any other terminal is using the channel before it starts a transmission. This “listen before transmit” principle has led to a class of efficient random access protocols for wired LANs

that are generally known as carrier sense multiple access (CSMA) schemes [Ref. 12.28]. A popular example is CSMA/CD (CSMA with collision detection), which is the standard for Ethernet (IEEE 802.3) LANs [Ref. 12.46].

However, designing MAC protocols for wireless networks raises a different set of challenges. Propagation path losses in the wireless channel cause the signal power to decline with distance. This introduces the following problems, which are the main factors that affect the efficiency of the MAC in a mobile ad hoc network:

Carrier Sensing Is Location-Dependent: Since the strength of the received signal depends on the distance from the transmitter, the same signal is not heard equally well by all terminals. Hence carrier sensing is not very effective in wireless. Typical problems of using carrier sensing to determine the availability of the wireless channel are:

- *The hidden terminal problem*: a node may be hidden or out of range from a sender but within range of its intended receiver. For instance, in Figure 12.2a, node C is out of range from A, and hence any transmission from C cannot be heard by A. So while C is transmitting to B, node A thinks that the channel is idle and simultaneously transmits a data packet to node B. This causes both packets to be lost at B because of interference, and the packets are considered to have suffered a "collision." A transmission from A to B will face the same consequence even if C is transmitting to some other node, such as D.
- The exposed terminal problem: this is the reverse problem, where a transmitting or "exposed" node is within range of a sender but is out of range of the destination. The problem is illustrated in Figure 12.2b, where node B, which wants to transmit a data packet to A, finds the channel to be busy due to the transmission from C to D. Hence, B might wait for the transmission from C to be over before transmitting to A, which is not necessary as the transmission from B would not interfere at D.

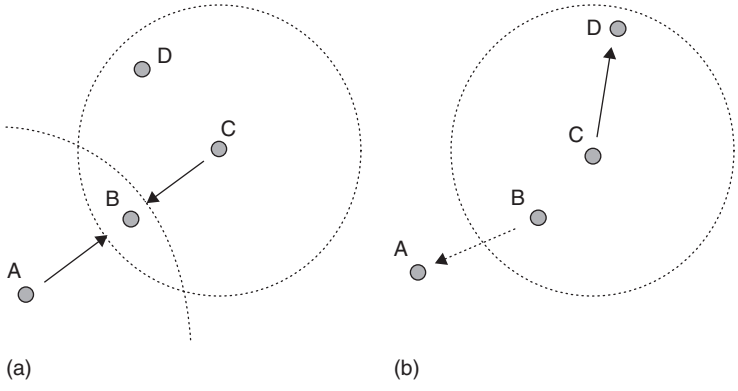


Figure 12.2: (a) The Hidden Terminal Problem, and (b) The Exposed Terminal Problem. The Dotted Circles Represent the Radio Range of the Transmitters

Both the hidden terminal and the exposed terminal problems arise due to the fact that carrier sensing is only performed at the transmitter, whereas its effect is determined by the interference power at the receiver, which are usually different due to propagation path loss characteristics.

Collision Detection Is Not Possible: A wireless transceiver cannot transmit and receive at the same time as the transmitted signal will always be far stronger than any received signal. Hence a wireless terminal cannot detect if its transmission has been successful. To inform the transmitting node about a successful packet transmission, the receiver sends an ACKNOWLEDGEMENT (ACK) packet back to the transmitter after it receives a data packet. If the transmitter does not receive an ACK within a fixed period of time, it assumes that the transmitted packet has been lost. However, this is learnt only after completing transmission of the data packet and waiting for a further no ACK timeout period.

Many different schemes have been designed for reducing these problems in wireless channel access. We first present the IEEE 802.11 standard that is the most popular scheme for wireless LANs, followed by a discussion on additional issues on the design of MAC protocols and current research directions.

12.1.1 IEEE 802.11

The IEEE 802.11 [Ref. 12.10] is an international standard of physical and MAC layer specifications for WLANs. It provides mandatory support for 1 Mb/s data rate with optional support for 2 Mb/s. These original specifications have been upgraded to higher data rates in succeeding versions, with the projected goal of going up to 54 Mb/s for future systems. The standard can be applied to both infrastructure-based WLANs, which use fixed access points for wireless communication with mobile terminals, as well as infrastructureless ad hoc networks. In the following, we discuss the main features of this standard with relation to ad hoc networks.

12.1.1.1 802.11 Physical Layer

IEEE 802.11 supports three different physical layers in order to allow designers to match price and performance to applications: one layer is based on infrared and two layers are based on radio transmission in the 2.4-GHz ISM band, an unlicensed band of radio frequencies available worldwide. The infrared specification is designed for indoor use only using line-of-sight and reflected transmissions of light waves in wavelengths from 850 to 950 nm. Both of the two RF specifications are based on spread spectrum, but employ different principles. While one uses frequency hopping (FH), the other is based on direct sequence (DS) spread spectrum. Either one can be used for the physical transmission system in ad hoc networks.

Frequency Hopping Spread Spectrum: As the name implies, a frequency-hopping spread spectrum radio hops from one carrier frequency to another during transmission. The transmission at any carrier frequency is narrowband. However, frequency spreading is

achieved by hopping from one carrier to another over a wide frequency band. The transmitter and receiver use the same sequence of carrier frequencies, which is pseudorandom (i.e., a long random sequence that repeats itself). The time for which the FH radio dwells in each frequency depends on the application requirements, government regulations, and adherence to standards. A *slow* FH system has a dwelling time that is longer than a bit period, whereas a *fast* FH system hops over many carrier frequencies during a single bit period. Since the hopping pattern is random, a FH system may experience interference during a few of the hops but achieve error-free transmission on other hops. One of the advantages of this property is that there is no hard limit on the total number of users that can be accommodated in a particular FH system. Rather, the limitation is decided by the amount of errors caused by multi-user interference that the users are willing to tolerate (known as the soft capacity). Such systems are especially beneficial in interference-limited communication systems, where the transmission capability is constrained by a large number of contending users who are not all active at the same time.

The 2.4 GHz ISM band in the United States (i.e., 2.4000 to 2.4835 GHz) has 79 channel frequencies in the hopping set, with a channel spacing of 1 MHz. The specified channel spacing allows a 1 Mb/s transmission rate using two-level Gaussian frequency shift keying (GFSK), which is the modulation scheme specified by the 802.11 standard. To achieve 2 Mb/s transmission rate, four-level GFSK modulation may be used, where two bits are encoded at a time using four frequencies. There are three different hopping sequence sets in the United States, with twenty-six hopping sequences in each set. All the terminals in any given ad hoc network must use the same hopping sequence. However, the availability of multiple sets allows multiple systems or networks to coexist in the same location.

Direct Sequence (DS) Spread Spectrum: The DS system achieves frequency spreading by multiplying each data bit by a sequence of chips ($+1/-1$ symbols that are shorter than a bit) before modulation. This has the effect of artificially increasing the transmission bandwidth. The receiver uses the same chip sequence to correlate the received signal. This technique achieves excellent interference rejection due to the auto-and cross-correlation properties of the random chip sequences. Usually the chip sequences are pseudorandom sequences having a long period. Multiple pairs of transmitters and receivers using different chip sequences can coexist in the same region. A DS system also has a soft capacity and can coexist with other narrowband radio systems without causing significant interference.

The IEEE 802.11 standard specifies an 11-chip Barker sequence for spreading each data bit. The modulation scheme is differential binary phase shift keying (DBPSK) for 1 Mb/s data rate, and differential quadrature phase shift keying (DQPSK) for 2 Mb/s. This effectively spreads the data stream over an 11 MHz band. Multiple systems can use different bands of frequencies whose center frequencies are separated by at least 30 MHz. As usual, all terminals of the same ad hoc network must use the same chip sequence (spreading code) for transmission as well as reception.

12.1.1.2 802.11 MAC

The 802.11 MAC is designed to provide mandatory asynchronous data service along with an optional time-bounded service that is only usable in an infrastructured wireless network with access points. The asynchronous data service is usable by both ad hoc networks and infrastructured wireless networks and supports "best effort" packet exchange without delay bounds.

The mandatory basic asynchronous service is provided by a method known as *carrier sense multiple access with collision avoidance* (CSMA/CA) and an optional channel reservation scheme based on a four-way handshake between the sender and receiver nodes. These two methods provide the mechanism for achieving distributed coordination amongst uncoordinated wireless terminals that do not use a fixed access point, and they are known as the distributed coordination function (DCF). A third method, known as the point coordination function (PCF), offers both asynchronous and time-bounded service, but it needs an access point to control medium access and avoid contention.

Basic DCF Using CSMA/CA: The basic channel access scheme uses two fundamental ideas to avoid collisions among contending transmitting stations:

- *Carrier sensing*: to determine that the medium is not being used by a neighboring transmitter (channel idle) before accessing the channel
- *Random backoff*: a terminal that senses the channel is busy, then waits for a random period of time to see the channel in the idle state before initiating transmission

A terminal that intends to transmit and senses the presence of a carrier (channel busy) waits till the end of the current transmission and considers the channel to be idle only when it detects the absence of the carrier for a certain duration of time, known as the *DCF inter-frame space* (DIFS). At the end of the DIFS period, in order to avoid collision with other terminals that might also be waiting for the current transmission to end before transmitting their packets, the terminal does not access the channel immediately. Instead, each terminal starts a backoff timer, which is initiated at a random value and counts down as long as the channel is sensed idle. The backoff timer is frozen whenever the channel is sensed as busy, resuming the countdown again after it goes idle (i.e., senses absence of the carrier for at least the DIFS period). The terminal initiates transmission only when its backoff timer reaches zero. The backoff interval is slotted, and may be expressed as $CW_{rand} \times slot\ time$, where CW_{rand} is a random integer chosen uniformly between 0 and CW and *slot time* is a predetermined slot duration. CW is the contention window, which can take one of the following sets of integer values: 7, 15, 31, 63, 127, 255. Initially a node uses the smallest value of CW and uses the next higher value in the set after each unsuccessful transmission.

In order to indicate that a transmission has been successful, a receiver transmits an ACK packet after a short inter-frame space (SIFS) period (which is shorter than DIFS) immediately following the reception of the data packet. In case an ACK is not received, the transmitter

assumes that the transmitted data packet is lost and it schedules a retransmission of the same. This will be continued for a maximum number of allowable retries at the MAC before the data packet is discarded.

An illustration of the access control scheme is shown in Figure 12.3. Here, nodes A, B, and C all have data packets to transmit when they find the channel busy (due to a transmission from some node X). After the channel is idle for a DIFS period, each node selects a random backoff period. In this illustration, the backoff timers of A, B, and C are chosen as 4, 1, and 3, respectively. So B's backoff timer reaches zero first, when it initiates transmission and the timers of both A and C are frozen. The transmissions of the data frames from A and C take place subsequently, as shown in Figure 12.3.

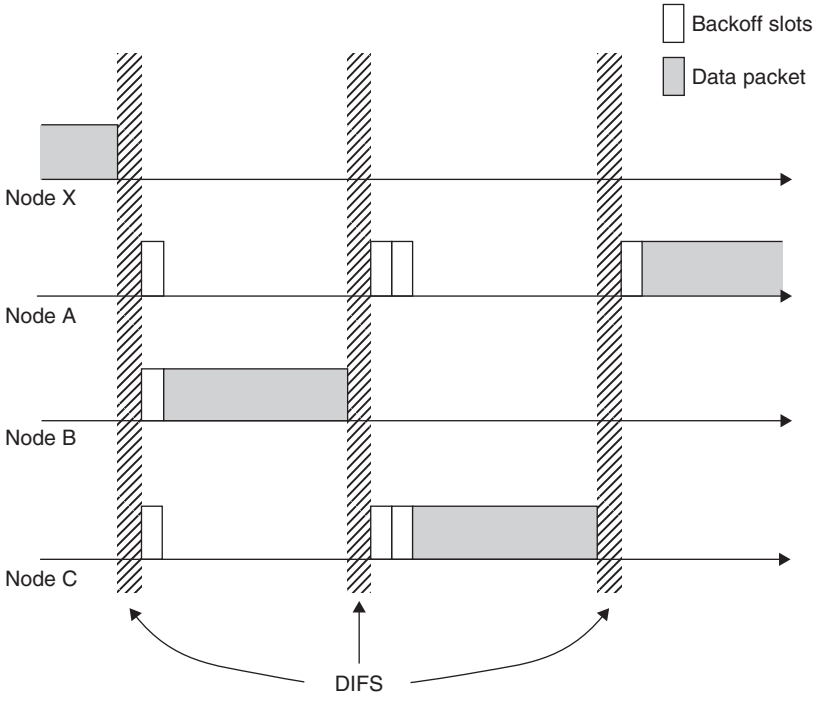


Figure 12.3: Illustration of the Basic CSMA/CA Protocol

According to this scheme, a collision may happen when multiple stations select the same backoff time. A large value *CW* will ensure a small probability of collision as it results in a smaller probability of two nodes selecting the same backoff time. However, a larger *CW* may cause a node to wait longer before transmission. When very few nodes are transmitting, a large value of *CW* causes inefficient usage of the channel. Hence, initially all nodes set the *CW* to the smallest value of 7. With heavier traffic, some of the transmissions will collide and eventually higher and higher values of *CW* may be chosen by the nodes to ensure collision-free transmission.

CSMA/CA with RTS/CTS Extension: Though the basic CSMA/CA scheme has excellent mechanisms to avoid collisions among a number of uncoordinated nodes that can hear one another, it does not solve problems due to hidden and exposed terminals. In order to address the hidden terminal problem, 802.11 has the option of adding the mechanism of an exchange of *request to send* (RTS) and *clear to send* (CTS) control packets between a transmitting and receiving nodes before initiating the transmission of a data packet.

The principle behind the use of the RTS and CTS packets can be seen from Figure 12.4. Here, node A, which intends to send a data packet to B, first broadcasts an RTS packet using the basic CSMA/CA scheme. The RTS frame contains the identity of the destination B, and the time that would be required for the entire transmission to complete. If B receives the RTS packet, it replies with a CTS packet after waiting for SIFS. The CTS packet also contains the time required for completion of the intended data exchange and the identity of the transmitting node. Upon receiving the CTS packet from B, A waits for SIFS and then transmits the data packet. When the data packet is received, B sends an ACK packet after SIFS, thus completing one entire data packet transfer protocol. All neighbors of A and B that receive either the RTS or the CTS learn about the intended exchange process and cooperate by remaining silent for the period of time that is required for the data exchange to be over.

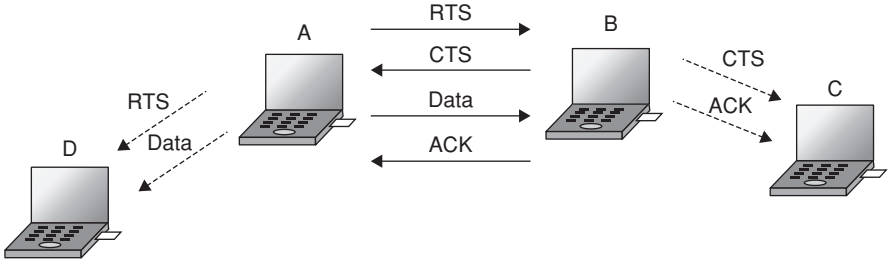


Figure 12.4: CSMA/CA with RTS/CTS Handshake

The exchange of RTS and CTS packets serves two purposes:

1. If A receives the CTS, it is ascertained that B is ready to receive and there are no interfering transmissions near node B. This process thus serves as a "virtual carrier sensing" mechanism.
2. All neighboring nodes of the destination, including those hidden from A (such as C), are expected to hear the CTS packet and remain silent for as long as it is required for the data transmission to be over. Neighbors of A (such as D) also remain silent for the period specified by the RTS so that their own transmissions do not experience any interference from the data packet to be transmitted from A. A silent period is implemented in a listening node by setting its *net allocation vector* (NAV) in accordance to the duration field in the RTS or CTS, which specifies the earliest possible time at which it can access the channel again. Hence, the RTS/CTS exchange effectively *reserves* the channel for the intended data transmission from A to B.

Even though the data packets have higher probability of success due to this channel reservation technique enacted by the RTS/CTS exchange, the RTS and CTS packets themselves are susceptible to the same rate of failure as that of the basic CSMA/CA scheme. Many of these control packets may suffer loss due to collisions and require retransmissions before the channel reservation is performed successfully. However, since the RTS and CTS control packets are shorter than the data packets, the scheme usually has a better throughput performance than the basic CSMA/CA in the presence of hidden terminals. Comprehensive analysis of the performance of the DCF under various conditions in mobile ad hoc networks have been reported [Refs. 12.5, 12.7, 12.58].

12.1.2 Additional Issues on MAC

Several concerns with the IEEE 802.11 MAC have motivated researchers to explore newer techniques to improve the channel utilization and throughput in mobile ad hoc networks. The basic access method of the 802.11 MAC protocol is susceptible to inefficiencies due to the hidden and exposed terminal problems. The RTS/CTS option reduces the hidden terminal problem but not the inefficiency caused by the exposed terminal problem. Some other concerns of 802.11 DCF using the RTS/CTS dialog are discussed in the following.

12.1.2.1 Additional Overhead of Control Packets

The transmission RTS and CTS control packets consume an additional amount of bandwidth. Usually this is justified when the size of the data packets is large and the advantage gained from channel reservations far outweighs the disadvantage induced by the additional overhead caused by the control packets. However, it has been observed that especially in higher loads and under high mobility, most of the channel bandwidth may be consumed by RTS and CTS transmissions [Ref. 12.21].

12.1.2.2 Collisions of Control Packets

Since the RTS and CTS packets are susceptible to collisions, the channel reservation scheme may fail, leading to loss of data packets as well. Figure 12.5 illustrates such a scenario. Here A starts an RTS-CTS dialog with B before transmitting a data packet to it. The CTS reply from B is received by A correctly, but it is not received by C, due to a collision with an RTS packet sent from D to E. Node A assumes that the channel is successfully reserved and proceeds with transmission of the data packet to B. This data transmission is vulnerable to interference from C, which has not been able to set its NAV accordingly, and may initiate a transmission to any of its neighbors before the data transmission is over.

Problems such as these are common because the RTS and CTS packets themselves are sent using the basic CSMA/CA access method, which is prone to the hidden and exposed terminal problems. A technique described by Garces and Garcia-Luna-Aceves [Ref. 12.17] tries to resolve this problem by making the duration of the CTS *longer* than the RTS packets. This

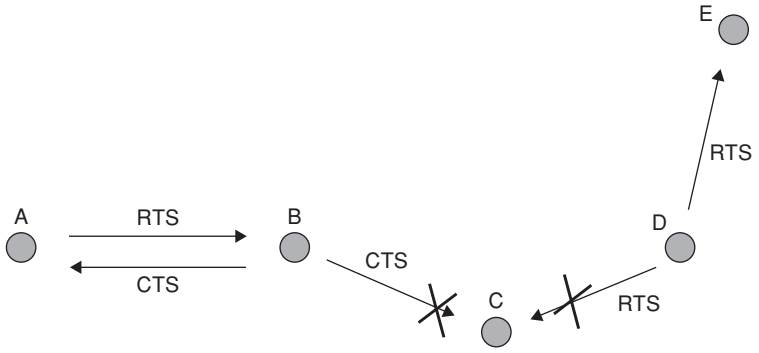


Figure 12.5: Example Where the Node C that is "Hidden" From A Misses the CTS Packet From B Due to a Collision From an RTS Packet From D

ensures that in the event that an RTS packet collides with a CTS at a receiver (such as in C in Figure 12.5), it would still be able to detect a part of the CTS packet. This might allow it to set its NAV to avoid interfering with the data exchange.

12.1.2.3 Radio Interference

Since wireless transmission is mostly limited by interference rather than noise, it is important to study the nature of interference and its effect on packet success probability. Figure 12.6 depicts the strengths of signals that would be received at node B from nodes A and C located at distances d1 and d2, respectively. Assuming that both transmitters use the same power Pt, the corresponding signal powers received at B, represented by Pr_A and Pr_C, respectively, depend on the path loss characteristics and the corresponding path lengths. The probability of error at the receiver depends on the total *signal-to-interference-plus-noise ratio* (SINR) of the corresponding packet at the receiver. The receiver noise is usually a constant parameter. The interference power is calculated by adding the powers of all radio signals at the receiver other than the power of the packet in question. The probability of bit error and consequently the

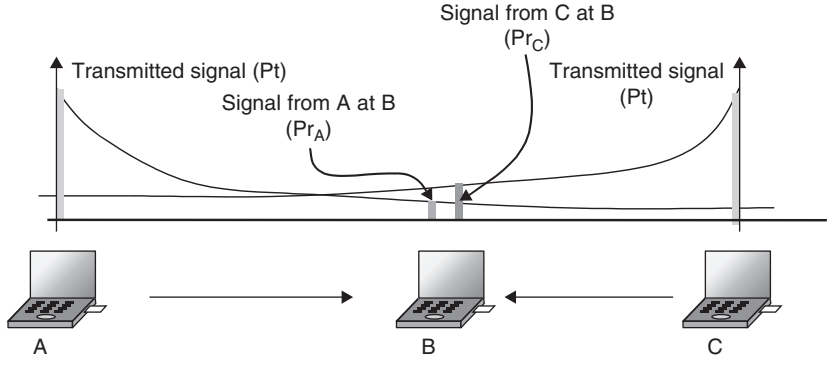


Figure 12.6: Effect of Propagation Path Loss on the Wireless Signal

packet error probability increases with decreasing values of the SINR. The minimum SINR required to correctly receive a packet depends on the radio technology, such as modulation, demodulation, coding, and so forth. A given radio usually has a specified *minimum SINR threshold* (SNR_{min}) for correctly receiving a packet. For instance, B will be able to receive the packet from A correctly if

$$\frac{Pr_A}{Pr_C + N} > SNR_{min} \tag{12.1}$$

For a given transmitter and corresponding receiver, the *transmission range* is defined as the maximum distance at which the received SINR is equal to $SINR_{min}$ in the absence of any interference, that is, the maximum distance at which reception will be error-free without interference. Usually radio channels are bidirectional, and hence a receiver will be able to receive a packet from a transmitter that is located within the transmission range (alternatively called the radio range).

A node determines the busy/idle state of a channel by comparing the strength of the carrier power to a predetermined *carrier-sense threshold* (T_{CS}). Typically, this threshold is chosen such that the carrier sensing range, or the distance within which all transmissions are detected, is at least as much as the transmission range of the nodes. A lower value of T_{CS} increases the carrier sensing range, but it also reduces frequency reuse by making larger number of nodes wait for their transmissions around a given transmitting node.

It is important to note that correct packet reception is not guaranteed whenever the receiver is within the transmission range of a transmitter. It also depends on the total amount of other interfering signals present. Typically, the interference power from a transmitter that is located at a distance less than the transmission range is expected to preclude the reception of any other packet without errors. Hence two simultaneous transmissions from nodes that are within range of a receiving node are said to have met with a "collision". The term "collision" has been borrowed from wireline networks, where any two simultaneously transmitted packets are lost irrespective of the location of the transmitters. In wireless networks, it relates to packet loss due to interference.

In wireless networks, packets may be lost due to interference from even those transmitters that are located outside the radio range of a receiver. An example is shown in Figure 12.7, where the combined interference from several transmitting nodes, all of which are out of range from node B, disrupts the reception of the packet from A to B.

12.1.2.4 Capture

Another concept used in wireless packet networks is packet capture, which refers to the mechanism with which a receiver can receive one of two simultaneously arriving packets if their received powers allow it [Ref. 12.31]. For instance, in Figure 12.6, even if both A and C

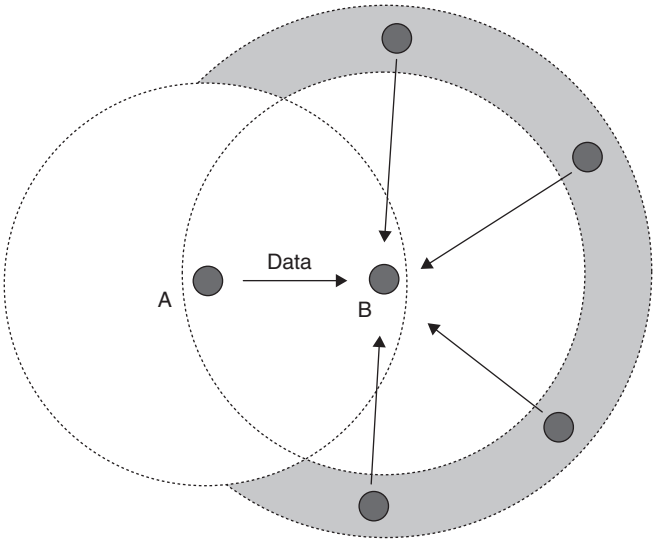


Figure 12.7: Illustration of the "Threat Zone," an Area Around the Destination B From Where Nodes Cannot Detect its CTS Packet. The Combined Interference From Transmissions From This Zone can Interfere with the Packet Reception at B

are transmitting at the same time, B can receive the packet from C as long as its power exceeds that from A by a sufficient margin. This is especially beneficial to the network performance under heavy traffic conditions when there are a large number of packet collisions and some of the collided packets are received successfully. A possible negative effect of the capture phenomena is that it can lead to unfair sharing of the channel. This can be seen in Figure 12.6, where transmitted packets from A will never be successful as long as C is transmitting, whereas the packets sent from C will always be captured at B.

12.1.2.5 Other MAC Protocols

Several solutions to these known problems have been suggested by various researchers. In the following, some of the notable concepts for new MAC protocols are summarized.

Collision Avoidance Techniques: The principle cause of packet loss in ad hoc networks is collisions, or interference caused by transmissions from hidden terminals. Several MAC protocols have been suggested that have features to avoid such collisions. One such technique is the transmission of a *busy tone* to indicate an ongoing data exchange process, which was first suggested in Tobagi and Kleinrock [Ref. 12.55]. Here, any node that hears an ongoing data transmission emits an out-of-band tone. A node hearing the busy tone will refrain from transmission, thereby increasing the distance of carrier sensing by a factor of two. Two other MAC protocols, the *Dual Busy Tone Multiple Access* [Ref. 12.9] and the *Receiver Initiated Busy Tone Protocol* [Ref. 12.59] also use this concept to avoid collisions. These schemes require additional complexity of narrowband tone detection and the use of separate channels.

Channel Reservation Techniques: The *Multiple Access with Collision Avoidance* (MACA) uses channel reservation based on the exchange of RTS and CTS control packets before transmission of the data packet [Ref. 12.27]. This scheme was incorporated in the IEEE 802.11 standard with the addition of a positive acknowledgement packet to indicate successful packet reception. Later, other protocols such as MACA for Wireless LANs (MACAW) [Ref. 12.3], Floor Acquisition Multiple Access (FAMA) [Ref. 12.17], and Collision Avoidance and Resolution Multiple Access (CARMA) [Ref. 12.18] also adopted the reservation scheme employing different variations of control packets.

Multiple Channel MAC: The concept of dividing the common medium into multiple orthogonal channels to reduce contention has been explored in [Refs. 12.24, 12.35, 12.38, 12.57]. When multiple channels are available, several concurrent transmissions are possible in the same neighborhood between distinct pairs of senders and receivers (Figure 12.8). If the same bandwidth is divided into *N* channels, either by frequency division or by using orthogonal CDMA codes, the traffic can be distributed over *N* channels. However, the transmission rate in each channel will also drop by a factor of *N*. It has been shown that such multichannel schemes can achieve a higher throughput by using an appropriate channel selection algorithm that allows each node to select the *best* channel available in its neighborhood. Several schemes for channel selection based on the exchange of RTS and CTS packets and carrier sensing over all channels have been explored [Refs. 12.24, 12.35, 12.38, 12.57]. The use of multiple channels

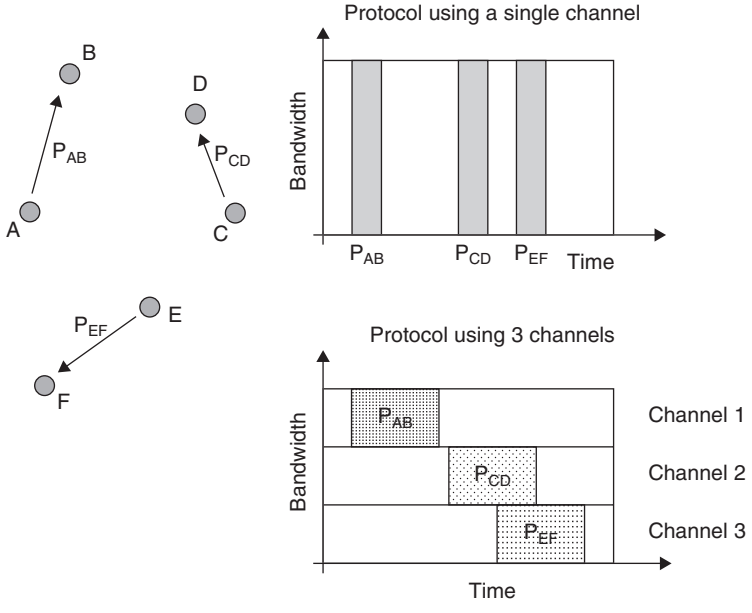


Figure 12.8: Illustration of Single-Channel and Multichannel MAC for the Concurrent Transmission of Three Packets: P_{AB}, P_{CD}, and P_{DE} From A to B, C to D, and E to F, Respectively

increases the hardware complexity, but it improves the throughput performance in the network by distributing the traffic over time as well as over bandwidth.

Use of Directional Antennas: Traditional ad hoc networks use omnidirectional antennas, as the direction for transmission and reception is variable. However, use of directional transmission provides several benefits for improving the link performance between a pair of communicating nodes:

1. A directional transmission can reduce the amount of interference to neighboring nodes. This can lead to a higher amount of frequency reuse and packet success probability.

2. A directional antenna can be used for receiving from a desired direction, reducing the amount of interference at the receiving node from adjacent transmitters. This further reduces the packet error probability.

3. Directional antennas have a higher gain due to their directivity. This can allow the transmitters to operate at a smaller transmission power and still maintain adequate SINR at the receiver. It will also reduce the average power consumption in the nodes [Ref. 12.36].

Despite these advantages, the usage of directional antennas in mobile ad hoc networks has additional design challenges. A mechanism for determining the direction for transmission and reception is required so that the mobile nodes can use directional antennas. Moreover, since all ad hoc networking protocols are traditionally designed for omnidirectional antennas, these protocols need to be adapted appropriately for proper functioning and maximizing the advantages that can be derived from directional transmissions and receptions. Many MAC and routing protocols that utilize directional antennas in ad hoc networks have been proposed in recent years [Refs. 12.30, 12.36, 12.37]. A comprehensive discussion on the various aspects of using directional antennas in ad hoc networks is given by Ramanathan [Ref. 12.48]. A central issue that concerns the applicability of directional antennas in mobile ad hoc networks is the comparatively larger size and cost of beam-forming antennas that are ideal for such applications. With advancements in technology and the possibility of shifting towards higher-frequency bands (such as the 5.8 GHz ISM band), it may be possible to design smaller as well as less expensive directional antennas. Hence, there is a growing interest in utilizing directional antennas in ad hoc networks.

12.2 Routing in Ad Hoc Networks

Movements of nodes in a mobile ad hoc network cause the nodes to move in and out of range from one another. As a result, there is a continuous making and breaking of links in

the network, causing the network connectivity (topology) to vary dynamically with time. Since the network relies on multihop transmissions for communication, this imposes major challenges for the network layer to determine the multihop route over which data packets can be transmitted between a given pair of source and destination nodes. Figure 12.9 demonstrates how the movement of a single node (C) changes the network topology, rendering the existing route between A and E (i.e., A–C–E) unusable. The network needs to evaluate the changes in the topology caused by this movement and establish a new route from A to E (such as A–D–C–E).

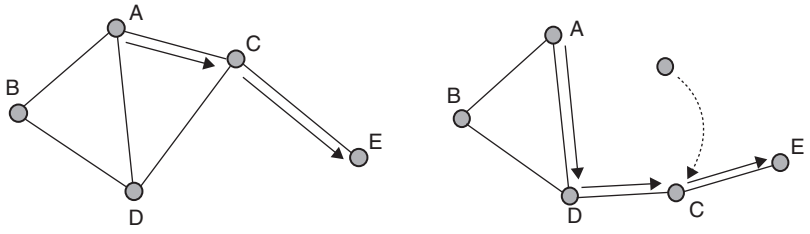


Figure 12.9: Illustration of the Change in the Route From A to E Due to the Movement of Node C

Because of the time-varying nature of the topology of mobile ad hoc networks, traditional routing techniques, such as the shortest-path and link-state protocols that are used in fixed networks, cannot be directly applied to ad hoc networks. A fundamental quality of routing protocols for ad hoc networks is that they must *dynamically* adapt to variations of the network topology. This is implemented by devising techniques for efficiently tracking changes in the network topology and rediscovering new routes when older ones are broken. Since an ad hoc network is infrastructureless, these operations are to be performed in a *distributed* fashion with the collective cooperation of all nodes in the network. Some of the desirable qualities of dynamic routing protocols for ad hoc networks are:

- *Routing overhead*: Tracking changes of the network topology requires exchange of control packets amongst the mobile nodes.

- These control packets must carry various types of information, such as node identities, neighbor lists, distance metrics, and so on, which consume additional bandwidth for transmission. Since wireless channel bandwidth is at a premium, it is desirable that the routing protocol minimizes the number and size of control packets for tracking the variations of the network.

- *Timeliness*: Since link breakages occur at random times, it is hard to predict when an existing route will expire. The timeliness of adaptation of the routing protocol is crucial. A broken route causes interruption in an ongoing communication until a new

route is established. Often the newly rediscovered route may be largely disjoint from the older route, which creates problems in rerouting the packets that were already transferred along the route and could not be delivered to the destination. Ideally, a new route should be determined before the existing one is broken, which may not be possible. Alternatively, a new route should be established with minimum delay.

- *Path optimality*: With constraints on the routing overhead, routing protocols for mobile ad hoc networks are more concerned with avoiding interruptions of communication between source and destination nodes rather than the optimality of the routes. Hence, in order to avoid excess transmission of control packets, the network may be allowed to operate with suboptimal (which are not necessarily the shortest) routes until they break. However, a good routing protocol should minimize overhead as well as the path lengths. Otherwise, it will lead to excessive transmission delays and wastage of power.

- *Loop freedom*: Since the routes are maintained in a distributed fashion, the possibility of loops within a route is a serious concern. The routing protocol must incorporate special features so that the routes remain free of loops.

- *Storage complexity*: Another problem of distributed routing architectures is the amount of storage space utilized for routing. Ad hoc networks may be applied to small portable devices, such as sensors, which have severe constraints in memory and hardware. Hence, it is desirable that the routing protocol be designed to require low storage complexity.

- *Scalability*: Routing protocols should be able to function efficiently even if the size of the network becomes large. This is not very easy to achieve, as determining an unknown route between a pair of mobile nodes becomes more costly in terms of the required time, number of operations, and expended bandwidth when the number of nodes increases.

Because of its many challenges, routing has been a primary focus of researchers in mobile ad hoc networks. The MANET working group in the IETF has been working on the issue of standardizing an IP-based routing standard for mobile ad hoc networks. Consequently, a large number of dynamic routing protocols applicable to mobile ad hoc networks have been developed. Reviews of prominent routing protocols for mobile ad hoc networks may be found in various references [Refs. 12.4, 12.8, 12.23, 12.50].

Based on when routing activities are initiated, routing protocols for mobile ad hoc networks may be broadly classified in three basic categories: (1) *proactive* or *table-driven* protocols, (2) *reactive* or *on-demand* routing protocols, and (3) *hybrid* routing protocols. Some representative examples of each class are shown in Figure 12.10.

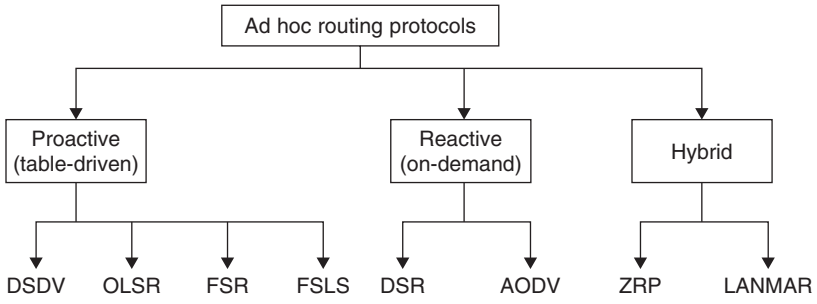


Figure 12.10: Classification and Examples of Ad Hoc Routing Protocols

12.2.1 Proactive Routing Protocols

Proactive protocols perform routing operations between all source destination pairs periodically, irrespective of the need of such routes. These protocols stem from conventional link state or distance-vector routing algorithms, and they attempt to maintain shortest-path routes by using periodically updated views of the network topology. These are typically maintained in routing tables in each node and updated with the acquisition of new information. Proactive protocols have the advantage of providing lower latency in data delivery and the possibility of supporting applications that have quality-of-service constraints. Their main disadvantage is due to the wastage of bandwidth in sending update packets periodically even when they are not necessary, such as when there are no link breakages or when only a few routes are needed.

12.2.1.1 Destination-Sequenced Distance-Vector Routing (DSDV)

DSDV [Ref. 12.44] is based on the classical Bellman-Ford algorithm [2] with adaptations that are specifically targeted for mobile networks. The Bellman-Ford algorithm uses the distance vector approach, where every node maintains a routing table that records the "next hop" for every reachable destination along the shortest route and the minimum distance (number of hops). Whenever there is any change in this minimum distance, the information is reported to neighboring nodes and the tables are updated as required.

To make this algorithm adequate for mobile ad hoc networks, DSDV added a *sequence number* with each distance entry to indicate the *freshness* of that entry. A sequence number is originated at the destination node and is incremented by each node that sends an update to its neighbors. Thus, a newer routing table update for the same destination will have a higher sequence number. Routing table updates are periodically transmitted throughout the network, with each node updating its routing table entries based on the latest sequence number corresponding to that entry. If two updates for the same destination have identical sequence numbers but different distances, then the shorter distance is recorded. The addition of sequence numbers removes

the possibility of long-lived loops and also the "counting-to-infinity" problem, where it takes a large number of update messages to ascertain that a node is not reachable [Ref. 12.44].

12.2.1.2 Optimized Link-State Routing Protocol (OLSR)

OLSR is a comparatively newer proactive routing protocol [Ref. 12.15]. It is an adaptation of conventional link-state routing in which each node tries to maintain information about the network topology. Each node determines the link costs to each of its neighbors by broadcasting HELLO messages periodically. Whenever there is a change in the link costs, the node broadcasts this information to all other nodes. In classical link-state algorithms, this is done by each node *flooding* the whole network with update packets containing updated link costs. Nodes use this information to apply a shortest-path algorithm (such as Dijkstra's shortest-path algorithm [Ref. 12.11]) to determine the best route to a specific destination.

OLSR optimizes the link-state protocol in two ways. First, it reduces the size of the update packets sent during the broadcasts by including only a subset of links to its neighbors. These are the links to a select set of neighbors known as the *multipoint relays* (MPR). The set of MPRs of a node consist of the minimum set of one-hop neighbors of that node so that the node can reach all of its two-hop neighbors by using these nodes as relay points. Each node computes its MPR set from the exchange of neighborhood information with all its neighbors. Second, instead of every neighbor broadcasting the update packets sent out by a node, only the MPR nodes participate in broadcasting these packets in OLSR. This minimizes the traffic of control packets during flooding. However, the savings of bandwidth achieved using these two techniques come at a cost of propagating incomplete topology information in the network. The updates include only MPR sets and not the sets of all neighbors of the broadcasting nodes. Hence, a shortest-path algorithm based on this partial topology information will generate routes containing the MPR nodes only. When the network is dense, that is, when each node has many neighbors, OLSR will work out to be efficient due to the reduction of control traffic for updates in the network.

12.2.1.3 Issues in Proactive Routing

The key characteristic of proactive routing protocols is that updates are sent periodically irrespective of need. Another issue is that they are table-driven. These two properties cause serious problems for making proactive routing protocols scale with network size. However, these protocols work well under heavy traffic and high mobility conditions as they try to maintain fresh routing information continuously.

Several new approaches have been proposed to make proactive protocols more scalable. One example is *Fisheye State Routing* (FSR) [Ref. 12.41], which is also an adaptation of link-state routing to ad hoc networks. FSR tries to limit routing load by avoiding flooding the network with routing information. Entire link-state information is only transmitted to the first-hop neighbors. In addition, it uses lower update rates for nodes that are located further away.

Hence, FSR maintains accurate route information on nodes that are close by, but the accuracy degrades with increasing distance of the destination from the source.

Overall, this technique saves the volume and size of routing traffic. A similar approach is adopted in the *Fuzzy Sighted Link-State algorithm* (FSLS) [Ref. 12.51]. As discussed, OLSR reduces routing load by broadcasting incomplete topology information. In general, these sacrifices lead to increased scalability of proactive routing protocols.

12.2.2 Reactive Routing Protocols

Reactive protocols are designed to minimize routing overhead. Instead of tracking the changes in the network topology to continuously maintain shortest path routes to all destinations, these protocols determine routes only when necessary. Typically, these protocols perform a *route discovery* operation between the source and the desired destination when the source needs to send a data packet and the route to the destination is not known. As long as a route is live, reactive routing protocols only perform *route maintenance* operations and resort to a new route discovery only when the existing one breaks. The advantage of this *on-demand* operation is that it usually has a much lower average routing overhead in comparison to proactive protocols. However, it has the disadvantage that a route discovery may involve *flooding* the entire network with query packets. Flooding is wasteful, which can be required quite frequently in case of high mobility or when there are a large number of active source-destination pairs. Moreover, route discovery adds to the latency in packet delivery as the source has to wait till the route is determined before it can transmit. Despite these drawbacks, on-demand protocols receive comparatively more attention than proactive routing protocols, as the bandwidth advantage makes them more scalable.

12.2.2.1 Dynamic Source Routing (DSR)

DSR is a reactive routing protocol that uses a concept called *source routing* [Ref. 12.25]. Each node maintains a *route cache* where it lists the complete routes to all destinations for which the routes are known. A source node includes the route to be followed by a data packet in its header. Routes are discovered on demand by a process known as *route discovery*. When a node does not have a route cache entry for the destination to which it needs to send a data packet, it initiates a route discovery by broadcasting a route REQUEST or QUERY message seeking a route to the destination. The REQUEST packet contains the identities of the source and the desired destination. Any node that receives a REQUEST packet first checks its route cache for an existing entry to the desired destination. If it does not have such an entry, the node adds its identity to the header of the REQUEST packet and transmits it. Eventually, the REQUEST packet will flood the entire network by traversing to all the nodes tracing all possible paths. When a REQUEST packet reaches the destination, or a node that has a known route to the destination, a REPLY is sent back to the source following the same route

that was traversed by that REQUEST packet in the reverse direction. This is done by simply copying the sequence of node identities obtained from the header of the REQUEST packet. The REPLY packet contains the entire route to the destination, which is recorded in the source node's route cache.

When an existing route breaks, it is detected by the failure of forwarding data packets on the route. Such a failure is observed by the absence of the link layer acknowledgement expected by the node where the link failure has occurred. On detecting the link failure, the node sends back an ERROR packet to the source. All nodes that receive the ERROR packet, including the source, delete all existing routes from their route caches that contain the specified link. If a route is still needed, a fresh route discovery is initiated.

12.2.2.2 Ad Hoc On-Demand Distance-Vector Routing (AODV)

AODV [Ref. 12.42] can be described as an on-demand extension of the DSDV routing protocol. Like DSDV, each route maintains routing tables containing the next hop and sequence numbers corresponding to each destination. However, the routes are created on demand, that is, only when a route is needed for which there is no "fresh" record in the routing table. In order to facilitate the determination of the freshness of routing information, AODV maintains the time since an entry has been last utilized. A routing table entry is "expired" after a certain predetermined threshold of time.

The mechanism for creating routes in AODV is somewhat different from that used in DSR. Here, when a node needs a route to some destination, it broadcasts a route REQUEST packet in which it includes the last known sequence number for that destination. The REQUEST packet is forwarded by all nodes that do not have a fresher route (determined by the sequence numbers) to the specified destination. While forwarding the REQUEST packet, each node records the earlier hop taken by the REQUEST packet in its routing table entry for the source (originator of the route discovery). Hence, a propagating REQUEST packet creates *reverse routes* to the source in the routing tables of all forwarding nodes. When the REQUEST packet reaches the desired destination or a node that knows a fresher route to it, it generates a route REPLY packet that is sent back along the same path that was taken by the corresponding REQUEST packet. The REPLY packet contains the number of hops to the destination as well as the most recent sequence number. Each node that forwards the REPLY packet enters the routing information for the destination node in its routing table, thus creating the *forward route* to the destination.

Routing table entries are deleted when an ERROR packet is received from one of the intermediate nodes on the route forwarding a data packet to the destination. When such an ERROR packet reaches the source, it may initiate a fresh route discovery to determine a fresh route to the destination.

12.2.2.3 Issues in Reactive Routing

Since reactive routing protocols only transmit routing packets when needed, these protocols are comparatively more efficient when there are fewer link breakages, such as under low mobility conditions. In addition, when there are only a few communicating nodes in the network, the routing functions are only concerned with maintaining the routes that are active. Because of these benefits, reactive or on-demand routing protocols have received more attention than proactive protocols for mobile ad hoc networks.

The main concern with reactive routing protocols is the need for flooding the entire network in search of a route when needed. Many optimizations have been suggested to reduce the excessive number of routing packets transmitted throughout the network during such flooding operations in reactive protocols. For instance, DSR has the option of broadcasting a *nonpropagating request packet* for route discovery, which is then broken into two phases. In the first phase, the source broadcasts a nonpropagating route request packet that only queries its first-hop neighbors for a known route to the destination. These packets are not forwarded by the neighbors. If none of the neighbors return a route, the source then proceeds to the second phase where a traditional propagating request packet is sent. The advantage of this scheme is that it avoids a networkwide flood of request packets when the route to the destination is known by one of the first-hop neighbors. A similar scheme is implemented in AODV using the concept of an *expanding ring search*. Here, increasingly larger neighborhoods, controlled by either hop-or time-constrained request packets, are searched to find the route to the destination. Some other techniques that perform similar optimizations are: *salvaging*, where an intermediate node in DSR uses an alternative route from its own cache when the original route is broken; and *promiscuous listening*, in which a node that overhears a packet not addressed to itself finds that it has a shorter route to the same destination and sends a *gratuitous reply* to the source with this new route. This increases the freshness of the route cache entries without additional route discoveries.

12.2.3 Hybrid Routing Protocols

The use of *hybrid routing* is an approach that is often used to obtain a better balance between the adaptability to varying network conditions and the routing overhead. These protocols use a combination of reactive and proactive principles, each applied under different conditions, places, or regions. For instance, a hybrid routing protocol may benefit from dividing the network into clusters and applying proactive route updates within each cluster and reactive routing across different clusters. Routing schemes that employ proactive route maintenance on top of reactive route discoveries have also been considered.

12.2.3.1 Zone Routing Protocol (ZRP)

ZRP [Ref. 12.22] divides the network into *zones* or clusters of nodes. The nodes within each zone maintain routing information for one another using a proactive algorithm such as a distance vector or link-state protocol. Hence, all nodes maintain updated routing tables

consisting of routes to all other nodes within the same zone (known as *intrazone routing*). Each zone also identifies a set of *peripheral nodes* that are located at the edges of the zone for communication with other zones. When a packet is to be sent to a node for which the source does not have an entry in its routing table, it is assumed that the destination is located in another zone. In that case, the node requests the peripheral nodes to send out a route request packet to all other zones in the networks. This is known as *interzone routing*, which uses a process that is similar to DSR except that the request packets are only handled by the peripheral nodes in the network. When the request packet reaches a peripheral node of the zone that contains the destination, a reply is sent back to the source. The overhead of flooding in such a route discovery is limited due to the involvement of peripheral nodes only. The proactive protocol in this hybrid framework limits the spread of periodic update packets within each zone. ZRP is especially suitable for large networks; however, the flooding of request packets during interzone route discoveries may still be a cause of concern.

12.2.3.2 Landmark Ad Hoc Routing Protocol (LANMAR)

LANMAR is designed for ad hoc networks that have the characteristics of group mobility, such as a group of soldiers moving together in a battlefield. Each group dynamically identifies a specific node within the group to be a *landmark* node. A proactive link-state routing protocol is used to maintain routing information within the group and a distance vector algorithm is used to do the same amongst all landmark nodes. Hence, each node has detailed topology information for all nodes within the group and distance and routing vector information to all landmarks. No routing information is maintained for nonlandmark nodes belonging to other groups. Packets to be sent to such a destination are forwarded towards the corresponding landmark. When the packet reaches the nodes within the group containing the destination, it is forwarded to the destination, possibly without going through its landmark. This scheme reduces the size of routing tables as well as the overhead of routing traffic forming a two-level routing hierarchy. Hence, it is expected to be more scalable than the so-called *flat* routing protocols.

12.2.4 Other Concepts in Ad Hoc Routing

There is an increasing list of new ideas and protocols for routing in mobile ad hoc networks. The MANET working group in the IETF publishes all significant developments and discussions by the group online in its mailing list [Ref. 12.20], which is the most comprehensive source of up-to-date information on research on ad hoc routing protocols. In addition to the representative protocols in the three broad categories of routing protocols described above, it is worthwhile to look at some of the other concepts that have been applied to routing in mobile ad hoc networks.

12.2.4.1 Geographic Position Aided Routing

The fundamental problems of routing in ad hoc networks arise due to the random movements of the nodes. Such movements make topological information stale, and hence, when an

on-demand routing protocol needs to find the route, it often has to flood the entire network looking for the destination. One of the ways of reducing the wastage of bandwidth in transmitting route request packets to every node in the network is to confine the search using geographical location information. Geographical positioning systems (GPS) can detect the physical location of a terminal using universal satellite-transmitted wireless signals. In recent times, GPS have become smaller, more versatile, and more cost-effective. Hence, several protocols have been proposed that assume the presence of a GPS receiver in each node and utilize the location information in routing [Refs. 12.1, 12.29, 12.39, 12.54].

One of the approaches for utilizing geographic location information in routing is to *forward data packets in the direction* of the location of the destination node, as proposed in various references [Ref. 12.1, 12.39, 12.54]. It may be required to define geographic location–specific addresses instead of logical node addresses to do that [Ref. 12.39].

An alternative concept is proposed in the Location Aided Routing (LAR) protocol [Ref. 12.29], which uses location information in on-demand routing *to limit the spread of request packets* for route discoveries. LAR uses information such as the last known location and speed of movements of a destination to determine a REQUEST ZONE, which is defined as a restricted area within which the REQUEST packets are forwarded in order to find the destination. Two different ways of defining REQUEST ZONES have been proposed. The idea is to allow route request packets to be forwarded by only those nodes that lie within the REQUEST ZONE, specified by the source. This limits the overhead of routing packets for route discovery, which would normally be flooded over the whole network.

A related protocol that uses *spatial locality* based on hop counts to confine the spread of request packets was proposed by Castaneda and Das [Ref. 12.6]. This protocol uses the concept that once an existing route is broken, a new route can be determined within a certain distance (measured in number of hops) from the old route. The protocol confines the spread of route request packets while searching for a new route to replace one that is freshly broken. For a new route discovery where no earlier routes were on record, the protocol still uses traditional flooding. However, this *query localization* technique for rediscovering routes still saves routing overhead.

12.2.4.2 Stability-Based Routing

A different approach to improve the performance of routing in mobile ad hoc networks is based on using routes that are selected on the basis of their *stability*. The Associativity-Based Routing (ABR) protocol [Ref. 12.56] maintains an *association stability metric* that measures the duration of time for which a link has been stable. While discovering a new route, the protocol selects paths that have a high aggregate-association stability. This is done with the idea that a long-lived link is likely to be stable for a longer interval than a link that has been relatively short-lived.

Signal Stability-Based Routing (SSR) [Ref. 12.12] uses signal strengths to determine stable links. It allows the discrimination between "strong" and "weak" links when a route request packet is received by a node. The request packet is forwarded by the node if it has been received over a strong link. This allows the selection of routes that are expected to be stable for a longer time.

12.2.4.3 Multipath Routing

On-demand or reactive routing protocols suffer from the disadvantage that data packets cannot be transmitted until the route discovery is completed. This delay can be significant under heavy traffic conditions when the REQUEST or the REPLY packet may take a considerable amount of time in traversing its path. This characteristic, along with the fact that each route discovery process consumes additional bandwidth for the transmission of REQUEST and REPLY packets, motivates us to find ways to reduce the frequency of route discoveries in on-demand protocols. One way of doing that is to maintain multiple alternate routes between the same source-destination pair such that when the primary route breaks, the transmission of data packets can be switched over to the next available path in the memory. Under the assumption that multiple paths do not break at the same time, which is most often true if the paths are sufficiently disjoint, the source may delay a fresh route discovery if the alternate paths are usable. As a result, many routing protocols have been designed to maintain multiple paths or routes for each pair of source and destination nodes.

The Temporally Ordered Routing Algorithm (TORA) [Ref. 12.40] provides multiple alternate paths by maintaining a "destination oriented" directed acyclic graph from the source. The DSR protocol also has an option of maintaining multiple routes for each destination in the route cache, so that an alternate route can be used upon failure of the primary route. Two multipath extensions of DSR were proposed by Nasipuri, Castaneda, and Das [Ref. 12.34] that aggressively determine multiple disjoint paths for each destination. Here, two different schemes for selecting alternative routes were considered, both benefiting from reducing the frequency of route discoveries caused by link breakages. Several other multipath routing protocols that derive benefits using the same principle have also been proposed [Refs. 12.16, 12.32, 12.45].

12.2.4.4 Preemptive Routing

A purely reactive routing protocol typically does not avoid a multihop communication from being interrupted *before* the route breaks due to a link failure. Most reactive routing protocols initiate a fresh route discovery when an ERROR packet is received at the source due to a link breakage. This introduces a pause in the communication until a new route is found. The goal of *preemptive routing* protocols is to avoid such pauses by triggering a route discovery and switching to a new (and, it is hoped, better) route before the existing route breaks. Such protocols can be viewed as a combination of proactive and reactive routing, where the route maintenance is performed proactively but the basic routing framework is reactive.

The crucial design issue in such protocols is to detect when to initiate a preemptive route discovery to find a "better" route. The protocol proposed by Goff and colleagues [Ref. 12.19] uses the technique of determining this by observing when the signal strength falls below a predetermined threshold. If the wireless channel is relatively static, then this correctly detects the initiation of link failure due to increasing distance between the two nodes in the link. However, multipath fading and shadowing effects might lead to false alarms while using this technique. Alternatively, using a time-to-live parameter was proposed by Nasipuri and colleagues [Ref. 12.33]. In this protocol, a preemptive route discovery is initiated when a route has been in use for a predetermined threshold of time. The preemption obviously makes the route discoveries more frequent than what would be observed in a purely reactive scheme. To keep the routing overhead low, the preemptive routing protocol presented by Nasipuri and colleagues [Ref. 12.33] proposes the use of query localization in the preemptive searches.

12.3 Conclusion

The mobile ad hoc network is one of the newest members in the family of wireless networks that span the planet. This chapter has aimed to provide the main issues and an overview of the developments in the MAC and routing protocols for mobile ad hoc networks. Although a vast amount of work has been done on it in the recent past, many questions still remain unanswered. Some of the issues that need further thought include:

- *MAC*: How can we design improved and robust MAC schemes that would dynamically adjust to variations of the wireless link characteristics and simultaneously cater to the need for higher data rates, quality-of-service requirements, and power savings, and that would be crucial in many future applications?
- *Routing*: By far the biggest issue in mobile ad hoc networking research is routing. With the rapid and diverse nature of growth of mobile ad hoc networks, the choice of the routing protocol is likely to depend on the network size, mobility, and application requirements. However, it will be interesting to see if an approach to generate a unified standard for ad hoc routing is achievable.
- *Transport*: The issues of transport layer protocols for mobile ad hoc networks require special attention. A discussion on these issues is outside the scope of this chapter. It is often said that optimizing ad hoc network performance requires a multilayer approach, where design problems at different layers of the protocol stack are addressed together for a unified solution. How can we arrive at such a design solution?
- *Scalability*: Many applications are already being conceived where hundreds of thousands of nodes are being considered for ad hoc networking. How do we design protocols for these large scale networks?

- *Internet connectivity*: What is the best paradigm for extending the reach of the Internet to mobile terminals that form a mobile ad hoc network with access points to the Internet?
- *Security*: All wireless networks are susceptible to security problems such as eavesdropping and jamming. How can we provide security to mobile ad hoc networks?
- *Power*: One of the major limitations of portability arises from limitations of battery power. In addition to developing improved battery technology, future ad hoc networking protocols have to be made more power efficient so that the network can survive longer without replacement of batteries.

These items are far from comprising a complete list of challenging research problems that ad hoc networking has posed. It is my hope that this chapter will inspire the reader to look into some of these in more detail.

References

[12.1] S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, "A Distance Routing Effect Algorithm for Mobility (DREAM)," *Proceedings of the ACM MOBICOM 1998* (October 1998): 76–84.

[12.2] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. (Prentice Hall, Upper Saddle River, NJ, 1987).

[12.3] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: A Media Access Protocol For Wireless Lans," *Proceedings of the SIGCOMM 1994* (August 1994): 212–25.

[12.4] J. Broch, D. A. Maltz, D. B. Johnson, Y-C. Hu, and J. Jetcheva, "A Performance Comparison Of Multi-Hop Wireless Ad Hoc Network Routing Protocols," *Proceedings of ACM MOBICOM* (October 1998): 85–97.

[12.5] F. Cali, M. Conti, and E. Gregori, "IEEE 802.11 Wireless LAN: Capacity Analysis and Protocol Enhancement," *Proceedings of IEEE INFOCOM 1998* (March/April 1998): 142–9.

[12.6] R. Castaneda and S. R. Das, "Query Localization Techniques for On-Demand Routing Protocols in Ad Hoc Networks," *Proceedings of the 1999 ACM Mobicom Conference* (August 1999): 186–94.

[12.7] H. S. Chhaya and S. Gupta, "Performance Modeling of Asynchronous Data Transfer Methods of IEEE 802.11MAC Protocol," *Proceedings of IEEE Personal Communications Conference* 3 (October 1996): 8–15.

[12.8] S. R. Das, R. Castaneda, J. Yan, and R. Sengupta, "Comparative Performance Evaluation of Routing Protocols for Mobile, Ad Hoc Networks," *7th International Conference on Computer Communications and Networks (IC3N)* (October 1998): 153–61.

[12.9] J. Deng and Z. J. Haas, "Dual Busy Tone Multiple Access (DBTMA): A New Medium Access Control for Packet Radio Networks," *Proceedings of IEEE ICUPS 1998* 2 (October 1998): 973–77.

[12.10] IEEE Standards Department, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE standard 802.11–1997, 1997.

[12.11] E. W. Dijkstra, "A Note on Two Problems in Connection with graphs," *Numerical Mathematics* 1 (October 1959): 269–71.

[12.12] R. Dube, C. D. Rais, K. Wang, and S. K. Tripathi, "Signal Stability Based Adaptive Routing (SSA) for Mobile Ad Hoc Networks," *IEEE Personal Communication* 4 (February 1997): 36–45.

[12.13] J. Haarsten, W. Allen, J. Inouye, O. Joeressen, and M. Naghshineh, "Bluetooth: Vision, Goals, and Architecture," *ACM SIGMOBILE Mobile Computing and Communications Review* 2 (October 1998): 38–45.

[12.14] K. J. Negus, J. Waters, J. Tourilhes, C. Romans, J. Lansford, and S. Hui, "HomeRF and SWAP: Wireless Networking for the Connected Home," *ACM SIGMOBILE Mobile Computing and Communications Review* 2 (October 1998): 28–37.

[12.15] P. Jacquet, P. Muhlethaler, and A. Qayyum, "Optimized Link State Routing Protocol," draft-ietf-manet-olsr-05.txt, 2000. IETF Internet Draft.

[12.16] D. Ganesan, R. Govindan, S. Shenker, and D. Estrin, "Highly-Resilient, Energy-Efficient Multipath Routing in Wireless Sensor Networks," *Proceedings of ACM/SIGMOBILE MOBIHOC 2001* (October 2001): 295–98.

[12.17] R. Garces and J. J. Garcia-Luna-Aceves, "Floor Acquisition Multiple Access with Collision Resolution," *Proceedings of the ACM/IEEE Mobile Computing and Networking Conference* (November 1996): 10–12.

[12.18] R. Garces and J. J. Garcia-Luna-Aceves, "Collision avoidance and resolution multiple access with transmission queues," *ACM Wireless Networks Journal* 5 (February, 1999): 95–109.

[12.19] T. Goff, N. B. Abu-Ghazaleh, D. S. Phatak, and R. Kahvecioglu, "Preemptive routing in ad hoc networks," *Proceedings of the ACM MOBICOM2001* (July, 2001): 43–52.

[12.20] IETF MANET Working Group. http://www.ietf.org/html.charters/manet-charter.html.

[12.21] Z. J. Haas, "On the Performance of a Medium Access Control Scheme for the Reconfigurable Wireless Networks," *Proceedings of IEEE MILCOM 1997* (November 1997).

[12.22] Z. J. Haas and M. R. Pearlman, "The Performance of Query Control Schemes for the Zone Routing Protocol," *ACM/IEEE Trans. Net.* 9 (August 2001): 427–38.

[12.23] X. Hong, K. Xu, and M. Gerla, "Scalable Routing for Mobile Ad Hoc Network," *IEEE Network Magazine* (July–August, 2002): 11–21.

[12.24] N. Jain, S. R. Das, and A. Nasipuri, "A Multichannel MAC Protocol with Receiver-Based Channel Selection for Multihop Wireless Networks," *Proceedings of the IEEE IC3N2001* (October 2001): 432–39.

[12.25] D. Johnson and D. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," in *Mobile computing*, ed. by T. Imielinski and H. Korth: Kluwer Academic, Dordrecht, The Netherlands, 353 (1996): 153–181.

[12.26] J. Jubin and J. D. Tornow, "The DARPA Packet Radio Network Protocols," *Proceedings of the IEEE* 75 (January 1987): 21–32.

[12.27] P. Karn, "MACA: A New Channel Access Method for Packet Radio," *Proceedings of ARRL/CRRL Amateur Radio 9th Computer Networking Conference* (1990): 134–40.

[12.28] L. Kleinrock and F. A. Tobagi, "Packet Switching in Radio Channels: Part-i–Carrier Sense Multiple Access Modes and Their Throughput-Delay Characteristics," *IEEE Transactions in Communications* COM-23 12 (December, 1975): 1400–16.

[12.29] Y. Ko and N. H. Vaidya, "Location-Aided Routing (LAR) in Mobile Ad Hoc Networks," *ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM)* (November 1998): 66–75.

[12.30] Y. B. Ko, V. Shankarkumar, and N. H. Vaidya, "Medium-Access Control Protocols Using Directional Antennas in Ad Hoc Networks," *Proceedings of IEEE INFOCOM2000* (March 2000).

[12.31] C. T. Lau and C. Leung, "Capture Models for Mobile Packet Radio Networks," *IEEE Transactions on Communications* 40 (May, 1992): 917–25.

[12.32] S.-J. Lee and M. Gerla, "Split Multipath Routing with Maximally Disjoint Paths in Ad Hoc Networks," *Proceedings of IEEE ICC2001* (2001).

[12.33] A. Nasipuri, R. Burleson, B. Hughes, and J. Roberts, "Performance of a Hybrid Routing Protocol for Mobile Ad Hoc Networks," *Proceedings of IEEE International Conference of Computer Communication and Networks (ICCCN 2001)* (October 2001): 296–302.

[12.34] A. Nasipuri, R. Castaneda, and S. R. Das, "Performance of Multi-path Routing for On-Demand Protocols in Mobile Ad Hoc Networks," *ACM/Baltzer Mobile Networks and Applications (MONET) Journal* 6 (August, 2001): 339–49.

[12.35] A. Nasipuri and S. R. Das, "Multichannel CSMA with Signal Power-Based Channel Selection for Multihop Wireless Networks," *Proceedings of IEEE Fall Vehicular Technology Conference (VTC 2000)* (September 2000): 211–18.

[12.36] A. Nasipuri, K. Li, and U. R. Sappidi, "Power Consumption and Throughput in Mobile Ad Hoc Networks Using Directional Antennas," *Proceedings of the IEEE International Conference on Computer Communications and Networks (IC3N)* (October 2002): 620–26.

[12.37] A. Nasipuri, S. Ye, J. You, and R. E. Hiromoto, "A MAC Protocol for Mobile Ad Hoc Networks Using Directional Antennas," *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC 2000)* 3 (September 2000): 1214–19.

[12.38] A. Nasipuri, J. Zhuang, and S. R. Das, "A Multichannel CSMA MAC Protocol for Multihop Wireless Networks," *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC 1999)* 3 (September 1999): 1402–6.

[12.39] J. C. Navas and T. Imielinski, "Geographic Addressing and Routing," *Proceedings of the ACM MOBICOM 1997* (1997): 66–76.

[12.40] V. Park and S. Corson, "Temporally Ordered Routing Algorithm (TORA) Version 1, Functional Specification," http://www.ietf.org/internet-drafts/draft-ietf-manet-tora-spec-01.txt (August, 1998). IETF Internet Draft.

[12.41] G. Pei, M. Gerla, and T.-W. Chen, "Fisheye State Routing: A Routing Scheme for Ad Hoc Wireless Networks," *Proceedings of the IEEE ICC* 1 (June 2000): 70–74.

[12.42] C. Perkins and E. Royer, "Ad Hoc On-Demand Distance-Vector (AODV) Routing," http://www.ietf.org/internet-drafts/draft-ietf-manetaodv-02.txt (November 1998). IETF Internet Draft.

[12.43] C. E. Perkins, *Ad Hoc Networking*: Addison Wesley, Boston, 2002.

[12.44] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," *Proceedings of the ACM SIGCOMM 1994 Conference* (August 1994): 234–44.

[12.45] D.S. Phatak and T. Goff, "A Novel Mechanism for Data Streaming across Multiple IP Links for Improving Throughput and Reliability in Mobile Environments," *Proceedings of the IEEE INFOCOM 2002* 2 (2002): 773–81.

[12.46] D. C. Plummer, "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48-Bit Ethernet Addresses for Transmission on Ethernet Hardware," RFC826, Standard (November 1982).

[12.47] G. J. Pottie and W. J. Kaiser, "Wireless Integrated Network Sensors," *Communications of the ACM* 43 (May 2000): 51–8.

[12.48] R. Ramanathan, "On the Performance of Beam-Forming Antennas in Ad Hoc Networks," *Proceedings of ACM/SIGMOBILE MOBIHOC 2001* (October 2001).

[12.49] J. Redi and B. Welsh, "Energy Conservation for Tactical Robot Networks," *Proceedings IEEE MILCOM* (1999): 1429–33.

[12.50] E. M. Royer and C. K. Toh, "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," *IEEE Personal Communication* 6 (April 1999): 46–55.

[12.51] C. Santivanez, R. Ramanathan, and I. Stavrakakis, "Making Link-State Routing Scale for Ad Hoc Networks," *Proceedings of 2001 ACM* (October 2001): 22–32.

[12.52] N. Schacham and J. Westcott, "Future Directions in Packet Radio Architectures and Protocols," *Proceedings of the IEEE* 75 (January 1987): 83–99.

[12.53] ETSI Secretariat, "Hiperlan Functional Specification," draft prETS 300 652, 1995.

[12.54] I. Stojmenovic and X. Lin, "GEDIR: Loop-Free Location-Based Routing in Wireless Networks," *Proceedings of the International Conference on Parallel and Dist. Comp. Systems* (November 1999).

[12.55] F. A. Tobagi and L. Kleinrock, "Packet Switching in Radio Channels: Part II—The Hidden Terminal Problem in Carrier Sense Multiple-Access and the Busy-Tone Solution," *IEEE Transactions in Communications* COM-23 (December, 1975): 1417–33.

[12.56] C.-K. Toh, "Associativity-Based Routing for Ad Hoc Mobile Networks," *Wireless Personal Communications Magazine* 4 (1997): 103–39.

[12.57] Y.-C. Tseng, S.-L. Wu, C.-Y. Lin, and J.-P. Shen, "A Multichannel MAC Protocol with Power Control for Multihop Mobile Ad Hoc Networks," *Proceedings of the 21st International Conference Distributed Computing Systems* (April 2001): 101–110.

[12.58] J. Weinmuller, M. Schlager, A. Festag, and A. Wolisz, "Performance Study of Access Control in Wireless Lans IEEE 802.11 DFWMAC and ETSI RES 10 Hiperlan," *Mobile Networks and Applications* 2 (June, 1997): 55–67.

[12.59] Z. J. Haas, "On the Performance of a Medium Access Control Scheme for the Reconfigurable Wireless Networks," *Proceedings of the IEEE MILCOM* 3 (November, 1997): 1558–64.

This page intentionally left blank

CHAPTER 13

Wireless Sensor Networks

Farid Dowla
Michael R. Moore

The purpose of this chapter is to provide a general approach to planning and implementing wireless sensor networks to support the arrays of sensors needed to operate plants, conduct scientific experiments, and test components. Using this brief tutorial, the reader should be able to acquire and organize the various sensor networks into a cohesive system and interface with vendors and subject matter experts as needed.

The primary purpose of sensor networks is to: (1) provide timely accurate data about the state of a plant so that the plant can run with maximum efficiency; (2) provide data to scientists as part of a complex experiment; or (3) provide trustworthy data for test and verification of components before they go into operation. Therefore, the final decisions about which kinds of networks to use should be based on the economics of lifetime cost versus the value of the data. Thus, the deployment of sensor networks must necessarily involve business and technical considerations. This chapter will focus mainly on the technical issues, but it will also alert the reader to some issues that especially impact cost.

In order to give a context for the subsequent technical discussions, this chapter begins with a very brief description of some of the ways in which sensor networks can be used to increase the efficiency of a plant.

13.1 Applications

Currently, many sensor networks are deployed to track the levels of the various vessels in a "tank farm" as part of tracking the inventory of chemicals available for the plant. In some plants, a much more encompassing inventory control system exists that includes radio-frequency identification (RFID) tagging and tracking. Currently, RFID tagging and tracking efforts are starting to be combined with sensor networking to provide total asset visibility [Ref. 13.1]. These networks can be used in conjunction with the purchasing system to provide just-in-time inventory control.

Another application area involves agile manufacturing. Many industries are providing more specialized products, therefore customers purchase fewer of each. In order to maintain quality while changing the settings of the process more often, accurate near–real-time measurements of critical product parameters are necessary. This can be accomplished with

sensor networks that use noninvasive sensors to monitor the quality of the product during the process and communicate the pertinent information back to the control mechanisms that may also be automated. A related application area involves process refinement. Scientists and manufacturing engineers are constantly finding new ways to control materials (e.g., tailored heating profiles, controlled chemical reactions, etc.). Sensor networks can be used to provide the feedback necessary to evaluate and improve these methods.

Still another application area involves compliance. For example, many plants have established International Organization for Standardization (ISO) 9000 procedures to be able to sell products in Europe. Compliance invariably involves verifying that operations are repeatable and well documented. This implies among other things that sensors constantly monitor the parameters of the materials and the state of the processes. The data from these sensors must be verified and reliable. Standards such as Institute of Electrical and Electronics Engineers (IEEE) 1451 provide for smart sensors that have an associated transducer electronic data sheet (TEDS). This data sheet can be used to track and update calibration data, correction factors, and other parameters that establish and verify the data from these sensors. Other standards such as IEEE 1588 establish the protocols for distributing time information along with the data that is also necessary for verifying that the sensor data is properly assigned to the concurrent state of the process.

13.2 Plant Network Layouts

Figure 13.1 highlights some key portions of the plant that might utilize sensor networks or associated data networks. First of all, a zone for the plant intranet is indicated that would need to be strategically placed, if it is wireless, at a height that is above normal human traffic and large metal structures at floor level, and below the level of overhead beams and cranes. Equivalently, in an office area, antenna repeaters can be placed on the ceilings of large spaces or at hallway intersections. These antennas must have an appropriate radiation pattern that covers most of the hemisphere below the ceiling.

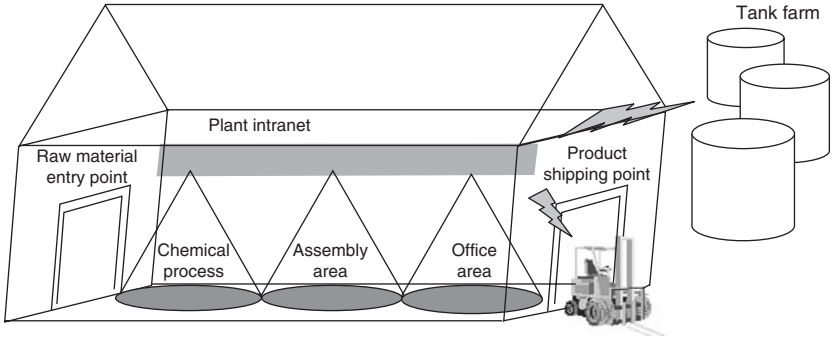


Figure 13.1: Overview of Plant Networks

Once a plant intranet is established with convenient access points (either wireless or wired), connection locations called access points can be established for the various subnets. These subnets provide tailored networking for data-intensive portions of the plant. That is, parts of the process or plant requiring large amounts of data to be used locally may utilize specialized sensor networks that only pass a subset of the data back through the plant intranet to the company's databases. This diversification of the network provides several benefits:

1. It allows cost-effective communication nodes to be tailored to their application rather than making all nodes carry the overhead and complexity of being all things to all users.
2. It provides an additional layer of security, especially against internal "hackers."
3. It makes spectrum management more manageable since each subset of the network can utilize a different portion of the EM spectrum or at least be allocated into regions (called micro-cells) within which only certain modulation schemes or frequency bands are utilized.

As shown in Figure 13.2, the material input and product output portals (as well as internal warehousing areas not shown) can be equipped with RFID portals. These portals typically include RF transceivers that irradiate items as they enter or leave the plant (or storeroom). Any items equipped with RFID tags of the proper protocol will be excited and read at the portal. These RFID readers would be typically connected to the plant intranet and thus incorporated into the inventory control databases. Efforts are underway to combine sensing (e.g., temperature) information along with the stored information on the RFID tags. Thus, the RFID systems are becoming a seamless part of the overall sensor networking picture.

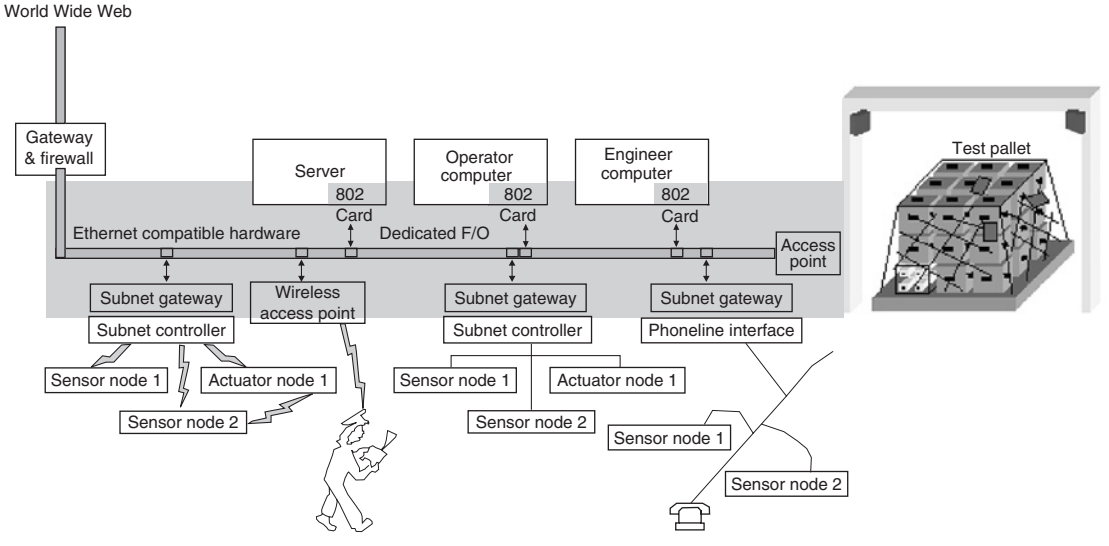


Figure 13.2: Plantwide Network Architecture

13.3 Plant Network Architecture

Figure 13.2 gives a generic plant architecture that connects several dedicated subnets into the plants' information infrastructure. In this example, a tethered IEEE 802-based (Ethernet) backbone is shown; however, all or portions of this network could be wireless, as discussed in Chapter 14. Four types of "subnets" are shown, as well as a single Internet Protocol (IP) capable wireless access point. The four subnets are (1) a wireless sensor network, (2) a tethered RF sensor subnet using a dedicated transmission line, (3) an RF sensor network that utilizes the existing phone-line infrastructure, and (4) an RFID portal and associated access point for tracking incoming materials or outgoing products.

Whether the plant intranet is tethered or wireless, it will be necessary to locate access points or gateways in the vicinity of individual "subnets" within the plant. These subnets may have a different set of requirements from the wide area networks/metropolitan area networks (WAN/MAN) that connect systems having a range of 1 km or more. For subnets that control or monitor a single process or large production unit (like an airplane wing), there is no need for global connectivity that would imply random access and IP support, but the throughput or synchronization requirements may be more stringent. Depending on how stringent the timing (nanosecond, microsecond, or millisecond) and how high the throughput, a random access standard operating in a controlled fashion may work. If not, a dedicated network with a central controller may be necessary.

13.4 Sensor Subnet Selection

The network designer should start by documenting the sensor networking requirements of the plant as well as estimating future areas of planned expansion, as shown in Table 13.1. Once all of the plant's sensor networking requirements have been established for the various regions of the plant, optimized subnets can be identified, purchased, and installed. One of the key lifetime costs of networks is the integration time and effort necessary to install the initial network as well as updating the hardware and software as the needs of the plant change with time. Therefore, it is my recommendation that open standards having a history of broad industry support such as IEEE 802 (Ethernet, 802.11b, etc.) be utilized for the plant intranet. This allows the network designer to deploy subnets that may or may not utilize IEEE 802

Table 13.1: Example Network Requirements Form

Subnet Description	# of Sensors	Total Estimated Throughput (bps)	Range (m)	Timing Resolution (seconds)
Tank farm	15	<100	2 k	10
Chemical process	6	~10 k	100	1
Vibration testing	100	~10 M	10	10^{-5}

protocols as long as manufacturers provide access points or bridges that translate the data from the subnet to IEEE 802 protocols.

In all of the following discussions, it is assumed that sensor nodes and networks can efficiently connect to the plant intranet. Although this chapter concentrates on the tradeoffs more particular to RF communications, the cost of integrating sensor networks into the plant operations should be a major consideration in the selection process. This cost will typically include most of the following: power cable installation, antenna installation, software driver development, user application software development, operator training, and maintenance. All of these costs must be considered when selecting a particular model of sensor network.

13.5 Functional Requirements

There are key functions that sensor networks should provide: safety, security, reliability, throughput, determinism, distributed intelligence, distributed controls, distributed communications, and data synchronization. A brief description of these follows.

Safety and security requirements are the most important issues when selecting any information system. Obviously, the safety and security features of the selected sensor network must be commensurate with the needs of the application. For instance, some applications require components and systems that are intrinsically safe.

In recent years, the users and developers of supervisory control and data acquisition (SCADA) systems have become increasingly aware of the necessity of securing their data and control links. Securing a wireless transmission may involve both RF signal means as well as bit-encryption means. For instance, spread-spectrum signaling makes it harder for a signal to be detected or intercepted, but this does not provide a very high level of data encryption. Any system requiring secure data should also employ message encryption means. Currently, some systems employ wired-equivalent privacy (WEP) encryption. The various encryption means are constantly being upgraded as hackers develop new methods of attacking them. Also, as mentioned earlier, networks must be protected from internal attacks since, "more than 70% of all corporate hacking is from inside the firewall . . . [by] a disgruntled employee" (p. 20) [Ref. 13.2]. This involves access controls and network architecture design. A detailed discussion of these issues is outside the scope of this chapter.

The need for reliability cannot be overstated when it comes to information networks. A recent report [Ref. 13.3] listed distrust of reliability as a key reason that more users did not employ wireless equipment:

> Reliability is a major concern. . . .This concern is not so much regarding failure of the products to work at all, but rather the reliability of the data transmission and reception. Most concern is over the effect of radiation resistance [EMC] . . . [Ref. 13.3].

Some typical reliability parameters are shown in Table 13.2.

Table 13.2: Some Example Reliability Criteria

Criteria	Value
Bit-error rate	10^{-5}
Probability of uncorrected errors	10^{-6}
Probability of undetected errors	10^{-9}
MTBF	5 years

The bit-error rate (BER) is a typical figure of merit for any communication link. It usually indicates the percentage of transmitted bits that are incorrectly received. However, most systems provide some level of error correction or at the very least error detection. The most critical of the three is the probability of undetected errors. Costly and sometimes dangerous electronic decisions can be made if undetected bit errors occur in a control system. The uncompensated BER and uncorrected errors affect the probability of a message getting through in a given amount of time. They also reduce the efficiency of the network, since retransmissions are required to compensate for detected but uncorrected errors. The bottom line is that the probability of undetected errors must be low enough not to cause dangerous or costly situations. Also, the BER and uncorrected error rates must be low enough to allow the network to support the necessary throughput.

Mean time between failure (MTBF) is a measure of how often the components fail. This number needs to be high enough that the user can reasonably expect the systems to run without replacement between major system or plant upgrades.

Another facet of reliability is the trustworthiness of the source and quality of the data. Even if the probability of undetected bit errors approaches zero and the data has been adequately encrypted, verification of the source and quality of the original data (before transmission) must be addressed. This involves establishing the quality of the sensor, the exact time at which the data was sampled, and sometimes whether or not the correct conversion algorithms were employed. Self-testing, self-calibrating, and the distribution of coordinated clock signals are some of the technical means employed to accomplish this type of reliability. These will be addressed in the next section, "Technical Tradeoffs and Issues."

Related to reliability are throughput and determinism. Throughput is usually quoted in bps (kbps or Mbps). When evaluating systems, the user must be aware of whether the stated values represent signaling rates or usable data rates. For instance, an RF link may have a signaling rate of 10 Mbps that only provides 2 Mbps of useful data bandwidth. This difference is due to the fact that many extra bits are transmitted to provide the necessary "hand-shaking" among

transmitter and receiver, error correction and detection, and security. Most manufacturers faithfully quote useful data bit rates, but the user must be aware of which of the two parameters they are comparing with their sensor communication needs (as typified by Table 13.1).

Determinism indicates that protocols are in place to guarantee that a message gets through within a given time window, or alternatively that the user is alerted if a message ever fails to get through. This figure of merit may be harder to ascertain, but for some applications it is critical. Sometimes average latency or guaranteed worst-case latency may be listed.

The last four functional requirements mentioned above—distributed intelligence, controls, and communications and data synchronization—allow interconnected systems to perform tasks that singly-optimized components could not. In general, distributed intelligence means that individual sensor nodes can accommodate algorithms or subroutines that reduce the raw data to key parameters, called feature vectors. This enables the user to obtain the necessary information without loading down the network with all of the "raw" data. Distributed controls, which often require peer-to-peer communications without going through the central computer, use this distributed intelligence to connect sensor nodes with actuator nodes. Distributed communications provides the peer-to-peer protocols and interfaces necessary for distributed controls. Finally, data synchronization enables the coordination of events (sampling of data, actuation of control events, etc.).

13.6 Technical Tradeoffs and Issues

This section focuses on the technical issues that must be understood when evaluating whether or not components and systems can achieve the functional requirements discussed above.

13.6.1 Bandwidth and Range

The most fundamental parameters for selecting the proper sensor network are bandwidth (or throughput) in bits per second (bps, kbps, Mbps) and range (meters, kilometers, feet, or miles). The physics of RF communications are such that throughput goes down with increased range, assuming all other parameters are held constant. In fact, many 802.11b devices are rated for four different throughputs that increase with decreasing guaranteed range. The tank farm example typically requires a relatively long range for sensor networks (1 km or more) but with a relatively low data rate (<1 kbps). On the other hand, some experiments performing modal analysis involving arrays of accelerometers may require several Mbps over a range of only a few meters or tens of meters. As mentioned, the user must distinguish between component bandwidth and the throughput rate of useful data.

13.6.2 Number of Sensors per Network

Related to the network bandwidth is the number of sensors that would reasonably be attached to a single access point of the plant intranet. In some systems each sensor will have its own

RF transmitter. In others, a sensor node may multiplex several sensors onto the communications bus. In both instances, the user must be aware of both the bus throughput maximum rates and the individual node/sensor maximum rates. Most busses can accommodate somewhere between 32 and 256 individual nodes. The user must decide which sensors need to be grouped onto a subnet to provide the necessary data coordination or control loops. Based on this the user must determine which subnet technologies support the requirements.

13.6.3 EMC

Electromagnetic compatibility (EMC) deals with compatibility in both directions. That is, it deals with limiting the emissions of the component being considered, as well as dealing with the potential susceptibility of the component to emissions from other devices. The Federal Communications Commission (FCC) rules focus mainly on limiting the emissions of electronic equipment such that they will not interfere with other devices. On the other hand, U.S. military standards (MIL-STDs) and European standards (e.g., International Electrotechnical Commission, or IEC) deal with both emissions and susceptibility.

First of all, wireless nodes must be selected within the constraints of the overall frequency planning of the facility. Typically, FCC guidelines are utilized as a general overview of which frequencies and power levels are acceptable within the planning of a region's spectrum use. However, local spectrum management authority must make judgment calls as to where and when various types of radio equipment can be used. For instance, some facilities do not allow handheld radios in the control room. For similar reasons, many hospitals restrict the locations within which cell phones can even be turned on. That is because any cell phone that is turned on is periodically transmitting at least a pilot signal so that it can be associated with a particular cell tower.

The determination of the proper use of wireless equipment within a plant is not limited to the knowledge and control of carrier frequencies (spectrum management) but should also include the selection of radio equipment with compatible modulation types (waveforms). In some instances, conventional narrowband transmitters such as handheld radios have interfered with important plant processing equipment, while direct-sequence spread-spectrum radio equipment was utilized in the same area without upsetting the process. This is because direct-sequence spread-spectrum waveforms "spread" the energy over a wider bandwidth (typically by factors of 10:1 to 1,000:1) such that the effective volts/Hz energy levels are lower.

Another modulation type called frequency-hopping spread-spectrum is often used to good effect to overcome jamming signals. It has been used heavily by the military and in some commercial applications.Acurrently popular line of components called Bluetooth utilizes frequency-hopping modulation. In this instance, the dwell time of the carrier is under a few msec; however, during that interval the signal is similar to narrowband transmitters. Frequency-hopping devices have been known to interfere with direct-sequence devices such

as wireless local area networks (LAN) utilizing IEEE 802.11b protocols. Because of this, some members of the IEEE 802.15 working groups have discussed separating IEEE 802.15.1 devices from direct-sequence devices by at least 6 feet.

Vendors should not only document FCC certification of their products, but they should also demonstrate that their units have worked in the presence of other wireless equipment. This wireless equipment should include handheld radios (which are typically narrowband licensed units), microwave ovens (some 2.45 GHz radio units are susceptible to the leakage fields from microwave ovens, even though their emanations are certified to be below permissible health limits), and cell phones. Also, the vendor should be able to show that their units do not upset existing plant equipment, especially other wireless links.

Some IEC documents call for the coordination of lightning protection experts, architects, and construction companies when building a plant to ensure that a thorough lightning protection system is in place. Analogously, an EMC expert should be consulted at plant startup and during major upgrades to ensure that all of the plant's wireless communication devices, equipment such as RF heaters, and other electronics sensitive to EM fields will peacefully coexist.

When there were only a few types of wireless devices, EMC was less of an issue than it is now. However, the current proliferation of wireless communication devices will cause intersystem EMC to become a major consideration when selecting wireless equipment. Note that intrasystem EMC is under the purview of the designer of the wireless components and that adhering to FCC rules provides some measure of intersystem EMC. History has proven that good engineering practices and well-planned governmental controls are not enough, however. Some systems will invariably interfere with others; good spectrum management and spatial separation will minimize this interference. The use of directive antennas, direct-sequence spread spectrum, power control, and all other available technical means should be employed to reduce the possibility of EMC problems.

13.6.4 Spectrum Management

As just mentioned, all wireless standards must be fully compliant with the government's FCC rules, especially concerning carrier frequencies and output power. The FCC has established license-free industrial, scientific, and medical (ISM) frequency bands suitable for wireless-sensor systems.

One choice to be made in the deployment of wireless sensor networks is in the frequency of transmission. Options include:

1. Licensed bands that require application procedures through the FCC.
2. The use of leased transmission facilities or services from independent providers.

3. Unlicensed systems restricted to specific ISM and similar bands, including bands around 915 MHz, 2.45 GHz, and 5–6 GHz.

When selecting a system utilizing licensed bands, the user must include the cost of obtaining FCC approvals into cost calculations. When utilizing the unlicensed bands, the user must be careful to manage the EMC issues already discussed.

13.6.5 Wireless Networking Standards

Table 13.3 [Ref. 13.4] shows a comparison of a few of the wireless networking standards adopted by IEEE 802, as well as by other standards organizations (e.g., IS-95 refers to the cell-phone CDMA standard). The network sizes shown include personal area networks (PANs), local area networks (LANs), and metropolitan area networks (MANs).

Table 13.3: Overview of Wireless Standards

Std	OFDM	FHSS	DSSS	GHz	Size	Mbps
IS-95			x	1+/−	Cell	1
Bluetooth		x		2.45	PAN	0.7
P802.15		x		2.45	PAN	0.7
P802.16b	x			5	MAN	54
802.11a	x			5	LAN	54
802.11		x	x	2.45	LAN	1, 2
802.11b			x	2.45	LAN	5.5, 11

As shown in Table 13.3, there are currently three common modulation techniques used in wireless information networks in addition to conventional narrowband techniques. These are frequency-hopping spread spectrum (FHSS), direct-sequence spread spectrum (DSSS), and the newest contender, orthogonal frequency division multiplex (OFDM). In my opinion, OFDM is more optimal for a few nodes streaming lots of data, and DSSS is better suited for lots of nodes handling less data per node [Ref. 13.5]. FHSS is often preferred over DSSS by the military, primarily because DSSS requires careful management of the transmit power of individual mobile nodes to overcome the near/far problem and FHSS does not.

13.6.6 Time Synchronization and Distribution

The synchronous sampling of sensors (or the determination of exactly when samples were taken) is also a very necessary component of a sensor-based network. In general, a sensor network must have a more tightly controlled, more deterministic time base than random-accessed general data networks can typically provide. Any protocol suitable for correlating samples from widely separated sensors must incorporate a highly robust mechanism to

synchronize various components of the system. For example, the individual transducer-to-bus interface modules (TBIMs) and the system bus controller (TBC) in the proposed IEEE P1451.3 standard coordinate their clocks via a dedicated 2 MHz spread-spectrum signal. Time resolutions required for some applications will possibly be in the microsecond range, although many scenarios will be far less demanding. Any sensor network system will need to provide means to achieve this more stringent timing resolution. Two approaches already proposed for system synchronization in the IEEE P1451.3 standard being developed include a dedicated sinusoidal sync signal on the cable and the simultaneous provision for transporting formatted timestamp data from the TBC to the remote TBIMs over the data channels. As mentioned, IEEE 1588 is one of the standards that defines a protocol for time distribution.

Time and data associations can be absolute or relative and can be established locally in the sensor node or remotely when the data arrives at the collection node. In high-end sensor nodes, a global positioning system (GPS) receiver or master clock could be used to generate an absolute time and date stamp that is communicated with the data. On the other hand, the sensor node could associate a relative number of clock ticks with the data and let the collection node add the absolute time information.

Whether it is relative or absolute time that is necessary, the mechanisms for establishing and distributing precision time must be carefully managed.

13.6.7 Power

Another key parameter in many applications is available power. If plant power is not conveniently located, then battery-operated or solar-powered nodes may be an option. While some users prefer battery-powered units, some prefer AC-powered units so that they will not require battery maintenance by plant personnel. When determining the cost of the systems to be installed, the cost of wiring in AC versus battery replacement costs should be compared. For instance, while current IEEE 802.11b-type nodes require too much power to be practical for battery power, some lower throughput nodes—especially those used for shorter ranges—are designed to be battery powered. Thus, the plant intranet (if it is wireless) should probably rely primarily on AC power with battery backup. It is more feasible to use battery power (or alternative power sources) for sensor networks that are used to optimize certain processes but are not critical to the safe and secure operation of the plant. Most plant engineers desire sensor communication nodes that do not need regular maintenance.

13.6.8 Key Smart Sensor Features

Through my involvement with sensor networking standards development, I have noted key features that enable sensors to be used in smart sensor networks: unique identification, efficient and standardized communication protocols, automatic networking, self-test and self-calibration, self-describing (e.g., TEDS), self-locating, and time coordination.

A universal unique identification (UUID) code that takes the place of the company name, model number, and serial number is critical because the history of the sensor cannot be traced or verified without it; its data would be suspect without an association with a unique identifier. In general, this UUID must be associated with electronic data that may be stored locally or remotely, as will be discussed in a moment. The Auto-ID center, an industry-funded research program headquartered at the Massachusetts Institute of Technology, has established an electronic product code (ePC) that could provide this UUID function. This ePC is intended to replace the UPC barcode that is almost ubiquitous in retail stores, as well as to provide unique identification of components before they become products. This center has also established an Object Name Server (ONS), similar to the Internet's Domain Name Server (DNS), which provides a link between the ePC and any available electronic database for that item.

Appropriate communication protocols are also important because the data is no good if it cannot be communicated. Also, even a well-thought-out protocol that is not widely accepted often demands many expensive hours of interface development. Therefore, a robust, widely accepted protocol is optimal.

Automatic networking refers to the sensor node functions of self-discovery, hot-swap, ad-hoc routing, and other methods that improve a node's ability to automatically and efficiently assimilate into an existing network. Nodes that can automatically establish themselves as a network and adapt to individual node replacements and even mobile platforms greatly increase the utility and decrease the administration costs of networking the sensors.

To the extent possible, the sensors and sensor nodes should have the ability to test their own health and calibration. If not, they should prompt the operators to take necessary actions on a periodic or as-needed basis.

All sensors should have some means of communicating their capabilities, as well as describing their unique identity. In IEEE 1451 standards this is described as having an accessible TEDS. The application software must have the appropriate graphic user interface (GUI) that allows the plant engineers (if not the operators) to have access to these parameters. They will be used in establishing the proper associations between the sensor measurements and the plant process. They will also be used to give the user confidence in the accuracy of the data. The TEDS can either be stored locally (embedded TEDS) or on a remote server (virtual TEDS).

The ideal sensor will have some means of informing the application layer of its absolute or relative location within the plant. Some manufacturers may choose to put GPS receivers on sensor nodes that will be exposed to satellites. Others may use barcodes or RFID tags to associate sensors with the plant process measurement. Still others may equip their nodes with an ID button that sends a message to the user's computer when activated. If no electronic means are available, the installers will have to manually enter the sensor-process association and location into the application database.

13.6.9 Tethered RF Links

Although this chapter focuses on wireless data links, the same RF waveforms may be utilized on tethered transmission lines. Many communication networks use one to five pairs of shielded wires to perform networking. These range from Ethernet that uses Category-5 cable to IEEE 1451.3 that uses two of the wires in a conventional phone cord. Other networks use coaxial cable (e.g., cable modems) and even AC wiring as their transmission medium. Since these networks unintentionally radiate some of their energy, they must also follow FCC guidelines.

In general, the AC wiring can handle low data-rate communications such as monitoring fluid levels. Data communications over telephone wiring can handle upwards of several Mbps. A consortium named Home Phone-line Networking Association (HomePNA) has developed products that communicate between PCs using existing phone lines. One implementation involves connecting the USB port of the PC to an HPNA box that then connects to another HPNA box via existing phone lines within a home, as shown in Figure 13.2. The HPNA boxes communicate using the CSMA protocol of IEEE 802.3 and their own unique physical layer protocol, which has been adopted as ITU-T G989.1 and ITU-T G989.2. The sensor networking standard IEEE 1451.3 has adopted HPNA as its physical layer. Note that in the case of "piggy-backed" systems, special circuits may be required at junction boxes and switches, since each separate phone number is wired independently and some AC circuits within a plant may be on different phases of the wiring.

Because of the uncertain quality of the lines, the utilization of existing phone and AC lines should be limited to noncritical sensor nodes. It can be very cost-effective for collecting reasonable amounts of data from locations that would otherwise require expensive wiring. However, it should not be used for the main "backbone" of the information infrastructure or for the networking data that requires highly critical or deterministic timing.

13.7 Conclusion

A general approach to planning and implementing the sensor networks to support the arrays of sensors needed to operate plants, conduct scientific experiments, and test components has been described. Using this brief tutorial, a user should be able to acquire and organize the various sensor networks and interface with vendors and subject matter experts as needed.

References

[13.1] The InterNational Committee for Information Technology Standards (INCITS), INCITS T20 (Real Time Locating Systems), http://www.autoid.org/INCITS/ncits_t20_2002.htm.

[13.2] E. Byres, "Can't Happen at Your Site?" *InTech Magazine* (1 February 2002): 20–22.

[13.3] "The North American Market for Wireless Monitoring and Control in Discrete and Process Manufacturing Applications," *Venture Development Corporate Report* (March 2002): 30.

[13.4] M. R. Moore, S. F. Smith, and K. Lee, "The Next Step: Wireless IEEE 1451Smart Sensor Networks," *Sensors Magazine* 18 (September 2001): 35–43.

[13.5] Ibid., p. 41.

CHAPTER 14

Reliable Wireless Networks for Industrial Applications

Farid Dowla
Robert Poor

Existing wired methods of providing communications in a factory or industrial setting have numerous shortcomings:

- Factory communication wiring can easily have an installed cost of $5 to $10 per foot. What if expensive communication wiring could be replaced with reliable wireless links?
- Productivity programs demand more and more information from smart devices. What if industrial gear could gain more local intelligence by sharing information with nearby sensors?
- More and more maintenance systems require remote data acquisition. What if it were possible to continually monitor the condition of all the equipment on the factory floor and predict failures before they happen, instead of locating them after the fact?

This chapter will describe various wireless approaches to factory and industrial communications.

14.1 Benefits of Using Wireless

An obvious problem that can be addressed with wireless solutions is simple wire replacement, where the radio frequency (RF) communication link emulates wire in an existing system. No changes are made to the system architecture. Rather, wireless links are used to transmit the same data that the physical wire once carried.

Consider an instrument connected by a serial cable to a control panel using Modbus as a communication protocol. Wireless RF links can replace the serial cable as the physical layer to carry Modbus packets back and forth, requiring no physical changes to the instrument, the control panel or the underlying software architecture. The serial cable is taken away, and a wireless transceiver is physically connected to the serial port at both the instrument and at the control panel. Neither the control panel nor the instrument can tell that it is not using a cable.

There are several economic benefits to this approach. Wiring cost can exceed $10 per installed foot. The labor required to run this cable and conduit is not cheap, installation cost is a growing concern for designers and facility managers, and labor rates continue to rise in most parts of the world. Also, if these cables were in a hazardous environment (such as a chemical processing plant), they would have to be isolated from potential contact with chemicals and placed inside conduit run through concrete walls to reach the instrumentation deployed throughout the plant. This would entail additional costs and design problems.

Another benefit of wireless is the speed of deployment. Wired systems can take days to weeks to be properly installed, isolated, and commissioned. Wireless networks require only the endpoints to be installed, saving hours or days for each instrument that is installed. Other instruments can be added as required without the need for expensive, disruptive cabling and labor.

A further benefit of the wireless system is the ease of reconfiguration and expansion. If there is the need for a plant expansion or relocation of instruments, there is no expensive conduit to be moved or added. If the instruments to be connected to the control panel need to be placed on mobile equipment, such as the mobile batch containers found in biotech, pharmaceutical, and other specialty chemical installations, wireless offers an attractive solution.

14.2 Issues in Deploying Wireless Systems

It may seem that using wireless systems to replace common links, like serial cables, is easy and straightforward. That is not often the case. Here are some common problems encountered when replacing wired systems with wireless:

1. RF links are not as dependable as wires. Anyone who has used a cell phone, portable radio, or CB knows firsthand about RF links. The signal is constantly changing as conditions change between the two points.
2. Expanding or moving an RF point is not always as easy as claimed, because a new position on the network may be out of range of the control point for the wireless network. This control point is commonly placed at the control panel in an industrial application.
3. Wireless installers sometimes offer the assistance of professional technicians who perform RF site surveys to determine control points for the wireless network based on planned coverage areas. Although useful, this adds highly skilled labor back into the installation cost and doesn't address ease of reconfiguration or expansion.
4. Some installations require additional wireless control points (sometimes referred to as wireless access points) in addition to the control panel.

This is a very serious set of problems to face when reliability is of prime importance. Thus, despite the increasing popularity of IEEE 802.11 wireless local area network (LAN) systems

and the promise of Blue-tooth systems, wireless communication has yet to be widely adopted in industrial applications. Although wireless systems seem like an obvious solution for industrial applications, in reality the cure can be worse than the disease—due to the fact that solutions based on these standards were not designed with the industrial environment in mind. Industrial users need a network architecture that takes the unique challenges of the industrial environment into account.

The ARBORNET research team at the Massachusetts Institute of Technology (MIT) Media Lab examined embedded wireless solutions for smart devices from 1997 to 2001. The study concluded that traditional cell phone-style wireless systems were simply inadequate for industrial applications and could never gain widespread acceptance in their current form, but that a fresh approach was needed and, ultimately, was identified.

14.2.1 Control and Sensing Networks versus Data Networks

To understand why standard wireless networks do not work well for industrial applications, it is important to distinguish between *control and sensing* networks and *data* networks.

Wireless data networks are primarily designed to link together computers, personal digital assistants (PDAs), printers, Internet access points, and other elements where large amounts of data are sent in both directions. In data networks, the emphasis is on speed: faster is better. The design and evolution of 802.11 networks is a good example. 802.11b, one of the first wireless LAN standards to be widely adopted, connects devices at up to 11 Mbps. One of the latest updates to this standard, 802.11a, will allow for data speeds up to 54 Mbps, enabling more rapid downloads of music and video files by end-users.

But wireless networks for industrial control and sensing must be, above all, reliable, adaptable, and scalable. Because industrial sensors send only a few of bits of data per second or minute, providing information like temperature, pressure, and flow, data rates of 11 Mbps or even 54 Mbps are rarely needed. Although speed is often the focus for data networks, the primary design objectives for industrial control and sensing networks are reliability, adaptability, and scalability.

14.2.2 Requirement #1: Reliability

For most industrial applications, reliability is crucial: wireless systems must be just as reliable as traditional copper wire. Depending on the application, garbled or dropped data can result in anything from a disruptive glitch to a devastating failure. Three factors determine the signal reliability between a radio transmitter and receiver:

1. Path loss
2. RF interference
3. Transmit power

Consider a conversation between two people. Path loss corresponds to how muted a person's voice becomes to another, either due to distance or the obstacles between them. The listener will have a hard time understanding the speaker who is too far away or talking through a closed door.

RF interference corresponds to ambient noise: it will be difficult for the listener to understand the speaker in a noisy environment. Many other factors—including receiver sensitivity and data encoding technique—affect the reliability of a link. However, between a given radio transmitter and receiver, the path loss, interference, and transmit power determine the bit-error rate.

The problems of path loss or interference can be overcome by moving closer to the listener or by shouting loud enough to be heard. In the wireless world, this corresponds to repositioning radios or by transmitting with a higher power. Unfortunately, neither of these are generally viable options. Increasing the transmit power creates a situation similar to people shouting over loud music at a party. A sophisticated antenna design that directs the RF signal towards the receiving radio might help. But this is much like using a megaphone to shout at the listener. It does improve the path loss situation, but may fail if the listener or the megaphone moves.

14.2.3 Requirement #2: Adaptability

The network should adapt to the existing environment. The environment should not have to be altered to make the system "wireless ready." For example, if you need a wireless link between a tank level sensor and a data logger, it is not practical to relocate the tank or the data logger just to create a reliable connection. In fact, a wireless link may be unsuitable for connecting tank level sensors and data collection points in preexisting structures as these are often immovable objects. If cables were already being used for this, more wire could always be run, though at a prohibitive cost.

In the wireless world, the network should integrate seamlessly with the environment. A key attribute of a good wireless network is that daily work activities and the facility layout are not a concern. It's not desirable to ask someone to move in order to hear them speak more clearly. Likewise, repositioning radios and equipment in order to increase communication reliability is not always a realistic option.

14.2.4 Requirement #3: Scalability

Any network, wired or wireless, should be able to scale gracefully as the number of end-points increases. Scalability is one of the attractions of fieldbuses over hardwired "home run" systems: once the trunk line is in place, adding new devices is relatively easy. In many multi-drop networks, adding a new device is as simple as wiring the device directly into the network cable or a termination block at one end of the network. Eliminating the need to "home run" wire the new device back to the control panel has reduced wiring.

In a wireless system, all of the devices on the network share the airwaves. Simply transmitting with more power can increase the reliability of a single transmit/receive pair, but as soon as multiple devices share the airwaves, this approach may actually decrease overall reliability. This is not unlike being in a large, noisy restaurant where people are speaking loudly to be heard and no one can understand anyone else. Similarly, transmitting with more power in order to increase reliability is not consistent with building scalable networks that support numerous endpoints.

14.3 Wireless Formats

14.3.1 Point-to-Point Links

Sometimes referred to as a "wireless bridge," a point-to-point link serves as a replacement for a single communication cable. A point-to-point link might be used to connect a programmable logic controller (PLC) to a remote monitoring station, as shown in Figure 14.1.

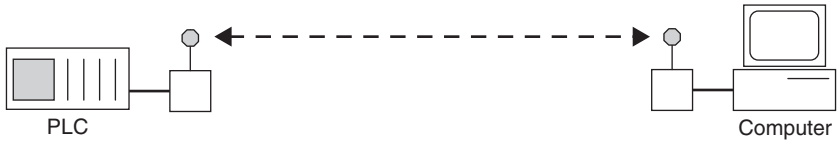


Figure 14.1: Point-to-Point Link

Point-to-point links can communicate reliably as long as the two endpoints are located sufficiently close to one another to escape the effects of RF interference and path loss. If a reliable connection is not initially achieved, it is sometimes possible to relocate the radios or boost the transmit power to achieve the desired reliability.

14.3.2 Point-to-Multipoint Links

Point-to-multipoint wireless systems, such as those based on IEEE 802.11 or Bluetooth, have one base station or access point that controls communication with all of the other wireless nodes in the network. Also referred to as a *hub and spoke* or *star* topology, this architecture has similarities to wired "home run" systems, in which all the signals converge on a single terminal block. A point-to-multipoint example is shown in Figure 14.2.

Signals in point-to-multipoint networks converge at a single endpoint. The reliability of these networks is set by the quality of the RF link between the central access point and each endpoint.

In industrial settings, it can be difficult to find a location for an access point that provides dependable communication with each endpoint. Moving an access point to improve communication with one endpoint will often degrade communication with other endpoints.

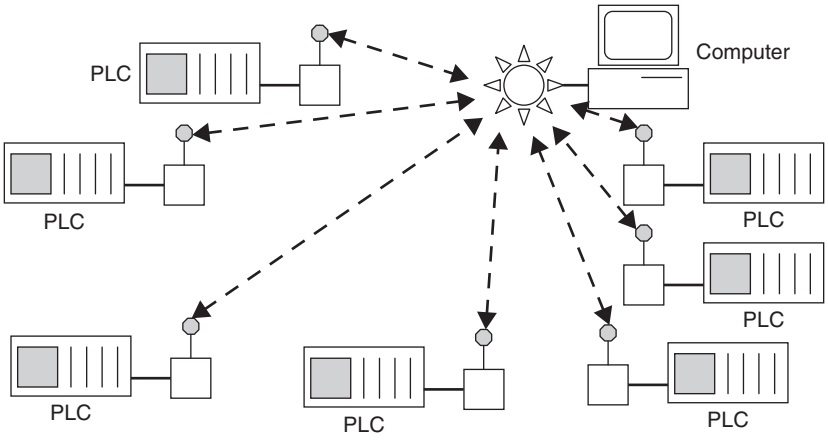


Figure 14.2: Point-to-Multipoint Network

Although it may be possible to wire together multiple access points in order to improve reliability, the cost of additional wiring can defeat the original reasons for choosing a wireless solution.

14.4 Wireless Mesh Networks

The wireless mesh network topology for industrial control and sensing that was developed by the MIT Media Lab and produced by the Ember Corporation is a "point-to-point-to-point" or "peer-to-peer" system called an ad hoc, multi-hop network. In a mesh network a node can send and receive messages, but it also functions as a router and can relay messages for its neighbors. Through this relaying process, a packet of wireless data will find its way to its ultimate destination, passing through intermediate nodes with reliable communication links. An example of a mesh network appears in Figure 14.3.

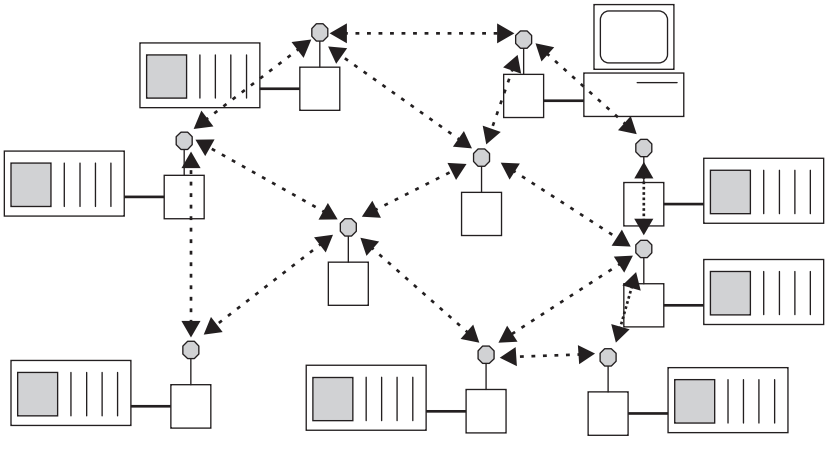


Figure 14.3: Wireless Mesh Network

There are some important things to notice in Figure 14.3:

1. The resemblance to a map of the Internet is not entirely coincidental. Like the Internet and other router-based communication networks, a mesh network offers multiple redundant communication paths throughout the network.

2. If a single node fails for any reason (including the introduction of strong RF interference), messages will automatically be routed through alternate paths.

3. In a mesh network, the distance between wireless nodes is short, which dramatically increases the link quality between nodes. If you reduce distance by a factor of two, the resulting signal is at least four times more powerful at the receiver. This makes links more reliable without increasing transmitter power in the individual nodes.

The advantages of mesh networks over point-to-point and point-to-multipoint configurations include:

- *More Nodes = Greater Reliability*: Notice the addition of freestanding "repeater" nodes in the middle of the network, as shown in Figure 14.3. In a mesh network, it is possible to extend distance, add redundancy, and improve the general reliability of the network simply by adding repeater nodes.

- *Mesh = Self-Configuring*: A network should not need a person to tell it how to get a message to its destination. A mesh network is self-organizing and does not require manual configuration. Because it is both self-configuring and self-healing, adding new gear or relocating existing gear is as simple as plugging in a wireless node and turning it on. The network discovers the new node and automatically incorporates it into the network without the need for a system administrator. A mesh network is not only inherently reliable, it is also highly adaptable. If your tank level sensor and data logger are placed too far apart for a solid RF communication link, you can quickly and easily lay down one or more repeater nodes to fill gaps in the network.

- *Mesh = Self-Healing*: On the Internet, if one router goes down, messages are sent through an alternate path by other routers. Similarly, if a device in a mesh network fails, messages are sent around it via other devices. The subtraction of one or more nodes does not necessarily affect its operation. A mesh network is self-healing because human intervention is not necessary for rerouting of messages.

- *Mesh = Redundant*: The actual meaning of "redundancy" in a real world is a matter of degree and must be carefully specified. In a mesh network, the degree of redundancy is essentially a function of node density. A mesh network can be deliberately

overdesigned simply by adding extra nodes, so that each device has two or more paths for sending data. This is a much simpler way of obtaining redundancy than is possible in most other types of systems.

- *Mesh = Scalable to Thousands of Nodes*: A mesh is also scalable and can handle hundreds or thousands of nodes. Since the operation of the network does not depend upon a central control point, adding multiple data collection points or gateways to other networks is convenient.

It is clear that reliability, adaptability, and scalability are the most important attributes of a wireless network for industrial control and sensing applications. Point-to-point networks can provide reliability, but they do not scale to handle more than one pair of endpoints. Point-to-multipoint networks can handle more endpoints, but the reliability is determined by the placement of the access point and the endpoints.

If environmental conditions result in poor reliability, it is difficult or impossible to adapt a point-to-multipoint network to increase the reliability. By contrast, mesh networks are inherently reliable, adapt easily to environmental or architectural constraints, and can scale to handle thousands of endpoints. These attributes are summarized in Table 14.1.

Table 14.1: Suitability in Industrial Applications

Topology	Reliability	Adaptability	Scalability
Point-to-Point	High	Low	None (2 endpoints)
Point-to-Multipoint	Low	Low	Moderate (7–30 endpoints)
Mesh Networks	High	High	Yes (1000s of endpoints)

14.5 Industrial Applications of Wireless Mesh Networks

14.5.1 Wire Replacement

At the beginning of this chapter, I gave an example of a wireless serial link replacement. This is most commonly done with point-to-point or point-to-multipoint technology, but mesh networks still provide complete transparency. The network does not know that copper has been replaced with an RF link, but the mesh network is inherently more reliable, more adaptable, and scalable.

14.5.2 Distributed Control

A specific opportunity for wireless, multi-hop, mesh networks is in distributed control systems. There has been a trend in recent years to place more intelligence throughout the control system. The IEEE 1451 standard *Smart Transducer Interface for Sensors and*

Actuators is evidence of this. Distributed intelligence is naturally served better by wireless multi-hop mesh networks, which do not require a central control topology.

The control of the wireless system is distributed throughout the network, allowing intelligent peers to communicate directly to other points on the network without having to be routed through some central point. Modular distributed control systems are easier to install and maintain. Since more of the system logic is at the instrument or subsystem level, clusters of instruments can interact and make local decisions. This is often done with small PLCs, which gather information from nearby instruments or sensors and then provide processing power and decision-making for this local instrument cluster. These clusters can then be connected as a group back into the main control system. The result is a less complex installation because individual instruments and points do not have to be directly connected to the main control panel.

14.5.3 Is Distributed Control Cheaper to Maintain?

Proponents insist that modular control systems are easier and less costly to maintain. The rationale is that highly modular control systems enable localized decision-making, which results in faster isolation of problems within the system. These problems can usually be diagnosed back to a single instrument cluster, allowing engineers and maintenance staff to focus their attention on one area of the system. Fast problem-solving means less downtime when something goes wrong. Likewise, when the system is operational, local decision-making by intelligent instruments and small PLCs identifies problems before they impact the entire system and cause bigger problems. Finally, these modular subsystems can be replaced or upgraded without affecting the entire system. These many factors make systems much cheaper and easier to operate and maintain.

Matching multi-hop, wireless mesh communication with distributed control facilitates a whole new dimension of interactions between sensors or sensor clusters. Sensors can now communicate directly to other devices on the network. This topology allows a tank level sensor to communicate directly with nearby valves, alerting them to open or close to prevent an overflow situation. Monitoring equipment could take readings from sensors without having to directly access the sensor with wired connections. This is useful in calibration and troubleshooting.

14.5.4 Diagnostic Monitoring

A third area of application for wireless, multi-hop mesh networks is in the diagnostic monitoring of devices. This monitoring can occur outside the normal control loop and wireless communication can be sent to notify the system user of any abnormal operation of the device. Take, for instance, the schematic of a sensor control loop shown in Figure 14.4. In this control loop, an additional signal is extracted and analyzed during the course of normal operation of

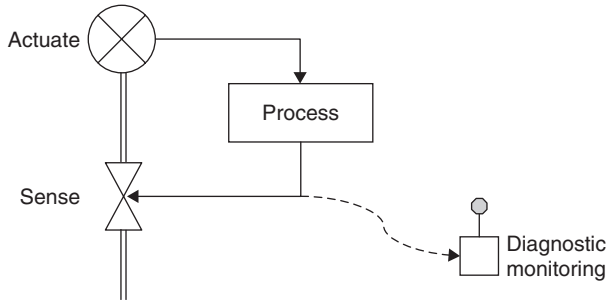


Figure 14.4: Schematic of Typical Sensor-Controlled Loop

the sensor. As the sensor operates, the signal is monitored for abnormalities without affecting the sensor's operation. If an abnormal signal or trend is observed, an alert is triggered.

The beauty of using a wireless link for onboard monitoring and alert is that the monitoring link remains independent of the control loop. By using a wireless, multi-hop mesh network, data can be routed dynamically to similar wireless devices. Surrounding devices can respond to the alert from the failing device, even as the alert is being sent to maintenance personnel. Another benefit of wireless is that maintenance personnel can directly connect to the diagnostic output of the sensor, without running wires. This can eliminate a huge task in the case of a tank level sensor in a large storage tank, or a temperature probe at the top of a tower stack at a chemical refinery. In a wireless, multi-hop mesh network, a user can get that data via any wireless node on the network.

By using a diagnostic device with additional processing power (e.g., a laptop computer, handheld computer, or handheld diagnostic device), maintenance personnel can check on configuration and other information about any node on the network. This information is a valuable tool for checking and verifying sensor operation when questionable data is received from a sensor through its primary control loop.

14.6 Case Study: Water Treatment

In order to validate wireless mesh networks in challenging industrial environments, Ember Corporation deployed a system in a water treatment plant. The environment was typical of such facilities, with significant wireless environment hurdles such as thick reinforced concrete walls segmenting giant tanks of water with large numbers of metal pipes running between tanks. The goal was to connect the instruments in the pipe gallery back to the control panel located in the control room on the third floor of the water filtration plant. Figure 14.5 provides a geographical representation of the instrumentation topology.

Figure 14.5 shows the approximate locations of eight instruments in the large pipe gallery along with four instruments in the small pipe gallery. The control room was located on the

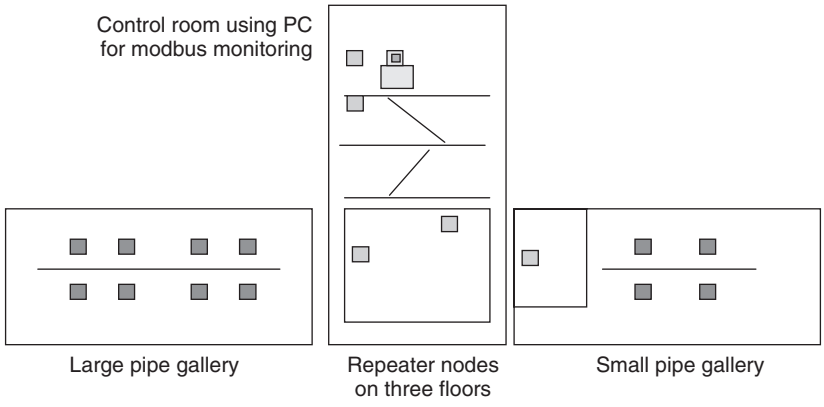


Figure 14.5: Schematic of Instrument Locations

third floor of an attached concrete building. Before wireless communication, a data collection PC in the control room communicated with process instruments over an RS-485 serial bus. The first step in converting this system to wireless networking was to replace this computer's bus connection with a wireless networking card connected to its serial port.

Each process instrument also had bus connections replaced with wireless networking cards, which self-configured on power-up and began attempting to send data to the control room. After all twelve instruments had wireless cards installed, it was possible to analyze the RF network traffic and determine where link reliability was below standards. These areas included spots where RF signals had to pass through reinforced concrete walls and where a single link spanned two flights of metal stairs. Improving these RF links was a simple matter of dropping down additional RF relay points. This step was made possible by the network's lack of a central wireless control point and each node's ability to cooperatively relay packets on behalf of its neighbors. After these repeater nodes were placed, the network was complete.

The time for complete installation was under 2 hours, compared to approximately 20 hours when each instrument had to be wired back to the control panel. The software on the PC did not discern any difference in the wireless communication network versus the wired serial cable network. The wireless network exhibited less than 0.1% packet loss before any attempt was made to resend lost packets through the network. This was accomplished via the mesh networking algorithms used by the wireless network. Neighboring nodes cooperatively relay packets over the best RF link.

14.7 Conclusion

Daily experience with some of the challenges of wireless consumer products, university research, and the commercial industry's slow adoption of wireless for use in enterprise

applications are indicators that products based on point-to-point and point-to-multipoint topologies are not well suited for use in industrial enterprise communication. Multi-hop mesh technology, however, is inherently reliable and redundant and can be extended to include thousands of devices. The real-world examples cited in this chapter demonstrate that mesh networks can be installed in hours instead of days or weeks and that these networks are highly dependable. Industrial systems can now benefit from a wireless format that satisfies the multiple conflicting demands of redundancy, distributed communication, flexibility, and reliability.

CHAPTER 15

Applications and Technologies

Alan Bensky

An important factor in the widespread penetration of short-range devices into the office and the home is the basing of the most popular applications on industry standards. In this chapter, we take a look at some of these standards and the applications that have emerged from them. Those covered pertain to HomeRF, Wi-Fi, HIPERLAN/2, Bluetooth, and Zigbee. In order to be successful, a standard has to be built so that it can keep abreast of rapid technological advancements by accommodating modifications that don't obsolete earlier devices that were developed to the original version. A case in point is the competition between the WLAN (wireless local area network) standard that was developed by the HomeRF Working Group based on the SWAP (shared wireless access protocol) specification, and IEEE specification 802.11, commonly known as Wi-Fi. The former used frequency-hopping spread-spectrum exclusively, and although some increase of data rate was provided for beyond the original 1 and 2 Mbps, it couldn't keep up with Wi-Fi, which incorporated new bandwidth efficient modulation methods to increase data rates 50-fold while maintaining compatibility with first generation DSSS terminals. Other reasons why HomeRF lost out to Wi-Fi are given below.

Many of the new wireless short-range systems are designed for operation on the 2.4 GHz ISM band, available for license-free operation in North America and Europe, as well as virtually all other regions in the world. Most systems have provisions for handling errors due to interference, but when the density of deployment of one or more systems is high, throughput, voice intelligibility, or quality of service in general is bound to suffer. We will look at some aspects of this problem and methods for solving it in relation to Bluetooth and Wi-Fi.

A relatively new approach to short-range communications with unique technological characteristics is ultra-wideband (UWB) signal generation and detection. UWB promises to add applications and users to short-range communication without impinging on present spectrum use. Additionally, it has other attributes including range finding and high power efficiency that are derived from its basic principles of operation. We present the main features of UWB communication and an introduction to how it works.

15.1 Wireless Local Area Networks (WLAN)

One of the hottest applications of short-range radio communication is wireless local area networks. While the advantage of a wireless versus wired LAN is obvious, the early versions

of WLAN had considerably inferior data rates so conversion to wireless was often not worthwhile, particularly when portability is not an issue. However, advanced modulation techniques have allowed wireless throughputs to approach and even exceed those of wired networks, and the popularity of highly portable laptop and handheld computers, along with the decrease in device prices, have made computer networking a common occurrence in multi-computer offices and homes.

There are still three prime disadvantages to wireless networks as compared to wired: range limitation, susceptibility to electromagnetic interference, and security. Direct links may be expected to perform at a top range of 50 to 100 meters depending on frequency band and surroundings. Longer distances and obstacles will reduce data throughput. Greater distances between network participants are achieved by installing additional access points to bridge remote network nodes. Reception of radio signals may be interfered with by other services operating on the same frequency band and in the same vicinity. Wireless transmissions are subject to eavesdropping, and a standardized security implementation in Wi-Fi called WEP (wired equivalent privacy), has been found to be breachable with relative ease by persistent and knowledgeable hackers. More sophisticated encryption techniques can be incorporated, although they may be accompanied by reduction of convenience in setting up connections and possibly in performance.

Various systems of implementation are used in wireless networks. They may be based on an industrial standard, which allows compatibility between devices by different manufacturers, or a proprietary design. The latter would primarily be used in a special purpose network, such as in an industrial application where all devices are made by the same manufacturer and where performance may be improved without the limitations and compromises inherent in a widespread standard.

15.1.1 The HomeRF Working Group

The HomeRF Working Group was established by prominent computer and wireless companies that joined together to establish an open industry specification for wireless digital communication between personal computers and consumer electronic devices anywhere in and around the home. It developed the SWAP specification—Shared Wireless Access Protocol, whose major application was setting up a wireless home network that connects one or more computers with peripherals for the purposes of sharing files, modems, printers, and other electronic devices, including telephones. In addition to acting as a transparent wire replacement medium, it also permitted integration of portable peripherals into a computer network. The originators expected their system to be accepted in the growing number of homes that have two or more personal computers.

Following are the main system technical parameters:

- Frequency-hopping network: 50 hops per second
- Frequency range: 2.4 GHz ISM band

- Transmitter power: 100 milliwatt
- Data rate: 1 Mbps using 2FSK modulation
  2 Mbps using 4FSK modulation
- Range: Covers typical home and yard
- Supported stations: Up to 127 devices per network
- Voice connections: Up to 6 full-duplex conversations
- Data security: Blowfish encryption algorithm (over 1 trillion codes)
- Data compression: LZRW3-A (Lempel-Ziv) algorithm
- 48-bit network ID: Enables concurrent operation of multiple co-located networks

The HomeRF Working Group ceased activity early in 2003. Several reasons may be cited for its demise. Reduction in prices of its biggest competitor, Wi-Fi, all but eliminated the advantage HomeRF had for home networks—low cost. Incompatibility with Wi-Fi was a liability, since people who used their Wi-Fi equipped laptop computer in the office also needed to use it at home, and a changeover to another terminal accessory after work hours was not an option. If there were some technical advantages to HomeRF, support of voice and connections between peripherals for example, they are becoming insignificant with the development of voice interfaces for Wi-Fi and the introduction of Bluetooth.

15.1.2 Wi-Fi

Wi-Fi is the generic name for all devices based on the IEEE specification 802.11 and its derivatives. It is promoted by the Wi-Fi Alliance that also certifies devices to ensure their interoperability. The original specification is being continually updated by IEEE working groups to incorporate technical improvements and feature enhancements that are agreed upon by a wide representation of potential users and industry representatives. 802.11 is the predominant industrial standard for WLAN and products adhering to it are acceptable for marketing all over the world.

802.11 covers the data link layer of lower-level software, the physical layer hardware definitions, and the interfaces between them. The connection between application software and the wireless hardware is the MAC (medium access control). The basic specification defines three types of wireless communication techniques: DSSS (direct sequence spread spectrum), FSSS (frequency-hopping spread spectrum) and IR (infra-red). The specification is built so that the upper application software doesn't have to know what wireless technique is being used—the MAC interface firmware takes care of that. In fact, application software doesn't have to know that a wireless connection is being used at all and mixed wired and wireless links can coexist in the same network.

Wireless communication according to 802.11 is conducted on the 2.400 to 2.4835 GHz frequency band that is authorized for unlicensed equipment operation in the United States and Canada and most European and other countries. A few countries allow unlicensed use in only a portion of this band. A supplement to the original document, 802.11b, adds increased data rates and other features while retaining compatibility with equipment using the DSSS physical layer of the basic specification. Supplement 802.11a specifies considerably higher rate operation in bands of frequencies between 5.2 and 5.8 GHz. These data rates were made available on the 2.4 GHz band by 802.11g that has downward compatibility with 802.11b.

15.1.3 Network Architecture

Wi-Fi architecture is very flexible, allowing considerable mobility of stations and transparent integration with wired IEEE networks. The transparency comes about because upper application software layers (see below) are not dependent on the actual physical nature of the communication links between stations. Also, all IEEE LAN stations, wired or wireless, use the same 48-bit addressing scheme so an application only has to reference source and destination addresses and the underlying lower-level protocols will do the rest.

Three Wi-Fi network configurations are shown in Figures 15.1 through 15.3. Figure 15.1 shows two unattached basic service sets (BSS), each with two stations (STA). The BSS is the basic building block of an 802.11 WLAN. A station can make *ad hoc* connections with other stations within its wireless communication range but not with those in another BSS that is outside of this range. In order to interconnect terminals that are not in direct range one with the other, the distributed system shown in Figure 15.2 is needed. Here, terminals that are in range of a station designated as an access point (AP) can communicate with other terminals not in direct range but who are associated with the same or another AP. Two or more such access points communicate between themselves either by a wireless or wired medium, and therefore data exchange between all terminals in the network is supported. The important thing here is that the media connecting the STAs with the APs, and connecting the APs among themselves are totally independent.

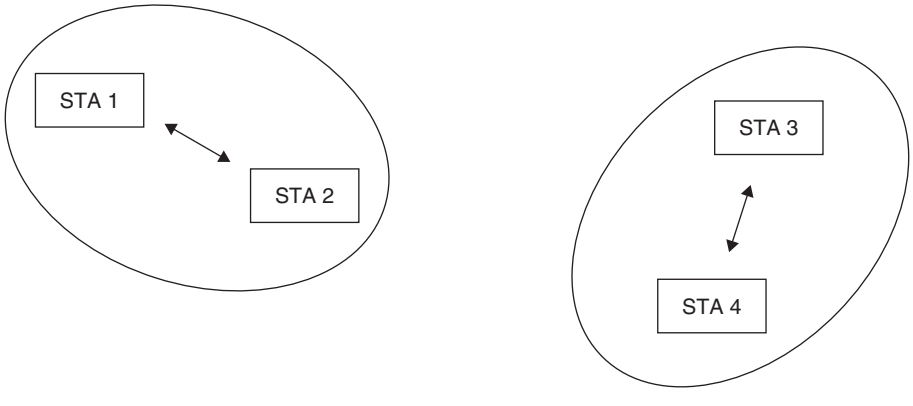


Figure 15.1: Basic Service Set

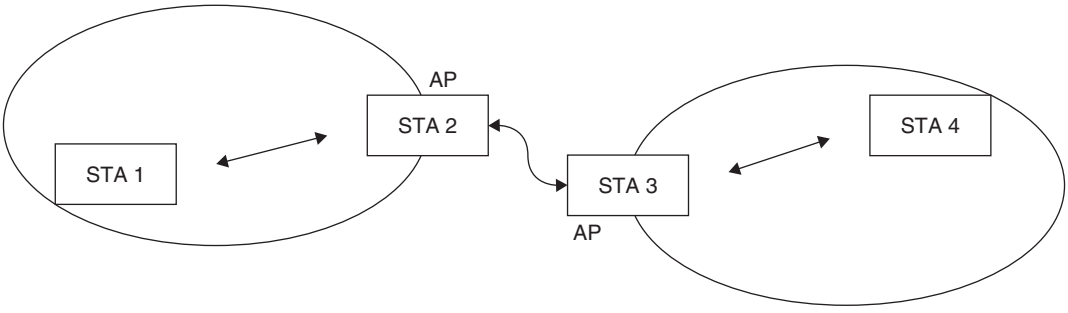


Figure 15.2: Distribution System and Access Points

A network of arbitrary size and complexity can be maintained through the architecture of the extended service set (ESS), shown in Figure 15.3. Here, STAs have full mobility and may move from one BSS to another while remaining in the network. Figure 15.3 shows another element type—a portal. The portal is a gateway between the WLAN and a wired LAN. It connects the medium over which the APs communicate to the medium of the wired LAN—coaxial cable or twisted pair lines, for example.

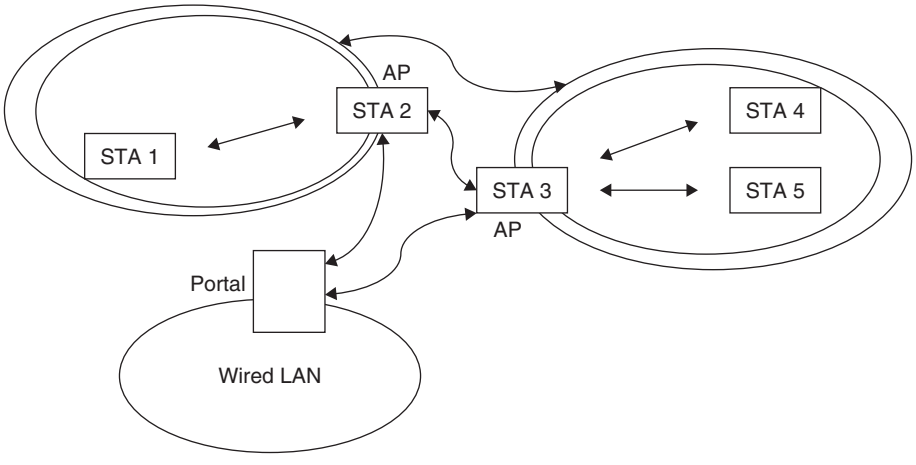


Figure 15.3: Extended Service Set

In addition to the functions Wi-Fi provides for distributing data throughout the network, two other important services, although optionally used, are provided. They are authentication and encryption. Authentication is the procedure used to establish the identity of a station as a member of the set of stations authorized to associate with another station. Encryption applies coding to data to prevent an eavesdropper from intercepting it.

802.11 details the implementation of these services in the MAC. Further protection of confidentiality may be provided by higher software layers in the network that are not part of 802.11.

The operational specifics of WLAN are described in IEEE 802.11 in terms of defined protocols between lower-level software layers. In general, networks may be described by the communication of data and control between adjacent layers of the Open System Interconnection Reference Model (OSI/RM), shown in Figure 15.4, or the peer-to-peer communication between like layers of two or more terminals in the network. The bottom layer, physical, represents the hardware connection with the transmission medium that connects the terminals of the network—cable modem, radio transceiver and antenna, infrared transceiver, or power line transceiver, for example. The software of the upper layers is wholly independent of the transmission medium and in principle may be used unchanged no matter what the nature of the medium and the physical connection to it. IEEE 802.11 is concerned only with the two lowest layers, physical and data link.

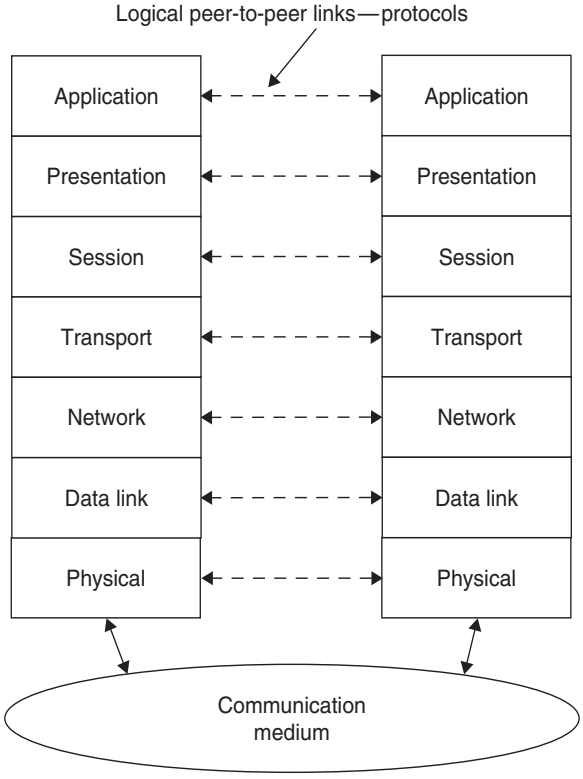


Figure 15.4: Open System Interconnection Reference Model

IEEE 802.11 prescribes the protocols between the MAC sublayer of the data layer and the physical layer, as well as the electrical specifications of the physical layer. Figure 15.5 illustrates the relationship between the physical and MAC layers of several types of networks

with upper-layer application software interfaced through a commonly defined logical link control (LLC) layer. The LLC is common to all IEEE local area networks and is independent of the transmission medium or medium access method. Thus, its protocol is the same for wired local area networks and the various types of wireless networks. It is described in specification ANSI/IEEE standard 802.2.

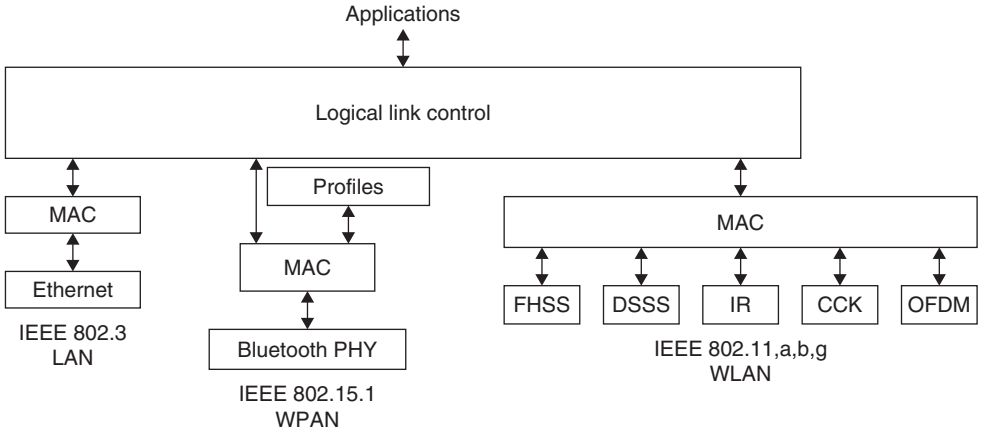


Figure 15.5: Data Link and Physical Layers (PHY)

The Medium Access Control function is the brain of the WLAN. Its implementation may be as high-level digital logic circuits or a combination of logic and a microcontroller or a digital signal processor. IEEE 802.11 and its supplements, (which may be generally designated 802.11x), prescribe various data rates, media (radio waves or infrared), and modulation techniques (FHSS, DSSS, CCK, ODFM). These are the principle functions of the MAC:

- Frame delimiting and recognition,
- Addressing of destination stations,
- Transparent transfer of data, including fragmentation and defragmentation of packets originating in upper layers,
- Protection against transmission error,
- Control of access to the physical medium,
- Security services—authentication and encryption.

An important attribute of any communications network is the method of access to the medium. 802.11 prescribes two possibilities: DCF (distributed coordination function) and PCF (point coordination function).

The fundamental access method in IEEE 802.11 is the DCF, more widely known as CSMA/CA (carrier sense multiple access with collision avoidance). It is based on a procedure during which a station wanting to transmit may do so only after listening to the channel and determining that it is not busy. If the channel is busy, the station must wait until the channel is idle. In order to minimize the possibility of collisions when more than one station wants to transmit at the same time, each station waits a random time-period, called a back off interval, before transmitting, after the channel goes idle. Figure 15.6 shows how this method works.

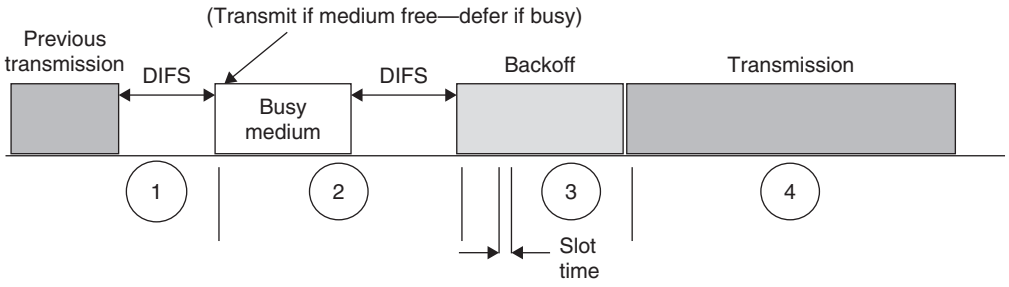


Figure 15.6: CSMA/CA Access Method

The figure shows activity on a channel as it appears to a station that is attempting to transmit. The station may start to transmit if the channel is idle for a period of at least a duration of DIFS (distributed coordination function interframe space) since the end of any other transmission (Section 1 of the figure). However, if the channel is busy, as shown in Section 2 of the figure, it must defer access and enter a back off procedure.

The station waits until the channel is idle, and then waits an additional period of DIFS. Now it computes a time-period called a back off window that equals a pseudo-random number multiplied by a constant called the "slot time." As long as the channel is idle, as it is in Section 3 of the figure, the station may transmit its frame at the end of the back off window, Section 4. During every slot time of the back off window the station senses the channel, and if it is busy, the counter that holds the remaining time of the back off window is frozen until the channel becomes idle and the back off counter resumes counting down.

Actually, the back off procedure is not used for every access of the channel. For example, acknowledgement transmissions and RTS and CTS transmissions (see below), do not use it. Instead, they access the channel after an interval called SIFS (short interframe space) following the transmission to which they are responding. SIFS is shorter than DIFS, so other stations waiting to transmit cannot interfere since they have to wait a longer time, after the previous transmission, and by then the channel is already occupied.

In waiting for a channel to become idle, a transmission contender doesn't have to listen continuously. When one hears another station access the channel, it can interpret the frame

length field that is transmitted on every frame. After taking into account the time of the acknowledgement transmission that replies to a data transmission, the time that the channel will become idle is known even without physically sensing it. This is called a virtual carrier sense mechanism.

The procedure shown in Figure 15.6 may not work well under some circumstances. For example, if several stations are trying to transmit to a single access point, two or more of them may be positioned such that they all are in range of the access point but not of each other. In this case, a station sensing the activity of the channel may not hear another station that is transmitting on the same network. A refinement of the described CSMA/SA procedure is for a station thinking the channel is clear to send a short RTS (request to send) control frame to the AP. It will then wait to receive a CTS (clear to send) reply from the AP, which is in range of all contenders for transmission, before sending its data transmission. If the originating station doesn't hear the CTS it assumes the channel was busy and so it must try to access the channel again. This RTS/CTS procedure is also effective when not all stations on the network have compatible modulation facilities for high rate communication and one station may not be able to detect the transmission length field of another. RTS and CTS transmissions are always sent at a basic rate that is common to all participants in the network.

The PCS is an optional access method that uses a master-slave procedure for polling network members. An AP station assumes the role of master and distributes timing and priority information through beacon management transmissions, thus creating a contention free access method. One use of the PCS is for voice communications, which must use regular time slots and will not work in a random access environment.

15.1.4 Physical Layer

The discussion so far on the services and the organization of the WLAN did not depend on the actual type of wireless connection between the members of the network. 802.11 and its additions specify various bit rates, modulation methods, and operating frequency channels, on two frequency bands, which we discuss in this section.

15.1.4.1 IEEE 802.11 Basic

The original version of the 802.11 specification prescribes three different air interfaces, each having two data rates. One is infrared and the others are based on frequency-hopping spread spectrum (FHSS) and direct-sequence spread-spectrum, each supporting raw data rates of 1 and 2 Mbps. Below is a short description of the IR and FHSS links, and a more detailed review of DSSS.

15.1.4.2 Infrared PHY

Infrared communication links have some advantages over radio wave transmissions. They are completely confined within walled enclosures and therefore eavesdropping concerns

are greatly relieved, as are problems from external interference. Also, they are not subject to intentional radiation regulations. The IEEE 802.11 IR physical layer is based on diffused infrared links, and the receiving sensor detects radiation reflected off ceilings and walls, making the system independent of line-of-site. The range limit is on the order of 10 meters. Baseband pulse position modulation is used, with a nominal pulse width of 250 nsec. The IR wavelength is between 850 and 950 nM. The 1 Mbps bit rate is achieved by sending symbols representing 4 bits, each consisting of a pulse in one of 16 consecutive 250 nsec slots. This modulation method is called 16-PPM. Optional 4-PPM modulation, with four slots per two-bit symbol, gives a bit rate of 2 Mbps.

Although part of the original IEEE 802.11 specification and having what seems to be useful characteristics for some applications, products based on the infrared physical layer for WLAN have generally not been commercially available. However, point-to-point, very short-range infrared links using the IrDA (Infrared Data Association) standard are very widespread (reputed to be in more than 300 million devices). These links work reliably line-of-site at one meter and are found, for example, in desktop and notebook computers, handheld PC's, printers, cameras and toys. Data rates range from 2400 Bps to 16 Mbps. Bluetooth devices will take over some of the applications but for many cases IrDA embedding will still have an advantage because of its much higher data rate capability.

15.1.4.3 FHSS PHY

While overshadowed by the DSSS PHY, acquaintance with the FHSS option in 802.11 is still useful since products based on it may be available. In FHSS WLAN, transmissions occur on carrier frequencies that hop periodically in pseudo-random order over almost the complete span of the 2.4 GHz ISM band. This span in North America and most European countries is 2.400 to 2.4835 GHz, and in these regions there are 79 hopping carrier frequencies from 2.402 to 2.480 GHz. The dwell on each frequency is a system-determined parameter, but the recommended dwell time is 20 msec, giving a hop rate of 50 hops per second. In order for FHSS network stations to be synchronized, they must all use the same pseudo-random sequence of frequencies, and their synthesizers must be in step, that is, they must all be tuned to the same frequency channel at the same time. Synchronization is achieved in 802.11 by sending the essential parameters—dwell time, frequency sequence number, and present channel number—in a frequency parameter set field that is part of a beacon transmission (and other management frames) sent periodically on the channel. A station wishing to join the network can listen to the beacon and synchronize its hop pattern as part of the network association procedure.

The FHSS physical layer uses GFSK (Gaussian frequency shift keying) modulation, and must restrict transmitted bandwidth to 1 MHz at 20 dB down (from peak carrier). This bandwidth holds for both 1 Mbps and 2 Mbps data rates. For 1 Mbps data rate, nominal frequency deviation is ± 160 kHz. The data entering the modulator is filtered by a Gaussian (constant phase delay)

filter with 3 dB bandwidth of 500 kHz. Receiver sensitivity must be better than −80 dBm for a 3% frame error rate.

In order to keep the same transmitted bandwidth with a data rate of 2 Mbps, four-level frequency shift-keying is employed. Data bits are grouped into symbols of two bits, so each symbol can have one of four levels. Nominal deviations of the four levels are ±72 kHz and ±216 kHz. A 500 kHz Gaussian filter smoothes the four-level 1 Megasymbols per second at the input to the FSK modulator. Minimum required receiver sensitivity is −75 dBm.

Although development of Wi-Fi for significantly increased data rates has been along the lines of DSSS, FHSS does have some advantageous features. Many more independent networks can be collocated with virtually no mutual interference using FHSS than with DSSS. As we will see later, only three independent DSSS networks can be collocated. However, 26 different hopping sequences (North America and Europe) in any of three defined sets can be used in the same area with low probability of collision. Also, the degree of throughput reduction by other 2.4 GHz band users, as well as interference caused to the other users is lower with FHSS. FHSS implementation may at one time also have been less expensive. However, the updated versions of 802.11—specifically 802.11a, 802.11b, and 802.11g—have all based their methods of increasing data rates on the broadband channel characteristics of DSSS in 802.11, while being downward compatible with the 1 and 2 Mbps DSSS modes (except for 802.11a which operates on a different frequency band).

15.1.4.4 DSSS PHY

The channel characteristics of the direct sequence spread spectrum physical layer in 802.11 are retained in the high data rate updates of the specification. This is natural, since systems based on the newer versions of the specification must retain compatibility with the basic 1 and 2 Mbps physical layer. The channel spectral mask is shown in Figure 15.7, superimposed on the simulated spectrum of a filtered 1 Mbps transmission. It is 22 MHz wide at the −30 dB points. Fourteen channels are allocated in the 2.4 GHz ISM band, whose center frequencies are 5 MHz apart, from 2.412 GHz to 2.484 GHz. The highest channel, number fourteen, is designated for Japan where the allowed band edges are 2.471 GHz and 2.497 GHz. In the US and Canada, the first eleven channels are used. Figure 15.8 shows how channels one, six and eleven may be used by three adjacent independent networks without co-interference. When there are no more than two networks in the same area, they may choose their operating channels to avoid a narrow-band transmission or other interference on the band.

In 802.11 DSSS, a pseudo-random bit sequence phase modulates the carrier frequency. In this spreading sequence, bits are called chips. The chip rate is 11 megachips per second (Mcps). Data is applied by phase modulating the spread carrier. There are eleven chips per data symbol. The chosen pseudo-random sequence is a Barker sequence, represented as 1,−1,1,1, −1,1,1,1,−1,−1,−1. Its redeeming property is that it is optimally detected in a receiver by

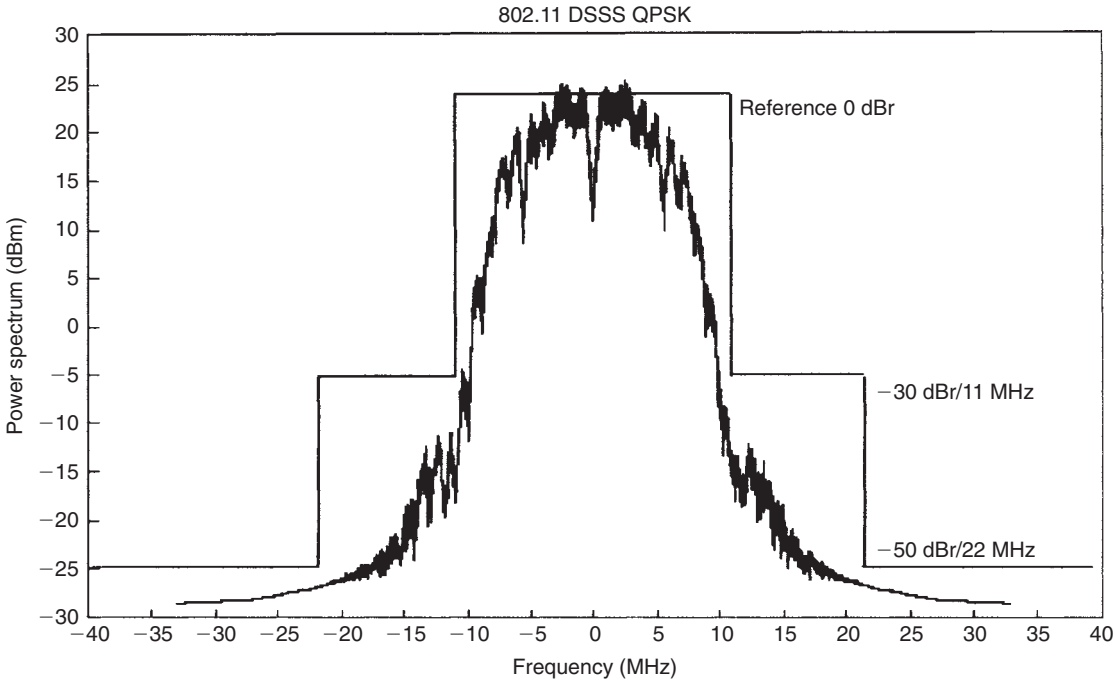


Figure 15.7: 802.11 DSSS Spectral Mask

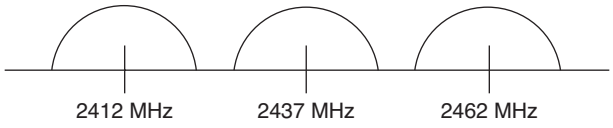


Figure 15.8: DSSS Non-interfering Channels

a matched filter or correlation detector. Figure 15.9 is one possible implementation of the modulator. The DSSS PHY specifies two possible data rates—1 and 2 Mbps. The differential encoder takes the data stream and produces two output streams at 1 Mbps that represent changes in data polarity from one symbol to the next. For a data rate of 1 Mbps, differential binary phase shift keying is used. The input data rate of 1 Mbps results in two identical output data streams that represent the changes between consecutive input bits. Differential quadrature phase shift keying handles 2 Mbps of data. Each sequence of two input bits creates four permutations on two outputs. The differential encoder outputs the differences from symbol to symbol on the lines that go to the inputs of the exclusive OR gates shown in Figure 15.9. The outputs on the *I* and *Q* lines are the Barker sequence of 11 Mcps inverted or sent straight through, at a rate of 1 Msps, according to the differentially encoded data at the exclusive OR gate inputs. These outputs are spectrum shifted to the RF carrier frequency (or an intermediate frequency for subsequent up-conversion) in the quadrature modulator.

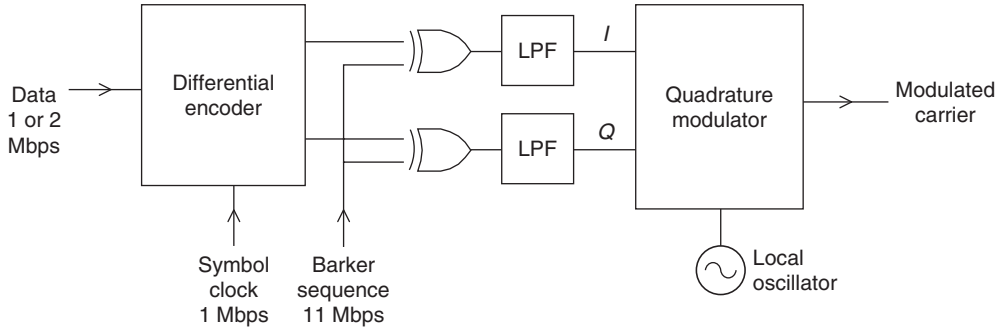


Figure 15.9: DSSS Modulation

Reception of DSSS signals is represented in Figure 15.10. The downconverted *I* and *Q* signals are applied to matched filters or correlation detectors. These circuits correlate the Barker sequence with the input signal and output an analog signal that represents the degree of correlation. The following differential decoder performs the opposite operation of the differential encoder described above and outputs the 1 or 2 Mbps data.

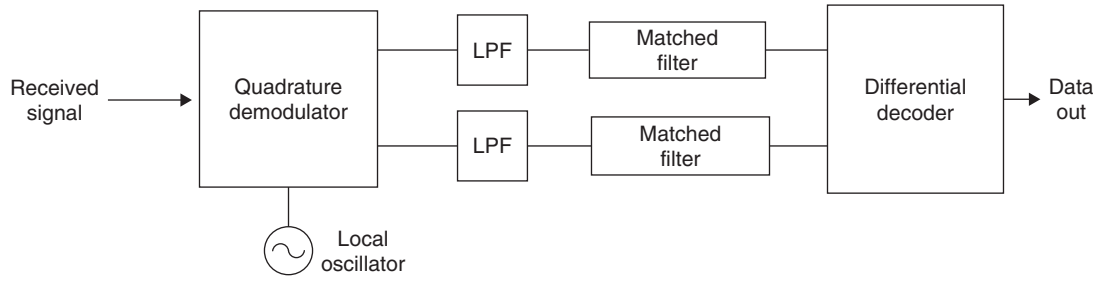


Figure 15.10: DSSS Reception

The process of despreading the input signal by correlating it with the stored spreading sequence requires synchronization of the receiver with transmitter timing and frequency. To facilitate this, the transmitted frame starts with a synchronization field (SYNC), shown at the beginning of the physical layer protocol data unit in Figure 15.11. Then a start frame delimiter (SFD) marks out the commencement of the following information bearing fields. All bits in the indicated preamble are transmitted at a rate of 1 Mbps, no matter what the subsequent data rate will be. The signal field specifies the data rate of the following fields in the frame so that the receiver can adjust itself accordingly. The next field, SERVICE, contains all zeros for devices that are only compliant with the basic version of 802.11, but some of its bits are used in devices conforming with updated versions. The value of the length field is the length, in microseconds, required to transmit the data-carrying field labeled MPDU (MAC protocol data unit). An error check field, labeled CRC, protects the integrity of the SIGNAL, SERVICE, and

LENGTH fields. The last field MPDU (MAC protocol data unit) is the data passed down from the MAC to be sent by the physical layer, or to be passed up to the MAC after reception. All bits in the transmitted frame are pseudo-randomly scrambled to ensure even power distribution over the spectrum. Data is returned to its original form by descrambling in the receiver.

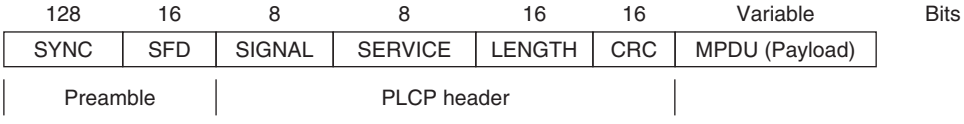


Figure 15.11: DSSS Frame Format

15.1.4.5 802.11b

The "b" supplement to the original 802.11 specification supports a higher rate physical layer for the 2.4 GHz band. It is this 802.11b version that provided the impetus for Wi-Fi proliferation. With it, data rates of 5.5 Mbps and 11 Mbps are enabled, while retaining downward compatibility with the original 1 and 2 Mbps rates. The slower rates may be used not only for compatibility with devices that aren't capable of the extended rates, but also for fall back when interference or range conditions don't provide the required signal-to-noise ratio for communication using the higher rates.

As previously stated, the increased data rates provided for in 802.11b do not entail a larger channel bandwidth. Also, the narrow-band interference rejection, or jammer resisting qualities of direct sequence spread-spectrum are retained. The classical definition of processing gain for DSSS as being the chip rate divided by the data bandwidth doesn't apply here. In fact, the processing gain requirement that for years was part of the FCC Rules paragraph 15.247 definition of direct sequence spread-spectrum was deleted in an update from August 2002, and at the same time reference to DSSS was replaced by "digital modulation."

The mandatory high-rate modulation method of 802.11b is called complementary code keying (CCK). An optional mode called packet binary convolutional coding (PBCC) is also described in the specification. Although there are similarities in concept, the two modes differ in implementation and performance. First the general principle of high-rate DSSS is presented below, applying to both CCK and PBCC, then the details of CCK are given.

As in the original 802.11, a pseudo-random noise sequence at the rate of 11 Mcps is the basis of high-rate transmission in 802.11b. It is this 11 Mcps modulation that gives the 22 MHz null-to-null bandwidth. However, in contrast to the original specification, the symbol rate when sending data at 5.5 or 11 Mbps is 1.375 Msps. Eight chips per symbol are transmitted instead of eleven chips per symbol as when sending at 1 or 2 Mbps. In "standard" DSSS as used in 802.11, the modulation, BPSK or QPSK, is applied to the group of eleven chips constituting a symbol. The series of eleven chips in the symbol is always the same (the Barker sequence

previously defined). In contrast, high-rate DSSS uses a different 8-chip sequence in each symbol, depending on the sequence of data bits that is applied to each symbol. Quadrature modulation is used, and each chip has an *I* value and a *Q* value which represent a complex number having a normalized amplitude of one and some angle, α, where α = arctangent (*Q*/*I*). α can assume one of four values divided equally around 360 degrees. Since each complex bit has four possible values, there are a total of $4^8 = 65536$ possible 8-bit complex words. For the 11 Mbps data rate, 256 out of these possibilities are actually used—which one being determined by the sequence of 8 data bits applied to a particular symbol. Only 16-chip sequences are needed for the 5.5 Mbps rate, determined by four data bits per symbol. The high-rate algorithm describes the manner in which the 256 code words, or 16 code words, are chosen from the 65536 possibilities. The chosen 256 or 16 complex words have the very desirable property that when correlation detectors are used on the *I* and *Q* lines of the received signal, downconverted to baseband, the original 8-bit (11 Mbps rate) or 4-bit (5.5 Mbps rate) sequence can be decoded correctly with high probability even when reception is accompanied by noise and other types of channel distortion.

The concept of CCK modulation and demodulation is shown in Figures 15.12 and 15.13. It's explained below in reference to a data rate of 11 Mbps. The multiplexer of Figure 15.12 takes a block of eight serial data bits, entering at 11 Mbps, and outputs them in parallel, with updates at the symbol rate of 1.375 MHz. The six latest data bits determine 1 out of 64 (2^6) complex code words. Each code word is a sequence of eight complex chips, having phase angles α_1 through α_8 and a magnitude of unity. The first two data bits, d_0 and d_1, determine an angle, α_8' which, in the code rotator (see Figure 15.12), rotates the whole code word relative to α_8 of the previous code word. This angle of rotation becomes the absolute angle α_8 of the present code word. The normalized *I* and *Q* outputs of the code rotator, which after filtering are input to a quadrature modulator for up-conversion to the carrier (or intermediate) frequency, are shown in equation [15.1]:

$$I_i = \cos(\alpha_i), \quad Q_i = \sin(\alpha_i) \qquad i = 1 \ldots 8 \tag{15.1}$$

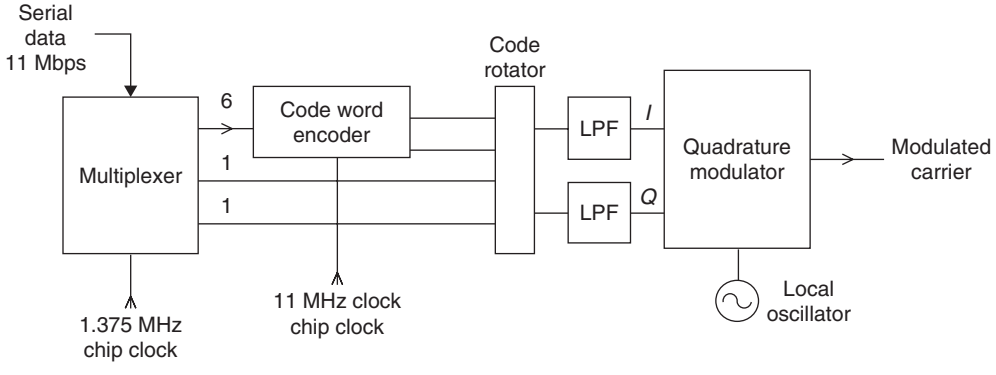


Figure 15.12: High-rate Modulator—11 Mbps

Data symbol: $d_0\ d_1\ d_2\ d_3\ d_4\ d_5\ d_6\ d_7$

Phase table	
d_i d_{i+1}	φ
0 0	0°
1 0	180°
0 1	90°
1 1	−90°

Phase $(d_0, d_1) = \varphi_1$
Phase $(d_2, d_3) = \varphi_2$
Phase $(d_4, d_5) = \varphi_3$
Phase $(d_6, d_7) = \varphi_4$

$\alpha_1 = \varphi_1 + \varphi_2 + \varphi_3 + \varphi_4$
$\alpha_2 = \varphi_1 + \varphi_3 + \varphi_4$
$\alpha_3 = \varphi_1 + \varphi_2 + \varphi_4$
$\alpha_4 = \varphi_1 + \varphi_4 + 180°$
$\alpha_5 = \varphi_1 + \varphi_2 + \varphi_3$
$\alpha_6 = \varphi_1 + \varphi_3$
$\alpha_7 = \varphi_1 + \varphi_2 + 180°$
$\alpha_8 = \varphi_1$

$I_i = \cos(\alpha_i)$
$Q_i = \sin(\alpha_i)$
$i = 1 \ldots 8$

Figure 15.13: Derivation of Code Word

Figure 15.13 is a summary of the development of code words a for 11 Mbps rate CCK modulation. High rate modulation is applied only to the payload—MPDU in Figure 15.11. The code word described in Figure 15.13 is used as shown for the first symbol and then every other symbol of the payload. However, it is modified by adding 180º to each element of the code word of the second symbol, fourth symbol, and so on.

The development of the symbol code word or chip sequence may be clarified by an example worked out per Figure 15.13. Let's say the 8-bit data sequence for a symbol is $\boldsymbol{d} = d_0 \ldots d_7 =$ 1 0 1 0 1 1 0 1. From the phase table of Figure 15.13 we find the angles φ: $\varphi_1 = 180º$, $\varphi_2 = 180º$, $\varphi_3 = -90º$, $\varphi_4 = 90º$. Now summing up these values to get the angle α_i of each complex chip, then taking the cosine and sine to get I_i and Q_i, we summarize the result in the following table:

i	1	2	3	4	5	6	7	8
α	0	180	90	90	−90	90	180	180
I	1	−1	0	0	0	0	−1	−1
Q	0	0	1	1	−1	1	0	0

The code words for 5.5 Mbps rate CCK modulation are a subset of those for 11 Mbps CCK. In this case, there are four data bits per symbol which determine a total of 16 complex chip sequences. Four 8-element code words (complex chip sequences) are determined using the last two data bits of the symbol, d_2 and d_3. The arguments (angles) of these code words are shown in Table 15.1. Bits d_0 and d_1 are used to rotate the code words relative to the preceding code word as in 11 Mbps modulation and shown in the phase table of Figure 15.13. Code words are modified by 180º every other symbol, as in 11 Mbps modulation.

Table 15.1: 5.5 Mbps CCK Decoding

d_3, d_2	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
00	90°	0°	90°	180°	90°	0°	−90°	0°
10	−90°	180°	−90°	0°	90°	0°	−90°	0°
01	−90°	0°	−90°	180°	−90°	0°	90°	0°
11	90°	180°	90°	0°	−90°	0°	90°	0°

The concept of CCK decoding for receiving high rate data is shown in Figure 15.14. For the 11 Mbps data rate, a correlation bank decides which of the 64 possible codes best fits each received 8-bit symbol. It also finds the rotation angle of the whole code relative to the previous symbol (one of four values). There are a total of 256 (64 × 4) possibilities and the chosen one is output as serial data. At the 5.5 Mbps rate there are four code words to choose from and after code rotation a total of 16 choices from which to decide on the output data.

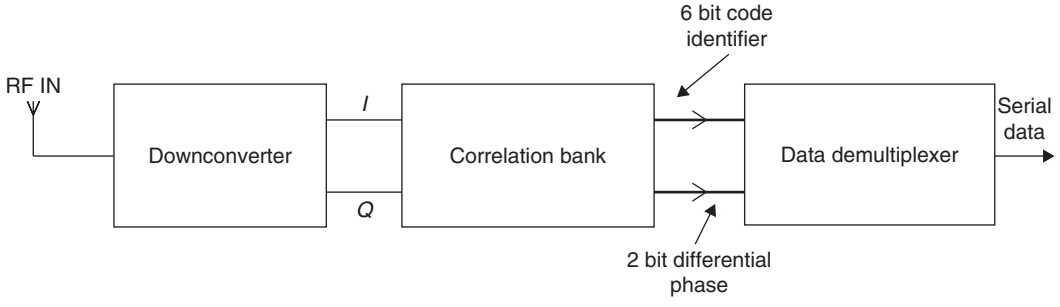


Figure 15.14: CCK Decoding

To maintain compatibility with earlier non high-rate systems, the DSSS frame format shown in Figure 15.11 is retained in 802.11b. The 128-bit preamble and the header are transmitted at 1 Mbps while the payload MPDU can be sent at a high rate of 5.5 or 11 Mbps. The long and slow preamble reduces the throughput and cancels some of the advantage of the high data rates. 802.11b defines an optional short preamble and header which differ from the standard frame by sending a preamble with only 72 bits and transmitting the header at 2 Mbps, for a total overhead of 96 μsec instead of 192 μsec for the long preamble and header. Devices using this option can only communicate with other stations having the same capability.

Use of higher data rates entails some loss of sensitivity and hence range. The minimum specified sensitivity at the 11 Mbps rate is −76 dBm for a frame-error rate of 8% when sending a payload of 1024 bytes, as compared to a sensitivity of −80 dBm for the same frame-error rate and payload length at a data rate of 2 Mbps.

15.1.4.6 802.11a and OFDM

In the search for ways to communicate at even higher data rates than those applied in 802.11b, a completely different modulation scheme, OFDM (orthogonal frequency division multiplexing) was adopted for 802.11a. It is not DSSS yet it has a channel bandwidth similar to the DSSS systems already discussed. The 802.11*a* supplement is defined for channel frequencies between 5.2 and 5.85 GHz, obviously not compatible with 802.11b signals in the 2.4 GHz band. However, since the channel occupancy characteristics of its modulation are similar to that of DSSS Wi-Fi, the same system was adopted in IEEE 802.11*g* for enabling the high data rates of 802.11a on the 2.4 GHz band, while allowing downward compatibility with transmissions conforming to 802.11b.

802.11a specifies data rates of 6, 9, 12, 18, 24, 36, 48, and 54 Mbit/s. As transmitted data rates go higher and higher, the problem of multipath interference becomes more severe. Reflections in an indoor environment can result in multipath delays on the order of 100 nsec but may be as long as 250 nsec, and a signal with a bit rate of 10 Mbps (period of 100 nsec) can be completely overlapped by its reflection. When there are several reflections, arriving at the receiver at different times, the signal may be mutilated beyond recognition. The OFDM transmission system goes a long way to solving the problem. It does this by sending the data partitioned into symbols whose length in time is several times the expected reflected path length time differences. The individual data bits in a symbol are all sent in parallel on separate subcarrier frequencies within the transmission channel. Thus, by sending many bits during the same time, each on a different frequency, the individual transmitted bit can be lengthened so that it won't be affected by the multipath phenomenon. Actually, the higher bit rates are accommodated by representing a group of data bits by the phase and amplitude of a particular transmitted carrier. A carrier modulated using quadrature phase shift keying (QPSK) can represent two data bits and 64-QAM (quadrature amplitude modulation) can present six data bits as a single data unit on a subcarrier.

Naturally, transmitting many subcarriers on a channel of given width brings up the problem of interference between those subcarriers. There will be no interference between them if all the subcarriers are orthogonal—that is, if the integral of any two different subcarriers over the symbol period is zero. It is easy to show that this condition exists if the frequency difference between adjacent subcarriers is the inverse of the symbol period.

In OFDM, the orthogonal subcarriers are generated mathematically using the inverse Fourier transform (IFT), or rather its discrete equivalent, the inverse discrete Fourier transform (IDFT). The IDFT may be expressed as shown in equation [15.2]:

$$x(n) = \frac{1}{N} \sum_{m=0}^{N-1} X(m)[\cos(2\pi mn/N) + j \cdot \sin(2\pi mn/N)] \tag{15.2}$$

$x(n)$ are complex sample values in the time domain, $n = 0 \ldots N - 1$, and $X(m)$ are the given complex values, representing magnitude and phase, for each frequency in the frequency domain. The IDFT expression indicates that the time domain signal is the sum of N harmonically related sine and cosine waves each of whose magnitude and phase is given by $X(m)$. We can relate the right side of the expression to absolute frequency by multiplying the arguments $2\pi mn/N$ by f_s/f_s to get:

$$x(n) = \frac{1}{N} \sum_{m=0}^{N-1} X(m)[\cos(2\pi m f_1 n t_s) + j \cdot \sin(2\pi m f_1 n t_s)] \tag{15.3}$$

where f_1 is the fundamental subcarrier and the difference between adjacent subcarriers, and t_s is the sample time $1/f_1$. In 802.11a OFDM, the sampling frequency is 20 MHz and $N = 64$, so $f_1 = 312.5$ kHz. Symbol time is $Nt_s = 64/f_s = 3.2\,\mu$sec.

In order to prevent intersymbol interference, 802.11a inserts a guard time of 0.8 μsec in front of each symbol, after the IDFT conversion. During this time, the last 0.8 μsec of the symbol is copied, so the guard time is also called a circular prefix. Thus, the extended symbol time that is transmitted is $3.2 + .8 = 4\,\mu$sec. The guard time is deleted after reception and before reconstruction of the transmitted data.

Although the previous equation, where $N = 64$, indicates 64 possible subcarriers, only 48 are used to carry data, and four more for pilot signals to help the receiver phase lock to the transmitted carriers. The remaining carriers that are those at the outside of the occupied bandwidth, and the DC term ($m = 0$ in equation (15.3)), are null. It follows that there are 26 ((48 + 4)/2) carriers on each side of the nulled center frequency. Each channel width is 312.5 kHz, so the occupied channels have a total width of 16.5625 (53×312.5 kHz) MHz.

For accommodating a wide range of data rates, four modulation schemes are used—BPSK, QPSK, 16-QAM and 64-QAM, requiring 1, 2, 4, and 6 data bits per symbol, respectively. Forward error correction (FEC) coding is employed with OFDM, which entails adding code bits in each symbol. Three coding rates: 1/2, 2/3, and 3/4, indicate the ratio of data bits to the total number of bits per symbol for different degrees of coding performance. FEC permits reconstruction of the correct message in the receiver, even when one or more of the 48 data channels have selective interference that would otherwise result in a lost symbol. Symbol bits are interleaved so that even if adjacent subcarrier bits are demodulated with errors, the error correction procedure will still reproduce the correct symbol. A block diagram of the OFDM transmitter and receiver is shown in Figure 15.15. Blocks FFT and IFFT indicate the fast Fourier transform and its inverse instead of the mathematically equivalent (in terms of results) discrete Fourier transform and inverse discrete Fourier transform (IFDT) that we used above because it is much faster to implement. Table 15.2 lists the modulation type and coding rate used for each data rate, and the total number of bits per OFDM symbol, which includes data bits and code bits.

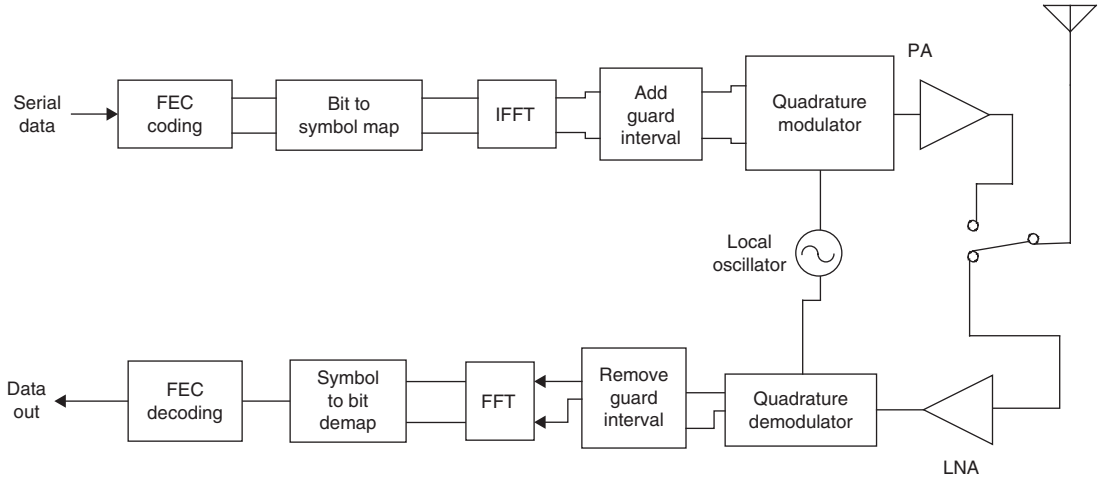


Figure 15.15: OFDM System Block Diagram

Table 15.2: OFDM Characteristics According to Data Rate

Data Rate Mbps	Modulation	Coding Rate	Coded Bits per Subcarrier	Coded Bits per OFDM Symbol	Data Bits per OFDM Symbol
6	BPSK	1/2	1	48	24
9	BPSK	3/4	1	48	36
12	QPSK	1/2	2	96	48
18	QPSK	3/4	2	96	72
24	16-QAM	1/2	4	192	96
36	16-QAM	3/4	4	192	144
48	64-QAM	2/3	6	288	192
54	64-QAM	3/4	6	288	216

The available frequency channels in the 5 GHz band in accordance with FCC paragraphs 15.401–15.407 for unlicensed national information infrastructure (U-NII) devices are shown in Table 15.3. Channel allocations are 5 MHz apart and 20 MHz spacing is needed to prevent co-channel interference. Twelve simultaneous networks can coexist without mutual interference. Power limits are also shown in Table 15.4.

Extension of the data rates of 802.11b to those of 802.11a, but on the 2.4 GHz band is covered in supplement 802.11g. The OFDM physical layer defined for the 5 GHz band is applied essentially unchanged to 2.4 GHz. Equipment complying with 802.11g must also have the lower-rate features and the CCK modulation technique of 802.11b so that it will be downward compatible with existing Wi-Fi systems.

Table 15.3: Channel Allocations and Maximum Power for 802.11a in United States

Band	Operation Channel Numbers	Channel Center Frequencies (MHz)	Maximum Power with up to 6 dBi antenna gain (mW)
U-NII lower band (5.15–5.25 GHz)	36 40 44 48	5180 5200 5220 5240	40
U-NII middle band (5.25–5.35 GHz)	52 56 60 64	5260 5280 5300 5320	200
U-NII upper band (5.725–5.825 GHz)	149 153 157 161	5745 5765 5785 5805	800

Table 15.4: HIPERLAN/2 Frequency Channels and Power Levels (Reference 15.1, ETSI TS 101 475 V1.3.1 (2001–12))

Center Frequency (MHz)	Radiated Power (mean EIRP) (dBm)
Every 20 MHz from 5180 to 5320	23
Every 20 MHz from 5500 to 5680	30
5700	23

15.1.5 HIPERLAN/2

While 802.11b was designed for compliance with regulations in the European Union and most other regions of the world, 802.11a specifically refers to the regulations of the FCC and the Japanese MPT. ETSI (European Telecommunications Standards Institute) developed a high-speed wireless LAN specification, called HIPERLAN/2 (high performance local area network), which meets the European regulations and in many ways goes beyond the capabilities of 802.11a. HIPERLAN/2 defines a physical layer essentially identical to that of 802.11a, using coded OFDM and the same data rates up to 54 Mbps. However, its second layer software level is very different from the 802.11 MAC and the two systems are not compatible. Built-in features of HIPERLAN/2 that distinguish it from IEEE 802.11a are the following:

- *Quality of service (QOS).* Time division multiple access/time division duplex (TDMA/TDD) protocol permits multimedia communication.
- *Dynamic frequency selection (DFS).* Network channels are selected and changed automatically to maintain communication reliability in the presence of interference and path disturbances.

- *Transmit power control (TPC)*. Transmission power is automatically regulated to reduce interference to other frequency band users and reduce average power supply consumption.
- *High data security*. Strong authentication and encryption procedures.

All of the above features of HIPERLAN/2 are being dealt with by IEEE task groups for implementation in 802.11. Specifically, the features of DFS and TPC are necessary for conformance of 802.11a to European Union regulations.

Frequency channels and power levels of HIPERLAN/2 are shown in Table 15.4.

15.2 Bluetooth

There are two sources of the Bluetooth specification. One is the Bluetooth Special Interest Group (SIG). The current version at this writing is Version 1.1. It is arranged in two volumes—Core and Profiles. Volume 1, the core, describes the physical, or hardware radio characteristics of Bluetooth, as well as low-level software or firmware which serves as an interface between the radio and higher level specific user software. The profiles in Volume 2 detail protocols and procedures for several widely used applications. The other Bluetooth source specification is IEEE 802.15.1. It is basically a rewriting of the SIG core specification, made to fit the format of IEEE communications specifications in general.

Bluetooth is an example of a wireless personal area network (WPAN), as opposed to a wireless local area network (WLAN). It's based on the creation of ad hoc, or temporary, on-the-fly connections between digital devices associated with an individual person and located in the vicinity of around ten meters from him. Bluetooth devices in a network have the function of a master or a slave, and all communication is between a master and one or more slaves, never directly between slaves. The basic Bluetooth network is called a piconet. It has one master and from one to seven slaves. A scatternet is an interrelated network of piconets where any member of a piconet may also belong to an adjacent piconet. Thus, conceptually, a Bluetooth network is infinitely expandable. Figure 15.16 shows a scatternet made up of three piconets. In it, a slave in one piconet is a master in another. A device may be a master in one piconet only.

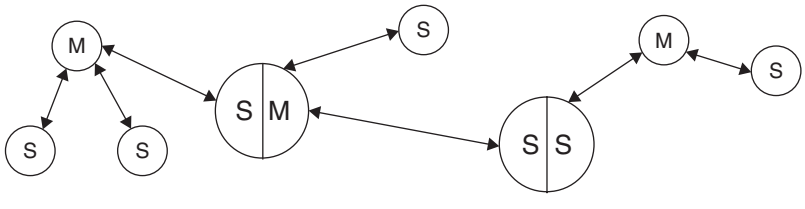


Figure 15.16: Bluetooth Scatternet

The basic RF communication characteristics of Bluetooth are shown in Table 15.5.

Table 15.5: Bluetooth Technical Parameters

Characteristic	Value	Comment
Frequency Band	2.4 to 2.483 GHz	May differ in some countries
Frequency Hopping Spread Spectrum (FHSS)	79 1-MHz channels from 2402 to 2480 MHz	May differ in some countries
Hop Rate	1600 hops per second	
Channel Bandwidth	1 MHz	20 dB down at edges
Modulation	Gaussian Frequency Shift Keying (GFSK)	
	Filter BT = 0.5	Gaussian Filter bandwidth = 500 kHz
	Nominal modulation index = 0.32	Nominal Deviation = 160 kHz
Symbol Rate	1 Mbps	
Transmitter Maximum Power		
Class 1	100 mW	Power control required
Class 2	2.5 mW	Must be at least 0.25 mW
Class 3	1 mW	No minimum specified
Receiver Sensitivity	−70 dBm for BER = 0.1%	

A block diagram of a Bluetooth transceiver is shown in Figure 15.17. It's divided into three basic parts: RF, baseband, and application software. A Bluetooth chip set will usually include the RF and baseband parts, with the application software being contained in the system's computer or controller. The user data stream originates and terminates in the application software. The baseband section manipulates the data and forms frames or data bursts for transmission. It also controls the frequency synthesizer according to the Bluetooth frequency-hopping protocol. The blocks in Figure 15.17 are general and various transmitter and receiver configurations are adopted by different manufacturers. The Gaussian low-pass filter block before the modulator, for example, may be implemented digitally as part of a complex signal *I*/*Q* modulation unit or it may be a discrete element filter whose output is applied to the frequency control line of a VCO. Similarly, the receiver may be one of several types, as discussed in Chapter 6. If a superheterodyne configuration is chosen, the filter at the output of the downconverter will be a bandpass type. A direct conversion receiver will use low pass filters in complex *I* and *Q* outputs of the downconverter. While different manufacturers employ a variety of methods to implement the Bluetooth radio, all must comply with the same strictly defined Bluetooth specification, and therefore the actual configuration used in a particular chipset should be of little concern to the end user.

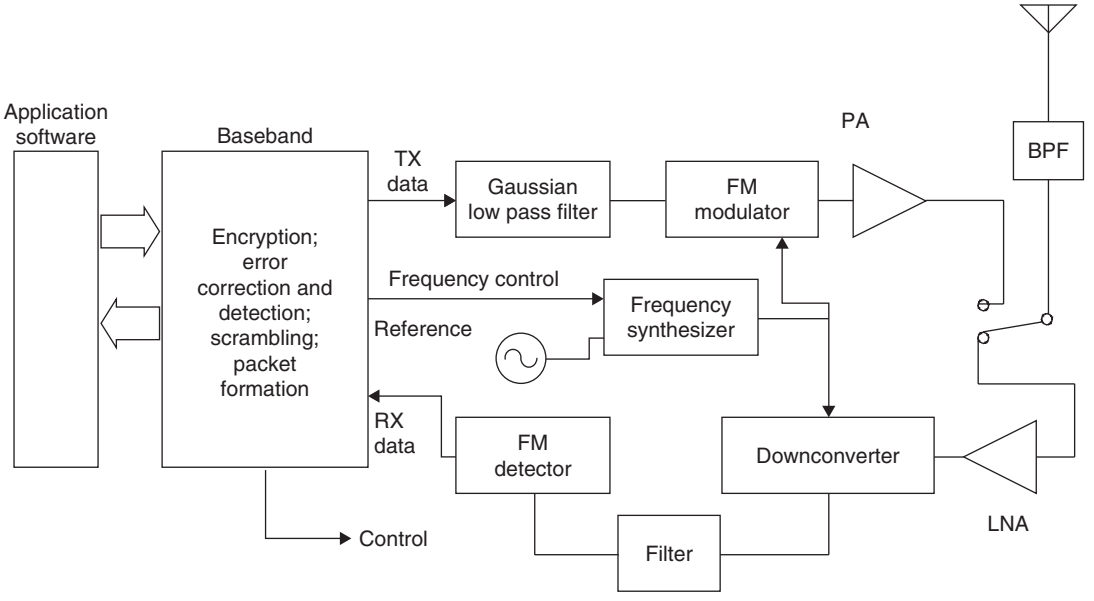


Figure 15.17: Bluetooth Transceiver

The Bluetooth protocol has a fixed-time slot of 625 microseconds, which is the inverse of the hop rate given in Table 15.5. A transmission burst may occur within a duration of one, three, or five consecutive slots on one hop channel. As mentioned, transmissions are always between the piconet master and a slave, or several slaves in the case of a broadcast, or point-to-multipoint transmission. All slaves in the piconet have an internal timer synchronized to the master device timer, and the state of this timer determines the transmission hop frequency of the master and that of the response of a designated slave. Figure 15.18 shows a sequence of transmissions between a master and two slaves. Slots are numbered according to the state, or phase, of the master clock, which is copied to each slave when it joins the piconet. Note that master transmissions take place during even numbered clock phases and slave transmissions during odd numbered phases. Transmission frequency depends on the clock phase, and if a device makes a three or five slot transmission (slave two in the diagram), the intermediate

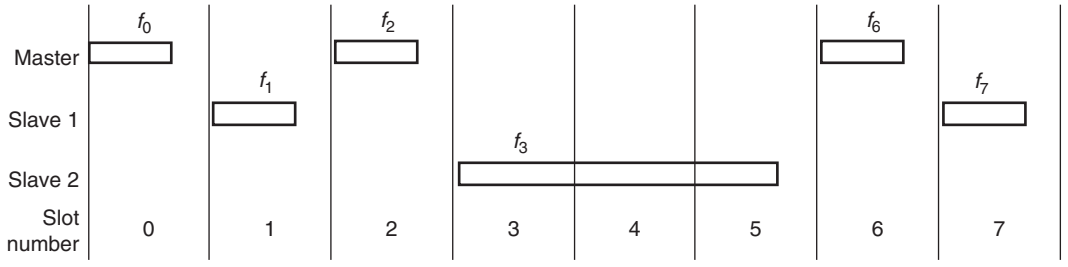


Figure 15.18: Bluetooth Timing

frequencies that would have been used if only single slots were transmitted are omitted (f_4 and f_5 in this case). Note that transmissions do not take up a whole slot. Typically, a single-slot transmission burst lasts 366 microseconds, leaving 259 microseconds for changing the frequency of the synthesizer, phase locked loop settling time, and for switching the transceiver between transmit and receive modes.

There are two different types of wireless links associated with a Bluetooth connection. An asynchronous connectionless link (ACL) is used for packet data transfer while a synchronous connection oriented link (SCO) is primarily for voice. There are two major differences between the two link types. When an SCO link is established between a master and a slave, transmissions take place on dedicated slots with a constant interval between them. Also, unlike an ACL link, transmitted frames are not repeated in the case of an error in reception. Both of these conditions are necessary because voice is a continuous real-time process whose data rate cannot be randomly varied without affecting intelligibility. On the other hand, packet data transmission can use a handshaking protocol to regulate data accumulation and the instantaneous rate is not usually critical. Thus, for ACL links the master has considerable leeway in proportioning data transfer with the slaves in its network. An ARQ (automatic repeat request) protocol is always used, in addition to optional error correction, to ensure the highest reliability of the data transfer.

Bluetooth was conceived for employment in mobile and portable devices, which are more likely than not to be powered by batteries, so power consumption is an important issue. In addition to achieving low-power consumption due to relatively low transmitting power levels, Bluetooth incorporates power saving features in its communication protocol. Low average power is achieved by reducing the transmission duty cycle, and putting the device in a low-power standby mode for as long a period as possible relative to transmit and receive times while still maintaining the minimum data flow requirements. Bluetooth has three modes for achieving different degrees of power consumption during operation: sniff, hold, and park. Even in the normal active mode, some power saving can be achieved, as described below.

Active Mode: During normal operation, a slave can transmit in a particular time slot only if it is specifically addressed by the master in the proceeding slot. As soon as it sees that its address is not contained in the header of the master's message, it can "go to sleep," or enter a low-power state until it's time for the next master transmission. The master also indicates the length of its transmission (one, three, or five slots) in its message header, so the slave can extend its sleep time during a multiple slot interval.

Sniff Mode: In this mode, sleep time is increased because the slave knows in advance the time interval between slots during which the master may address the slave. If it's not addressed during the agreed slot, it returns to its low-power state for the same period and then wakes up and listens again. When it is addressed, the slave continues listening during subsequent master transmission slots as long as it is addressed, or for an agreed time-out period.

Hold Mode: The master can put a slave in the hold mode when data transfer between them is being suspended for a given period of time. The slave is then free to enter a low-power state, or do something else, like participate in another piconet. It still maintains its membership in the original piconet, however. At the end of the agreed time interval, the slave resynchronizes with the traffic on the piconet and waits for instructions from the master.

Park Mode: Park has the greatest potential for power conservation, but as opposed to hold and sniff, it is not a directly addressable member of the piconet. While it is outside of direct calling, a slave in park mode can continue to be synchronized with the piconet and can rejoin it later, either on its own initiative or that of the master, in a manner that is faster than if it had to join the piconet from scratch. In addition to saving power, park mode can also be considered a way to virtually increase the network's capacity from eight devices to 255, or even more. When entering park mode, a slave gives up its active piconet address and receives an 8-bit parked member address. It goes into low-power mode but wakes up from time to time to listen to the traffic and maintain synchronization. The master sends beacon transmissions periodically to keep the network active. Broadcast transmissions to all parked devices can be used to invite any of them to rejoin the network. Parked units themselves can request re-association with the active network by way of messages sent during an access window that occurs a set time after what is called a "beacon instant." A polling technique is used to prevent collisions.

15.2.1 Packet Format

In addition to the data that originates in the high-level application software, Bluetooth packets contain fields of bits that are created in the baseband hardware or firmware for the purpose of acquisition, addressing, and flow control. Packet bits are also subjected to data whitening (randomization), error-correction coding, and encryption as defined for each particular data type. Figure 15.19 shows the standard packet format.

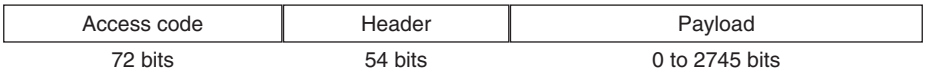


Figure 15.19: Bluetooth Packet

The access code is used for synchronization, d-c level compensation, and identification. Each Bluetooth device has a unique address, and it is the address of the device acting as master that is used to identify transmitted packets as belonging to a specific piconet. A 64-bit synchronization word sandwiched between a four-bit header and four-bit trailer, which provide d-c compensation, is based on the master's address. This word has excellent correlation properties so when it is received by any of the piconet members it provides synchronization and positive identification that the packet of which it is a part belongs to their network. All message packets sent by members of the piconet use the same access code.

The header contains six fields with link control information. First, it has a three-bit active member address which identifies to which of the up to seven slaves a master's message is destined. An all zero address signifies a broadcast message to all slaves in the piconet. The next field has four bits that define the type of packet being sent. It specifies, for example, whether one, three, or five slots are occupied, and the level of error correction applied. The remaining fields involve flow control (handshaking), error detection and sequencing. Since the header has prime importance in the packet, it is endowed with forward-error correction having a redundancy of times three.

Following the header in the packet is the payload, which contains the actual application or control data being transferred between Bluetooth devices. The contents of the payload field depend on whether the link is an ACL or SCO. The payload of ACL links has a payload header field that specifies the number of data bytes and also has a handshaking bit for data-buffering control. A CRC (cyclic redundancy check) field is included for data integrity. As stated above, SCO links don't retransmit packets so they don't include a CRC. They don't need a header either because the SCO payload has a constant length.

The previous packet description covers packets used to transfer user data, but other types of packets exist. For example, the minimum length packet contains only the access code, without the four-bit trailer, for a total of 68 bits. It's used in the inquiry and paging procedures for initial frequency-hopping synchronization. There are also NULL and POLL packets that have an access code and header, but no payload. They're sent when slaves are being polled to maintain synchronization or confirm packet reception (in the case of NULL) in the piconet but there is no data to be transferred.

15.2.2 Error Correction and Encryption

The use of forward error correction (FEC) improves throughput on noisy channels because it reduces the number of bad packets that have to be retransmitted. In the case of SCO links that don't use retransmission, FEC can improve voice quality. However, error correction involves bit redundancy so using it on relatively noiseless links will decrease throughput. Therefore, the application decides whether to use FEC or not.

As already mentioned, there are various types of packets, and the packet type defines whether or not FEC is used. The most redundant FEC method is always used in the packet header, and for the payload in one type of SCO packet. It simply repeats each bit three times, allowing the receiver to decide on the basis of majority rule what data bit to assign to each group of incoming bits.

The other FEC method, applied in certain type ACL and SCO packets, uses what's called a (15, 10) shortened Hamming code. For every ten data bits, five parity bits are generated. Since out of every 15 transmitted bits only ten are retrieved, the data rate is only two-thirds what it

would be without coding. This code can correct all single errors and detect all double errors in each 15-bit code word.

Wireless communication is susceptible to eavesdropping so Bluetooth incorporates optional security measures for authentication and encryption. Authentication is a procedure for verifying that received messages are actually from the party we expect them to be and not from an outsider who is inserting false messages. Encryption prevents an eavesdropper from understanding intercepted communications, since only the intended recipient can decipher them. Both authentication and implementation routines are implemented in the same way. They involve the creation of secret keys that are generated from the unique Bluetooth device address, a PIN (personal identification number) code, and a random number derived from a random or pseudo-random process in the Bluetooth unit. Random numbers and keys are changed frequently. The length of a key is a measure of the difficulty of cracking a code. Authentication in Bluetooth uses a 128-bit key, but the key size for encryption is variable and may range from 8 to 128 bits.

15.2.3 Inquiry and Paging

A distinguishing feature of Bluetooth is its *ad hoc* protocol and connections are often required between devices that have no previous knowledge of their nature or address. Also, Bluetooth networks are highly volatile, in comparison to WLAN for example, and connections are made and dissolved with relative frequency. To make a new connection, the initiator—the master—must know the address of the new slave, and the slave has to synchronize its clock to the master's in order to align transmit and receive channel hop-timing and frequencies. The inquiry and paging procedures are used to create the connections between devices in the piconet.

By use of the inquiry procedure, a connection initiator creates a list of Bluetooth devices within range. Later, desired units can be summoned into the piconet of which the initiator is master by means of the paging routine.

As mentioned previously, the access code contains a synchronization word based on the address of the master. During inquiry, the access code is a general inquiry access code (GIAC) formed from a reserved address for this purpose. Dedicated inquiry access codes (DIAC) can also be used when the initiator is looking only for certain types of devices. Now a potential slave can lock on to the master, provided it is receiving during the master's transmission time and on the transmission frequency. To facilitate this match-up, the inquiry procedure uses a special frequency hop routine and timing. Only 32 frequency channels are used and the initiator transmits two burst hops per standard time slot instead of one. On the slot following the transmission inquiry bursts, the initiator listens for a response from a potential slave on two consecutive receive channels whose frequencies are dependent on the previously transmitted frequencies.

When a device is making itself available for an inquiring master, it remains tuned to a single frequency for a period of 1.28 seconds and at a defined interval and duration scans the channel for a transmission. At the end of the 1.28-second period, it changes to another channel frequency. Since the master is sending bursts over the whole inquiry frequency range at a fast rate—two bursts per 1250 microsecond interval—there's a high probability the scanning device will catch at least one of the transmissions while it remains on a single frequency. If that channel happens to be blocked by interference, then the slave will receive a transmission after one of its subsequent frequency changes. When the slave does hear a signal, it responds during the next slot with a special packet called FHS (frequency hop synchronization) in which is contained the slave's Bluetooth address and state of its internal clock register. The master does not respond but notes the slave's particulars and continues inquiries until it has listed the available devices in its range. The protocol has provisions for avoiding collisions from more than one scanning device that may have detected a master on the same frequency and at the same time.

The master makes the actual connection with a new device appearing in its inquiry list using the page routine. The paging procedure is quite similar to that of the inquiry. However, now the master knows the paged device's address and can use it to form the synchronization word in its access code. The designated slave does its page scan while expecting the access code derived from its own address. The hopping sequence is different during paging than during inquiry, but the master's transmission bursts and the slave's scanning routine are very similar.

A diagram of the page state transmissions is given in Figure 15.20. When the slave detects a transmission from the master (Step 1), it responds with a burst of access code based on its own Bluetooth address. The master then transmits the FHS, giving the slave the access code information (based on the master's address), timing and piconet active member address (between one and seven) needed to participate in the network. The slave acknowledges FHS receipt in Step 4. Steps 5 and 6 show the beginning of the network transmissions which use the normal 79 channel hopping-sequence based on the master's address and timing.

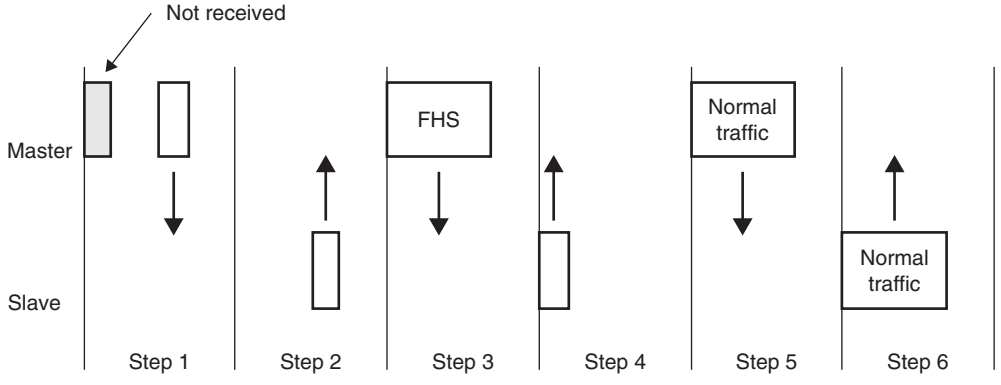


Figure 15.20: Paging Transmissions

15.3 Zigbee

Zigbee is the name of a standards-based wireless network technology that addresses remote monitoring and control applications. Its promotion and development is being handled on two levels. A technical specification for the physical and data link layers, IEEE 802.15.4, was drawn up by a working group of the IEEE as a low data rate WPAN (wireless personal area network). An association of committed companies, the Zigbee Alliance, is defining the network, security, and application layers above the 802.15.4 physical and medium access control layers, and will deal with interoperability certification and testing.

The distinguishing features of Zigbee to which the IEEE standard addresses itself are

- Low data rates—throughput between 10 and 115.2 Kbps
- Low power consumption—several months up to two years on standard primary batteries
- Network topology appropriate for multisensor monitoring and control applications
- Low complexity for low cost and ease of use
- Very high reliability and security

These will lend themselves to wide-scale use embedded in consumer electronics, home and building automation and security systems, industrial controls, PC peripherals, medical and industrial sensor applications, toys and games and similar applications. It's natural to compare Zigbee with the other WPAN standard, Bluetooth, and there will be some overlap in implementations. However, the two systems are quite different, as is evident from the comparison in Table 15.6.

15.3.1 Architecture

The basic architecture of Zigbee is similar to that of other IEEE standards, Wi-Fi and Bluetooth for example, a simplified representation of which is shown in Figure 15.21. On the bottom are the physical layers, showing two alternative options for the RF transceiver functions of the specification. Both of these options are never expected to exist in a single device, and indeed their transmission characteristics—frequencies, data rates, modulation system—are quite different. However, the embedded firmware and software layers above them will be essentially the same no matter what physical layer is applied. Just above the physical layers is the data link layer, consisting of two sublayers: medium access control, or MAC, and the logical link control, LLC. The MAC is responsible for management of the physical layer and among its functions are channel access, keeping track of slot times, and message delivery acknowledgement. The LLC is the interface between the MAC and physical layer and the upper-application software.

Table 15.6: Comparison of Zigbee and Bluetooth

	Bluetooth	Zigbee
Transmission Scheme	FHSS (Frequency Hopping Spread Spectrum)	DSSS (Direct Sequence Spread Spectrum)
Modulation	GFSK (Gaussian Frequency Shift Keying)	QPSK (Quadrature Phase Shift Keying) or BPSK (Binary Phase Shift Keying)
Frequency Band	2.4 GHz	2.4 GHz, 915 MHz, 868 MHz
Raw Data Bit Rate	1 MBPS	250 KBPS, 40 KBPS or 20 KBPS (depends on frequency band)
Power Output	Maximum 100 mW, 2.5 mW, or 1 mW, depending on class	Minimum capability 0.5 mW; maximum as allowed by local regulations
Minimum Sensitivity	−70 dBm for 0.1% BER	−85 dBm (2.4 GHz)or −92 dBm (915/868 MHz) for packet error rate < 1%
Network topology	Master-Slave 8 active nodes	Star or Peer-Peer 255 active nodes

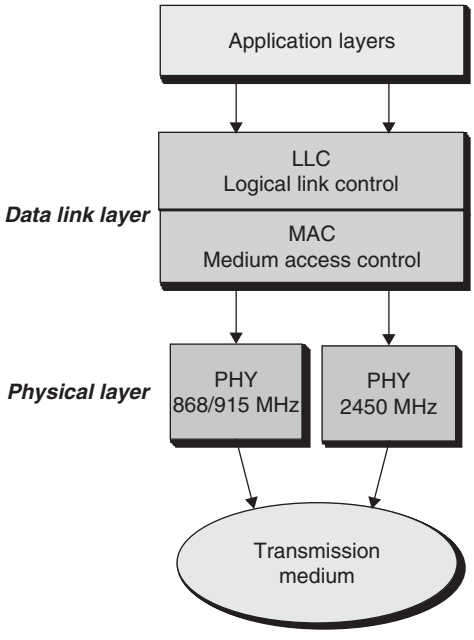


Figure 15.21: Zigbee Architecture

Application software is not a part of the IEEE 802.15.4 specification and it is expected that the Zigbee Alliance will prepare profiles, or programming guidelines and requirements for various functional classes in order to assure product interoperability and vendor independence. These profiles will define network formation, security, and application requirements while keeping in mind the basic Zigbee features of low power and high reliability.

15.3.2 Communication Characteristics

In order to achieve high flexibility of adaptation to the range of applications envisioned for Zigbee, operation is being specified for three unlicensed bands—2.4 GHz, 915 MHz and 868 MHz, the latter two being included in the same physical layer. Those two bands are generally mutually exclusive, their use being determined by geographic location and regional regulations. The following 27 transmitting channels are defined:

Channel Number	Center Frequency Range	Channel Width
0	868.3 MHz	600 kHz
1 to 10	906 to 924 MHz	2 MHz
11 to 27	2405 to 2480	5 MHz

Data rates and modulation types for each of the bands are shown in Table 15.7.

Table 15.7: Data Rates and Modulation

PHY (MHz) (MHz)	Frequency Band (MHz)	Spreading Parameters		Data Parameters		
		Chip Rate (kcps)	Modulation	Bit Rate (kbps)	Symbol Rate (ksps)	Symbols
868/915	868–868.6	300	BPSK	20	20	Binary
	902–928	600	BPSK	40	40	Binary
2450	2400–2483.5	2000	Offset-QPSK	250	62.5	16-ary Orthogonal

In both physical layers, the modulation is DSSS (direct sequence spread spectrum). The spreading parameters are defined to meet communication authority regulations in the various regions as well as desired data rates. For example, the chip rate of 600 Kbps on the 902–928 band allows the transmission to meet the FCC paragraph 15.247 requirement of minimum 500 kHz bandwidth at 6 dB down for digital modulation. However the chip rate, and with it the data rate, has to be reduced on Channel 0 in order to meet the confines of the 868 to 868.6 MHz channel allowed under ERC recommendation 70–03 and ETSI specification EN 300–220. On the 2400 to 2483.5 MHz band, the bit rate of 250 Kbps allows a throughput, after

considering the overheads involved in packet transmissions, to attain 115.2 Kbps, a rate used for some PC peripherals for example.

The spreading modulation used on the 2450 MHz physical layer has similarity in principle to that used on IEEE 802.11b (high-rate Wi-Fi) to increase the data bit rate without raising the chip rate, thereby achieving a desired carrier bandwidth. Sixteen different, almost orthogonal 32-bit long spreading sequences are available for transmission at 2 Mchips/second. Each consecutive sequence of four data bits determines which of the sixteen spreading sequences is sent. On reception, the receiver can identify the spreading sequence and thus decode the data bits. The modulation used, O-QPSK (offset quadrature phase shift keying) with half-sine wave pulse shaping is essentially equivalent to a form of frequency shift keying, MSK (minimum shift keying). It is fairly easy to generate and has a relatively narrow bandwidth for the given chip rate. This latter feature allows a large number of nonoverlapping channels that can be used, with proper upper layer software, on the crowded 2.4 GHz band to avoid interference.

Other physical layer characteristics of Zigbee are output power and receiver sensitivity. The devices must be capable of radiating at least −3 dBm although output may be reduced to the minimum necessary in order to limit interference to other users. Maximum power is determined by the regulatory authorities. While much higher powers are allowed, it may not be practical to transmit over, say 10 dBm, because of absolute limits on spurious radiation and the general objective of low-cost and low-power consumption. Minimum receiver sensitivity for the 868/915 MHz physical layer is specified as −92 dBm and −85 dBm on 2.4 GHz. These limits are for a packet error rate of one percent.

15.3.3 Device Types and Topologies

Two device types, of different complexities, are defined. A full function device (FFD) will be able to implement the full protocol set and can act as a network coordinator. Devices capable of minimal protocol implementation are reduced function devices (RFD). Due to the distinction between device types, networks in which most members require only minimum functionality, such as switches and sensors, can be made significantly less costly and have lower power consumption than if all devices were constrained to have maximum capability.

Flexibility in network configuration is achieved through two topologies—star and peer-to-peer that are depicted in Figure 15.22. A network may have as many as 255 members, one of which is a PAN (personal area network) coordinator. The function of the PAN coordinator, in addition to any specific application it may have, is to initiate, terminate, or route communication around the network. It also provides synchronization services. In a star network, each device communicates directly with the coordinator. The coordinator must be a FFD, and the others can be FFDs or RFDs. Relatively simple applications, like PC peripherals and toys, would typically use the star topology.

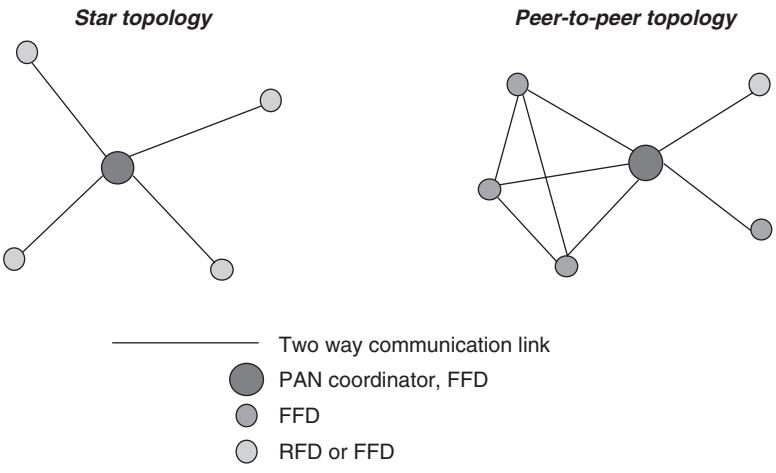


Figure 15.22: Network Topologies

In the peer-to-peer topology, any device can communicate with any other device as long as it is in range. RFDs cannot participate, since an RFD can only communicate with a FFD. More complicated structures can be set up as a combination of peer-to-peer groups and star configurations. There is still just one PAN coordinator in the whole network. One example of such a structure is a cluster-tree network shown in Figure 15.23. In this arrangement devices on the network extremities may well be out of radio range of each other, but they can still communicate by relaying messages through the individual clusters.

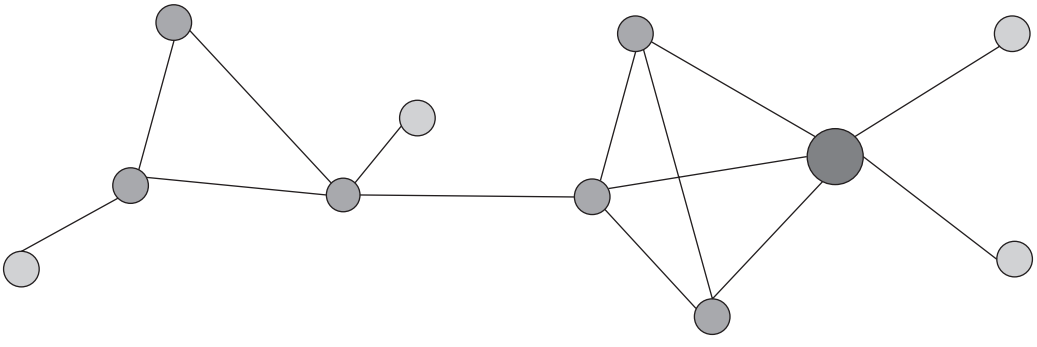


Figure 15.23: Cluster-Tree Topology

15.3.4 Frame Structure, Collision Avoidance, and Reliability

Zigbee frame construction and channel access are similar to those of WLAN 802.11 (Wi-Fi) but are less complex. The transmitted packet has the basic construction shown in Figure 15.24. The purpose of the preamble is to permit acquisition of chip and symbol timing. The PHY

header, which is signaled by a delimiter byte, notifies the baseband software in the receiver of the length of the subsequent data. The PSDU (PHY service data unit) is the message that has been passed down through the higher protocol layers. As shown, it can have a maximum of 127 bytes although monitoring and control applications will typically be much shorter. Included in the PSDU are information on the format of the message frame, a sequence number, address information, the data payload itself, and at the end, two bytes that serve as a frame check sequence. Reliability is assured since the receiver performs an independent calculation of this frame check sequence and compares it with the number received. If any bits have been changed by interference or noise, the numbers will not match. Only if a match occurs, the receiving side returns an acknowledgement to the originator of the message. Lacking an acknowledgement, the transmission will be repeated until it is successfully received.

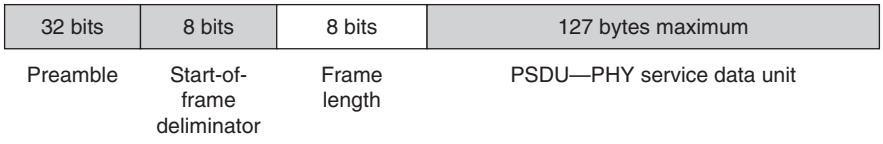


Figure 15.24: Transmission Packet

In order to avoid two or more stations trying to transmit at the same time, a carrier sense multiple access with collision avoidance (CSMA-CA) routine is employed, similar to that used in Wi-Fi, IEEE 802.11. The Zigbee receiver monitors the channel and only if it is idle it may initiate a transmission. If the channel is occupied, the terminal must wait a random back off period before it can again attempt access.

Acknowledgement messages are sent without using the collision avoidance mechanism.

15.3.5 Zigbee Applications

While the promoters of Zigbee aim to cover a very large market for those applications that require relatively low data rates, there will remain applications for which the compromises inherent in a general specification are not acceptable, and producers will continue to develop devices with proprietary specifications and characteristics. However, the open specification and a recognized certification of conformity are an advantage in many situations. For example, a home burglar alarm system would accept wireless sensors produced by different manufacturers, which will facilitate future expansion or allow installers to add sensors of types not available from the original system manufacturer. Use of devices approved according to a recognized standard gives the consumer some security against obsolescence.

Although Zigbee claims to be appropriate for most control applications, it will not fit all of them, and will not necessarily take advantage of all the possibilities of the unlicensed device regulations. Its declared maximum range of some 50 to 75 meters will fall short of the

requirements of many systems. Given the meager maximum power allowed, greater range means reduced bandwidth and reduced data rate. In fact, a great many of the applications envisaged by Zigbee can get by very well with data rates of hundreds or a few thousand bits per second, and by matching receiver sensitivity to these rates, ranges of hundreds of meters can be achieved.

One partial answer to the range question is the deployment of the Zigbee network in a cluster-tree configuration, as previously described. Adjacent nodes serve as repeaters so that large areas can be covered, as long as the greatest distance between any two directly communicating nodes does not exceed Zigbee's basic range capability. For example, in a multi-floor building, sensors on the top floor can send alarms to the control box in the basement by passing messages through sensors located on every floor and operating as relay stations.

No doubt that there will be competition between Bluetooth and Zigbee for use in certain applications, but the overall deployment and the reliability of wireless control systems will increase. The proportion of wireless security and automation systems will increase because the new standard will provide a significant boost in reliability, security, and convenience, as compared to most present solutions.

15.4 Conflict and Compatibility

With the steep rise of Bluetooth product sales and the already large and growing use of wireless local area networks, there is considerable concern about mutual interference between Bluetooth-enabled and Wi-Fi devices. Both occupy the 2.4 to 2.4835 GHz unlicensed band and use wideband spread-spectrum modulating techniques. They will most likely be operating concurrently in the same environments, particularly office/commercial but also in the home.

Interference can occur when a terminal of one network transmits on or near the receiving frequency of a terminal in another collocated network with enough power to cause an error in the data of the desired received signal. Although they operate on the same frequency band, the nature of Bluetooth and Wi-Fi signals are very different. Bluetooth has a narrowband transmission of 1 MHz bandwidth which hops around pseudo-randomly over an 80 MHz band while Wi-Fi (using DSSS) has a broad, approximately 20 MHz, bandwidth that is constant in some region of the band. The interference phenomenon is apparent in Figure 15.25. Whenever there is a frequency and time coincidence of the transmission of one system and reception of the other, it's possible for an error to occur. Whether it does or not depends on the relative signal strengths of the desired and undesired signals. These in turn depend on the radiated power outputs of the transmitters and the distance between them and the receiver. When two terminals are very close (on the order of centimeters), interference may occur even when the transmitting frequency is outside the bandwidth of the affected receiver.

Bluetooth and Wi-Fi systems are not synchronous and interference between them has to be quantified statistically. We talk about the probability of a packet error of one system caused

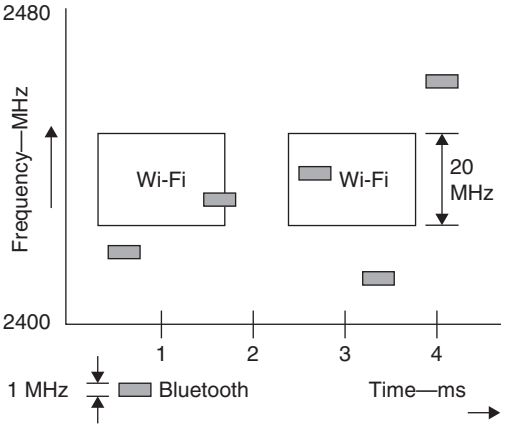


Figure 15.25: Wi-Fi and Bluetooth Spectrum Occupation

by the other system. The consequence of a packet error is that the packet will have to be retransmitted once or more until it is correctly received, which causes a delay in message throughput. Voice transmissions generally don't allow packet retransmission because throughput cannot be delayed, so interference results in a decrease in message quality.

Following are parameters that affect interference between Bluetooth and Wi-Fi:

- *Frequency and time overlap*. A collision occurs when the interferer transmits at the same time as the desired transmitter and is strong enough to cause a bit or symbol error in the received packet.
- *Packet length*. The longer the packet length of the Wi-Fi system, relative to a constant packet length and hop rate of Bluetooth, the longer the victim may be exposed to interference from one or more collisions and the greater the probability of a packet error.
- *Bit rate*. Generally, the higher the bit rate, the lower the receiver sensitivity and therefore the more susceptible the victim will be to packet error for given desired and interfering signal strengths. On the other hand, higher bit rates usually result in reduced packet length, with the opposite effect.
- *Use factor*. Obviously, the more often the interferer transmits, the higher the probability of packet error. When both communicating terminals of the interferer are in the interfering vicinity of the victim the use factor is higher than if the terminals are further apart and one of them does not have adequate strength to interfere with the victim.
- *Relative distances and powers*. The received power depends on the power of the transmitter and its distance. Generally, Wi-Fi systems use more power than Bluetooth, typically 20 mW compared to 1 mW. Bluetooth Class 1 systems may transmit up to

100 mW, but their output is controlled to have only enough power to give a required signal level at the receiving terminal.

- Signal-to-interference ratio of the victim receiver, SIR, for a specified symbol or frame error ratio.
- Type of modulation, and whether error-correction coding is used.

A general configuration for the location of Wi-Fi and interfering Bluetooth terminals is given in Figure 15.26. In this discussion, only transmissions from the access point to the mobile terminal are considered. We can get an idea of the vicinity around the Wi-Fi mobile terminal in which operating Bluetooth terminals will affect transmissions from the access point to the mobile terminal by examining the following parameters:

CI_{cc}, CI_{ac}—Ratio of signal carrier power to co-channel or adjacent channel interfering power for a given bit or packet error rate (probability).

P_{WF}, P_{BT}—Wi-Fi and Bluetooth radiated power outputs.

$PL = Kd^r$—Path loss which is a function of distance d between transmitting and receiving terminals, and the propagation exponent r. K is a constant.

d_1—Distance between Wi-Fi mobile terminal and access point.

d_2—Radius of area around mobile terminal within which an interfering Bluetooth transmitter signal will increase the Wi-Fi bit error rate above a certain threshold.

PR_{WF}, PR_{BT}—Received powers from the access point and from the Bluetooth interfering transmitters.

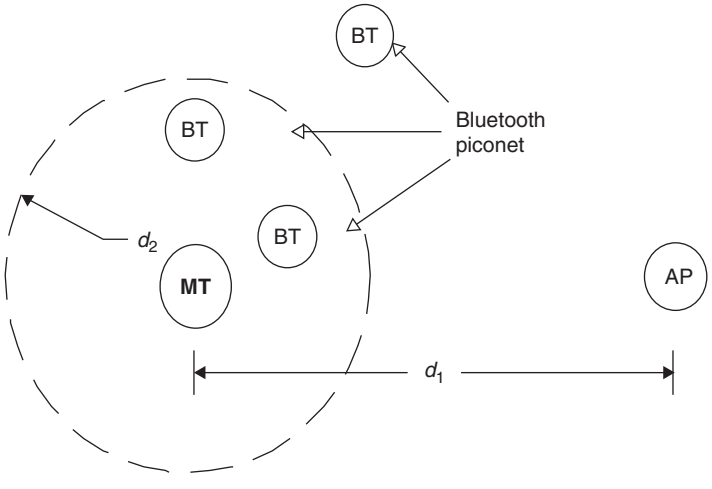


Figure 15.26: Importance of Relative Terminal Location

d_2 as a function of d_1 is found as follows, using power in dBm:

$$\begin{aligned} &1)\ PR_{\mathrm{WF}} = P_{\mathrm{WF}} - 10\log(\mathrm{K}d_1^r);\ PR_{\mathrm{BT}} = P_{\mathrm{BT}} - 10\log(\mathrm{K}d_2^r) \\ &2)\ CI_{\mathrm{CC}} = PR_{\mathrm{WF}} - PR_{\mathrm{BT}} = P_{\mathrm{WF}} - 10\log(\mathrm{K}d_1^r) - P_{\mathrm{BT}} + 10\log(Kd_2^r) \\ &3)\ (CI_{\mathrm{CC}} - P_{\mathrm{WF}} + P_{\mathrm{BT}})/10r = \log(d_2/d_1) \\ &4)\ d_2 = d_1 \cdot 10^{(CI_{\mathrm{CC}} - P_{\mathrm{WF}} + P_{\mathrm{BT}})/(10\cdot r)} \end{aligned} \tag{15.4}$$

As an example, the interfering area radius d_2 is now calculated from equation (15.3) using the following system parameters:

$$CI_{\mathrm{CC}} = 10\,\mathrm{dB}, \quad P_{\mathrm{WF}} = 13\,\mathrm{dBm}, \quad P_{\mathrm{BT}} = 0\,\mathrm{dBm}, \quad r = 2 \qquad \text{(free space)} \tag{15.5}$$

In this case, if a Wi-Fi terminal is located 15 meters from an access point, for example, all active Bluetooth devices within a distance of 10.6 meters from it have the potential of interfering. Only co-channel interference is considered. Adjacent channel interference, if significant, would increase packet error probability because many more Bluetooth hop channels would cause symbol errors. However, the adjacent channel CI_{ac} is on the order of 45 dB lower than CI_{cc} and would be noticed only when Bluetooth is several centimeters away from the Wi-Fi terminal.

The effect of an environment where path loss is greater than in free space can be seen by using an exponent $r = 3$. For the same Wi-Fi range of 15 meters, the radius of Bluetooth interference becomes 11.9 meters.

While equation (15.3) does give a useful insight into the range where Bluetooth devices are liable to deteriorate Wi-Fi performance, its development did involve simplifications. It considered that the signal-to-interference ratio that causes the error probability to exceed a threshold is constant for all wanted signal levels, which isn't necessarily so. It also implies a step relationship between signal-to-interference ratio and performance degradation, whereas the effect of changing interference level is continuous. The propagation law used in the development is also an approximation.

15.4.1 Methods for Improving Bluetooth and Wi-Fi Coexistence

By dynamically modifying one or more system operating parameters according to detected interference levels, coexistence between Bluetooth and Wi-Fi can be improved. Some of these methods are discussed below.

15.4.1.1 Power Control

Limiting transmitter power to the maximum required for a satisfactory level of performance will reduce interference to collocated networks. Power control is mandatory for Class 1 Bluetooth

systems, where maximum power is 100 mW. The effect of the power on the interference radius is evident in equation (15.1). For example, in a Bluetooth piconet established between devices located over a spread of distances from the master, the master will use only the power level needed to communicate with each of the slaves in the network. Lack of power control would mean that all devices would communicate at maximum power and the collocated Wi-Fi system would be exposed to a high rate of interfering Bluetooth packets.

15.4.1.2 Adaptive Frequency Hopping

Wi-Fi and Bluetooth share approximately 25 percent of the total Bluetooth hop-span of 80 MHz. Probably the most effective way to avoid interference between the two systems is to restrict Bluetooth hopping to the frequency range not used by Wi-Fi. When there is no coordination or cooperation between collocated networks, the Bluetooth piconet master would have to sense the presence of Wi-Fi transmissions and modify the frequency-hopping scheme of the network accordingly. A serious obstacle to this method was lifted by a change to the FCC regulations governing spread-spectrum transmissions in the 2.4 GHz band. Previously, frequency hopping devices were committed to hopping over at least 75 pseudo-randomly selected hop channels. In August 2002, paragraph 15.247, according to which Bluetooth and Wi-Fi devices are regulated, was changed to allow a minimum of 15 nonoverlapping channels in the 2400 to 2483 MHz band. In addition, the regulation allows employing intelligent hopping techniques, when less than 75 hopping frequencies are used, to avoid interference with other transmissions, and also suppression of transmission on an occupied channel provided that there are a minimum of 15 hops. The Bluetooth specification is due to be modified to take advantage of the adaptive frequency hopping method of avoiding interference.

There are situations where adaptive frequency hopping may not be effective or may have a negative effect. When two or more adjacent Wi-Fi networks are operating concurrently, they will utilize different 22 MHz sections of the 2.4 GHz band—three nonoverlapping Wi-Fi channels are possible. In this case, Bluetooth may not be able to avoid collisions while using a minimum of 15 hop frequencies. In addition, if there are several Bluetooth piconets in the same area, collisions among themselves will be much more frequent than when the full 79 channel hopping sequences are used.

15.4.1.3 Packet Fragmentation

The two interference-avoiding methods described above are applicable primarily for action by the Bluetooth network. One method that the Wi-Fi network can employ to improve throughput is packet fragmentation. By fragmenting data packets and sending more, but shorter transmission frames, each transmission will have a lower probability of collision with a Bluetooth packet. Although reducing frame size increases the percentage of overhead bits in the transmission, when interference is heavy the overall effect may be higher throughput than

if fragmentation was not used. Increasing bit rate for a constant packet length will also result in a shorter transmitted frame and less exposure to interference.

The methods mentioned above for reducing interference presume no coordination between the two different types of collocated wireless networks. However, devices are now being produced, in laptop and notebook computers for example, that include both Wi-Fi and Bluetooth, sometimes even in the same chipset. In this case collaboration is possible in the device software to prevent inter-network collisions.

15.5 Ultra-wideband Technology

Ultra-wideband (UWB) technology is based on transmission of very narrow electromagnetic pulses at a low repetition rate. The result is a radio spectrum that is spread over a very wide bandwidth—much wider than the bandwidth used in the spread-spectrum systems previously discussed. Ultra-wideband transmissions are virtually undetectable by ordinary radio receivers and therefore can exist concurrently with existing wireless communications without demanding additional spectrum or exclusive frequency bands.

These are some of the advantages cited for ultra-wideband technology:

- Very low spectral density—Very low probability of interference with other radio signals over its wide bandwidth,
- High immunity to interference from other radio systems,
- Low probability of interception/detection by other than the desired communication link terminals,
- High multipath immunity,
- Many high data rate ultra-wideband channels can operate concurrently,
- Fine range-resolution capability,
- Relatively simple, low-cost construction, based on nearly all digital architectures.

Transmission and reception methods are unique, and are described briefly below.

Differing from conventional radio communication systems, which use up conversion and down conversion to pass information signals between baseband and bandpass frequency channels where wireless propagation occurs, UWB signal generation and detection use baseband techniques. An example of a UWB "carrier" is a Gaussian monopulse, shown in Figure 15.27 [Ref. 15.2]. Its power spectrum is shown in Figure 15.28. If the time scale in Figure 15.27 is in nanoseconds, then the width of the pulse is 0.5 nanoseconds and the 3 dB bandwidth of the power spectrum is approximately 3.2 GHz with maximum power density at 2 GHz.

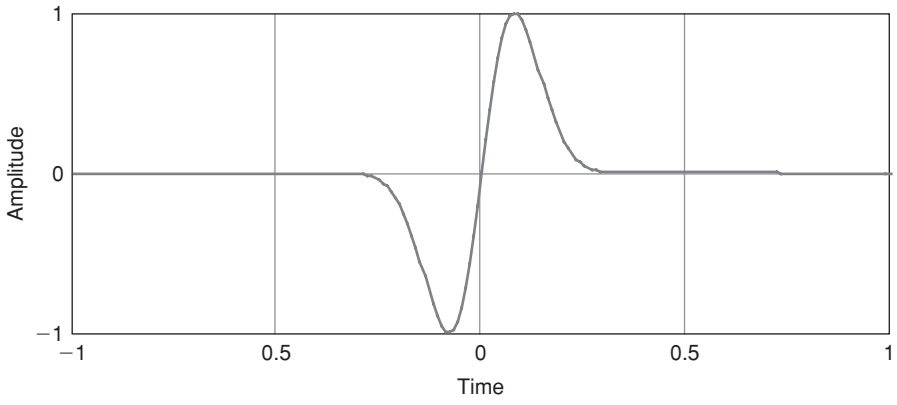


Figure 15.27: UWB Monopulse

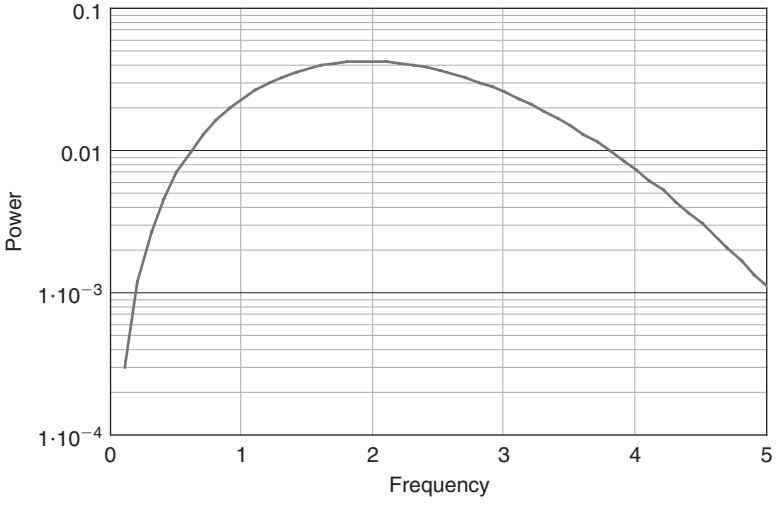


Figure 15.28: Spectrum of Monopulse

In order to pass information over a UWB communication link, trains of pulses must be transmitted with some characteristic of a pulse or group of pulses varied in order to distinguish between "0" and "1." The time between consecutive pulses should be determined in a pseudo-random manner in order to smooth the energy spikes in the frequency spectrum. Reception of the transmitted pulse train is done by correlating the received signal with a similar sequence of pulses generated in the receiver. A large number of communication links can be maintained simultaneously and independently by using different pseudo-random sequences for each link.

A pulse similar to that of Figure 15.27 can be generated by applying an impulse, or perhaps more conveniently a step-voltage or current, to a linear band limited network. Figure 15.29 is a simulation of a sequence of UWB pulses created by stimulating a bandpass filter with a pseudo-randomly spaced sequence of impulses. The figure also shows the power spectrum of

that sequence. The network that creates the individual UWB pulses includes the transmitter antenna, the propagation channel, and the receiving antenna, whose characteristics, in terms of impulse response or amplitude and phase vs. frequency must be known and accounted for in designing the system.

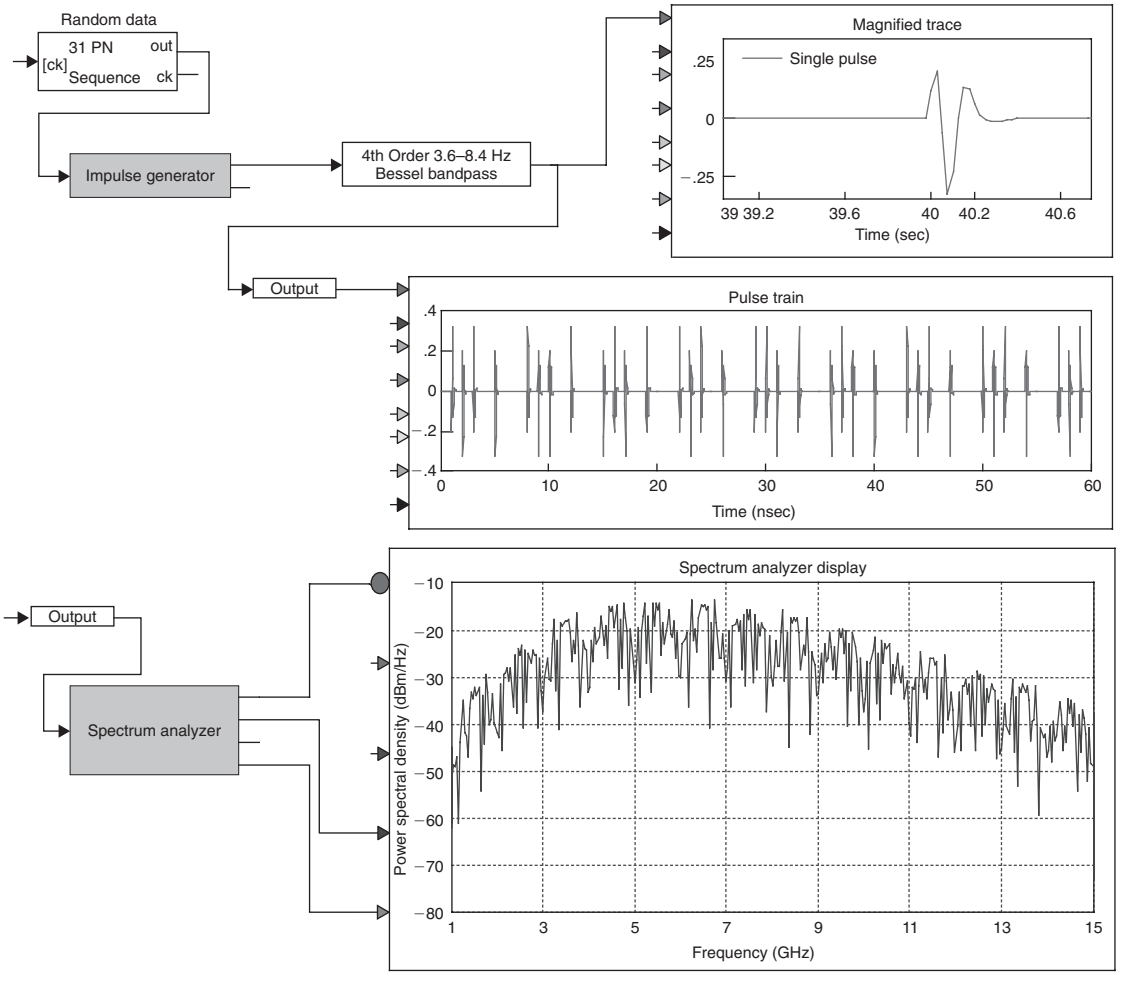


Figure 15.29: Simulated Sequence of UWB Pulses

There are several ways of representing a UWB pulse as "1" or "0." One method is to advance or retard the transmitted pulse with respect to the expected time of arrival of the pulse in the receiver according to the agreed pseudo-random time sequence. Another method is to send the pulse with or without inversion. In both cases the correlation of the received pulse with a "template" pulse generated in the receiver will result in a different polarity, depending on whether a "1" or a "0" was transmitted.

Detection of UWB bits is illustrated in Figure 15.30. A "1" monopulse is represented by a negative line followed by a positive line, and a "0" monopulse by the inverse—a positive line and a negative line. The synchronized sequence generated in the receiver is drawn on the second line and below it the result of the correlation operation $\int f(t) \cdot g(t)\, dt$ where $f(t)$ is the received signal and $g(t)$ is the locally generated sequence. By sampling this output at the end of each bit period and then resetting the correlator, the transmitted sequence is reconstructed in the receiver.

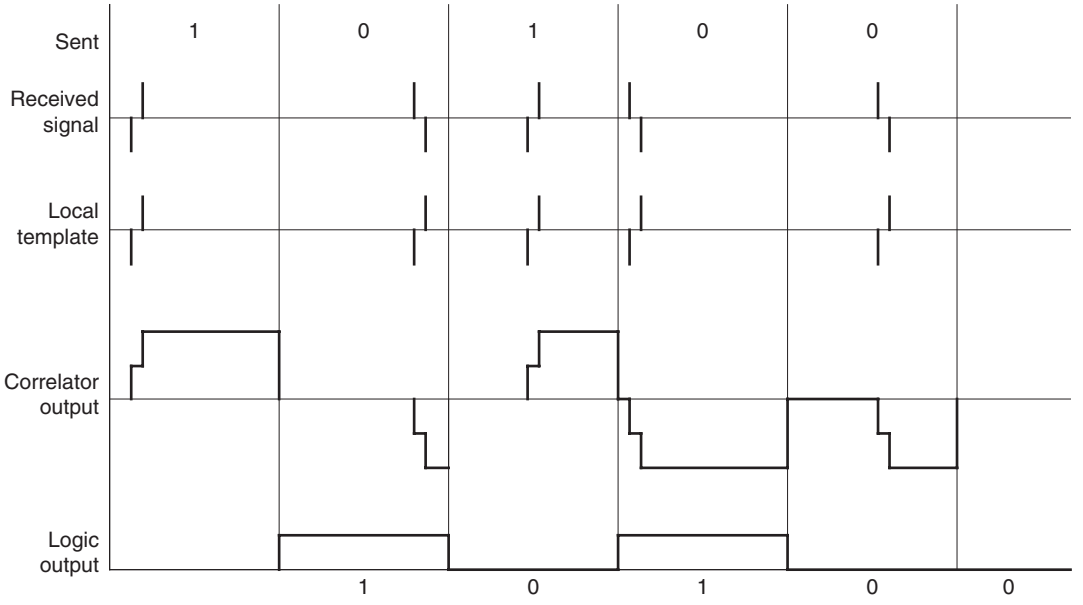


Figure 15.30: Detection of UWB Bit Sequence

As mentioned above, an individual bit can be represented by more than one sequential monopulse. Doing so increases the processing gain by the number of monopulses per bit. Processing gain is also an inverse function of the pulse duty cycle. This is because, for constant average power, the power in the pulses contributing to each bit must be raised by (1/*duty-cycle*). By gating out the noise except during the interval of the expected incoming pulse, the signal-to-noise ratio will only be a function of the power in the pulse, regardless of the duty cycle. An example may make the explanation clearer. Let's say we are sending data a rate of 10 Mbps. A UWB pulse in the transmitter is 200 picoseconds wide; 20 pulses represent one bit. The time between bits is $1/(10*10^6) = 100$ nanoseconds. So the time between pulses is (100 ns)/20 = 5 ns. The duty cycle is (200 ps)/(5 ns) = 25. Now the processing gain attributed to the number of UWB pulses per bit is $10\log(20) = 13$ dB. That due to the duty cycle is $10\log(25) = 14$ dB. Total processing gain is 13 + 14 = 27 dB.

A simplified block diagram of a UWB system is shown in Figure 15.31. A key to the generation of UWB pulses is the ability to create short impulse or step functions with rise

times on the order of tens or at the most hundreds of picoseconds, and to detect the UWB pulses that result from their application. High speed integrated circuits can be employed or special circuit elements, such as tunnel diodes or step recovery diodes, can be incorporated.

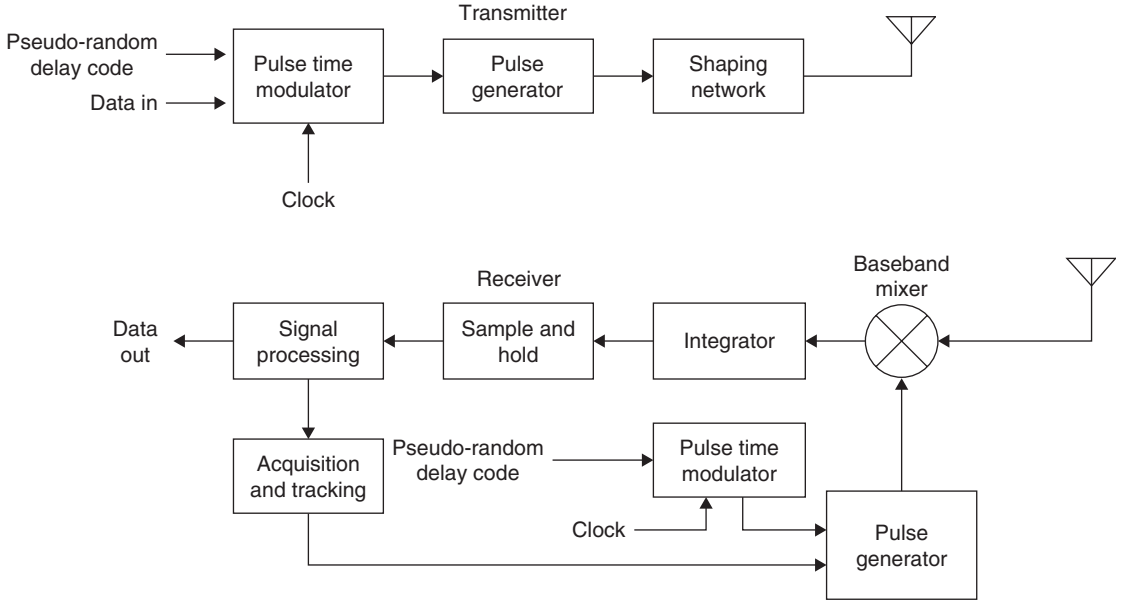


Figure 15.31: UWB Simplified Block Diagram

Conditions for using UWB in Europe are also being considered by the European Union. Due to the fear of interference with vital wireless services in the wide bandwidths covered by UWB radiation, spectral density limits allowed by the FCC and presumably to be permitted in the European Union are relatively low, on the order of spurious radiation limits for conventional unlicensed transmissions. However, as we have seen, high-processing gains can be achieved with UWB and communication ranges on the order of tens or hundreds of meters at highdata rates can be expected. Also, the FCC has indicated its intention to monitor the effects of UWB transmissions on other services, once equipment has been put into service in significant quantities, and the agency may be expected to modify or make its limits more lenient if interference is found not to be a problem. In any case, the unique characteristics of UWB are attractive enough to make this technology an important part of the offerings for short-range wireless communication in the years to come.

15.6 Summary

The annually increasing volumes for Bluetooth and Wi-Fi products, stimulated in a large part by the acceptance of industrial standards by the major manufacturers, are causing prices to fall

on complex integrated circuits as well as the basic RF components. This trend will open the way for the use of these parts in other short-range applications such as security and medical call systems. Use of sophisticated and proven two-way hardware and link protocols for these and other technically "low end" applications will open them up to much higher usage than they now command. A basic impediment to wireless use will still remain, however, and that is the problem of battery replacement. Reduced voltage and power consumption for integrated circuits will help, as will sophisticated wake-up protocols as are already built-in to Bluetooth. Range limitations may have to be dealt with by a greater use of repeaters than common in today's systems. Another area where advancements are affecting short-range radio is antennas. Both the use of higher frequencies and new designs are reducing antenna size and eliminating a visual reminder of the difference between wired and wireless devices.

The unconventional ultra-wideband technology, since its approval by the FCC, is opening up new civilian applications for short-range wireless, notably in the areas of distance measurement, concealed object location, and high precision positioning systems. Because of its high-interference immunity, and its property of not causing interference, it may successfully compete with and complement other technologies used for short-range radio applications such as personal communications systems, security sensors, and RFID tags.

In summary, advances in short-range radio communication developments in one area feeds its expansion in other areas. Overall, short-range radio will continue to play a major part in the ongoing communication revolution.

References

[15.1] ETSI TS 101 475 V1.3.1 (2001–12) Technical Report, HIPERLAN/2 Physical (PHY) Layer.

[15.2] Petroff, Alan and Withington, Paul, "Time Modulated Ultra Wideband (TM-UWB) Overview," Presented at Wireless Symposium/Portable by Design, Feb 25, 2000, San Jose, California (http://www.time-domain.com).

CHAPTER 16

System Planning

Ron Olexa

Now that we have reviewed the basics of radio operation, propagation, and predictive and actual performance measurements, it's time to see how this information is used as part of the design criteria in a system to actually provide services to a customer base.

System design must consider far more than just the RF aspects of the system. If the system is to function optimally and be cost effective, such diverse topics as equipment selection, real estate, construction, interconnect, power, and maintenance must be considered. Each of these topics has an initial capital cost and, with the exception of construction, an ongoing cost.

16.1 System Design Overview

Because of the myriad interactions you will encounter in designing a system, a flowchart is helpful for identifying the selection criteria for each of the key aspects of system design. Because there are so many different unique business opportunities that can be served with wireless data systems, it's impossible to review them all in this book. Instead I'll look at three distinctly different models, and walk through a design exercise for each. The first example system will be a single AP "hotspot" or small office LAN. The second example will be a far more complex MultiAP office LAN or "hotzone" requiring frequency reuse in its implementation. The third system will be a Wireless ISP (WISP) type system that is expected to cover a large outdoor area and provide Internet connectivity to a large, geographically dispersed user base. The WISP system could be composed of a single site covering a small town, or potentially hundreds or thousands of sites covering multiple counties or MSAs. Fundamentally they are all the same, though the complexities and need for managed spectrum grow with the size of the deployment.

Regardless of the scale of the system being deployed, there are a number of individual activities that have interaction with each other. For example, selecting locations for installing the radio hardware will be influenced by cost, coverage, and capacity needs of the system. Cost, coverage, and capacity are influenced by the selection of radio hardware and the frequency of operation. So, you can begin to imagine the complexity involved with the design of a large system. Each individual topic surrounding system design has its own associated flowchart which identifies activities and go/no go decision points. As well, each flow chart

must consider other parallel activities occurring under another separate topic, so that you assure that decisions and compromises made in the pursuit of one area of design do not negatively impact system viability because they ignored key factors of a separate decision matrix that they affected.

To simplify the overall decision matrix, I'll present individual flowcharts for each key activity. These flowcharts will show precursor or parallel activities that will need to be considered or reviewed when making final decisions surrounding individual key activities. After discussion of all the planning criteria, I'll show how these factors apply to real-world systems by using the flowcharts and decision matrices to plan actual systems, and show some of the trade-offs.

16.2 Location and Real Estate Considerations

Of course, the first thing you need to know is where the system will be deployed, what it needs to cover, and how much capacity is needed. In a hotspot or office LAN system, a physical address and suite number are necessary. In addition, a floor plan or other identifying criteria showing the area(s) to be covered should be acquired. Also discover as much about the property as possible. Blueprints, building drawings or other documentation concerning of the type of construction present in the building will be useful in the exercise of estimating coverage. Another key bit of information will be the name and contact information of the building owner, property manager, or other entity that may require coordination or approval of work on the premise.

If the system is of the WISP variety, there are additional needs. Since a building no longer defines the coverage, the physical area to be covered should be identified on a map or area image. This area should be inspected for available towers or multistory buildings that could be used to locate equipment and antennas. Latitude/longitude and height of antenna mounting locations of these buildings or towers should be identified. In addition, it is important to ascertain who is the owner or property manager of target buildings or towers.

You should also remember that many jurisdictions use zoning and permit processes for any communication facility, regardless of whether it uses licensed or unlicensed spectrum. It is critical to discover and comply with any local zoning or building and safety requirements early in your planning process. Failure to do so may lead to significant delays in deployment or, worst case, the local jurisdiction fining you and forcing you to cease operations and remove the equipment.

Because of the area to be covered, numerous possible equipment locations will exist. Determining where to concentrate your efforts requires a rapid assessment of which properties are best from an RF design standpoint. Assuming you have the specific operating characteristics of your equipment, the use of a propagation-modeling tool can prove valuable for assessing coverage from each of the location options.

In order to use such a tool, the parameters of the system need to be known, and assumptions need to be made about the conditions present at the CPE location. For example, if the CPE is

to be located indoors near the user's computer, additional path loss due to the construction of the building in which the CPE resides must be considered. If the CPE can be mounted outside, clear of local obstacles, then the propagation losses will be significantly lower, and the site's coverage greater.

As examples of this, the propagation plots in Figure 16.1 were computer generated using the Longley-Rice propagation model. The plots are based upon the same transmit power output and receive sensitivity. Only antenna gain and placement at the CPE has been changed. Figure 16.1a shows the coverage achieved from the base station or access point (the terms base station and access point can be and are used interchangeably. While the term access point was once unique to 802.11 hardware, it can now be seen referring to any number of base station products supporting wireless data) to a CPE unit using a 0 dBi gain omni antenna at street level, while Figure 16.1b shows the coverage achieved from the same base station to CPE using a 15 dBi antenna mounted at 15 feet elevation (sufficient to clear the roof of a one story home).

Figure 16.1a

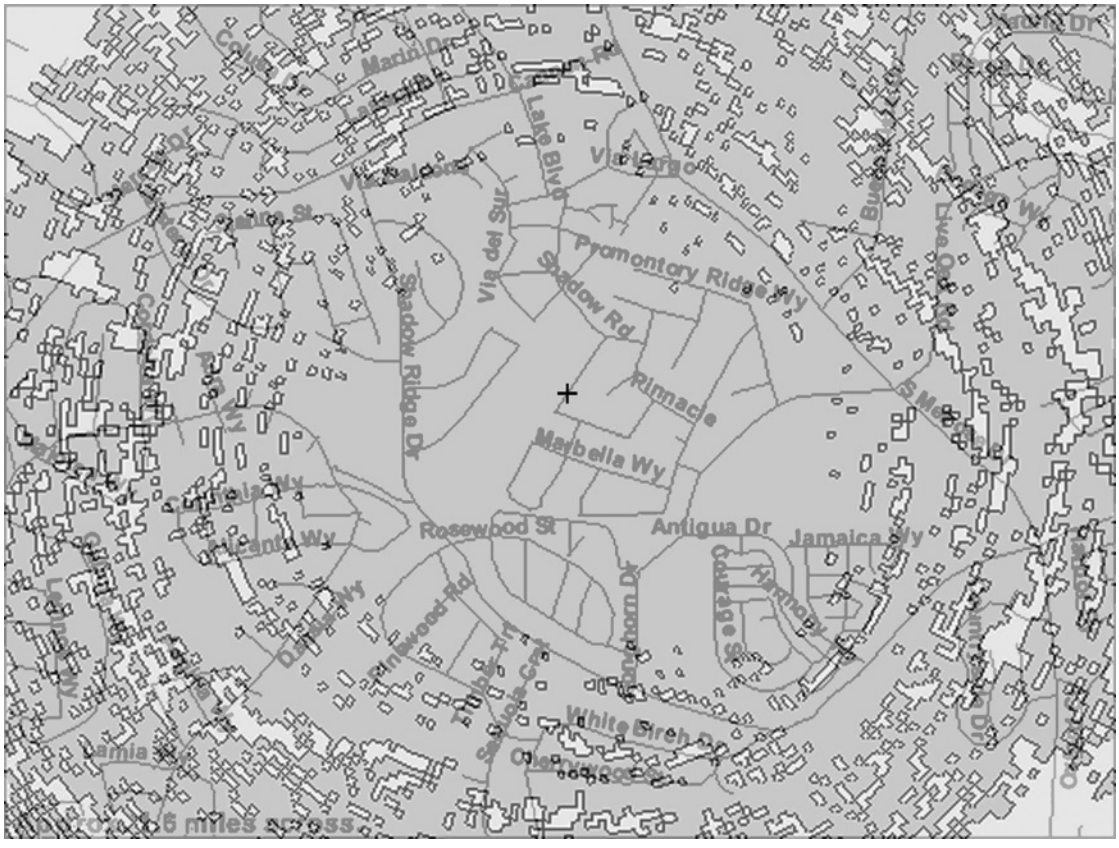


Figure 16.1b

The Longley-Rice model predicts long-term median transmission loss over irregular terrain relative to free-space transmission loss. The model was designed for frequencies between 20 MHz and 40 GHz and for path lengths between 0.1 km and 2000 km.

The Longley-Rice model is used in a number of commercial propagation models that are utilized for analyzing signal propagation in such commercial applications as cellular and PCS communication systems. The model is well known, and accepted by the FCC as a method of predicting coverage of broadcast facilities.

The model uses actual terrain data and predicts the median signal strength across this terrain based upon a combination of distance and terrain obstacles. The model also has a generalized morphology (land use) attenuation factor that can be attached to each site. This allows additional attenuation to be added to accommodate such things as building density and foliage.

While the coverage plots in this book show coverage by the use of a grey tone to indicate coverage area, the actual plots generated by the software are in full color and consist

of a number of colors, ranging from green to red. Each shade is associated with a 3 dB range of signal strength, with greens being high signal strength and red being low signal strength.

To use the prediction plot to design a system, you must determine three factors:

- How much fade margin does the system need?
- How much building attenuation must be overcome?
- Will the client device be externally mounted with a high gain antenna?

Since radio propagation is continually effected by multipath, the signal is always in a state of flux. This flux is known as fading. Even a stationary transmitter and receiver will see path fade between them based upon objects like trucks, cars and people moving in the environment. It is good engineering practice to use 8 to 10 dB of fade margin in a system design.

If the signal is to be received indoors, the building itself becomes an additional source of attenuation. This can range from 5 to 7 dB for wood frame construction to over 25 dB for office buildings with metallized glass facades.

Finally, the above factors need to be subtracted from the baseline receive sensitivity of the client device. In an 802.11b system, receive sensitivity on a high quality card ranges from −92 dBm for 1 Mbps throughput to −83 dBm for 11 Mbps throughput.

If the client device can make use of a high gain directional antenna, then this gain can be added to the path loss, or to make it simple, just add it to the receive sensitivity. So with a 15 dB gain Yagi antenna the −92 dBm sensitivity increases to (−92 dBm − 15 dB) or −107 dBm. You have not really increased the receive sensitivity, but you have added gain to the receive chain, which for all intents accomplishes the same thing.

So, a reliable communication link to a base client card at 1 Mbps requires a consistent −92 dB signal. Because of the fading nature of radio waves, it is necessary to add the fade margin to the −92 dBm minimum signal. This means that the signal required for a reliable link will need to be (−92 + 10) or −82 dBm. This signal level is sufficient to offer outdoor communication, but offers no margin to penetrate buildings. Additional signal is required for this. Again, 5 to 10 dB of additional signal strength will be necessary to penetrate light construction, so the required signal strength rises another 10 dB to −72 dBm.

If the client device makes use of a 15 dB gain antenna located outside above roofline, only the fade margin needs to be considered. Thus, a signal of (−92 dBm − 15 dB + 10 dB) or −97 dBm is required for a reliable communication path.

Looking at the contours and the legend, you can identify the areas that will be served with signals of the strength discussed above. Those points that show sufficient signal strength to meet the minimum requirements will generally be capable of supporting a reliable communications link.

Using the modeling tool can allow rapid review of a number of candidate radio site locations, and allow you to select the optimal locations for providing coverage to the target area.

As a location is identified, a propagation model should be run with the specifics of that location such as longitude, latitude, and height. This will give you a way to evaluate the anticipated coverage of each location and quickly identify which properties are ideal candidates and which are poor choices. By using the model for first pass identification of the best sites, you can focus your efforts on those sites first.

Another key consideration in the real estate realm is the availability of space, power, and interconnect for your system. Often the ideal RF location may be lacking one or more or these elements. Determining what additional costs and complexities are associated with getting the required space, power, and interconnect to the desired location is a critical part of the selection process. The flowchart in Figure 16.2 identifies the critical aspects that should be considered when selecting a site. As with most everything else in a radio-based system, there will rarely be a perfect solution, so compromises on certain factors related to sites will need to be made. For example, the site may not provide optimal coverage, but may be the only site in the area for which a zoning variance can be obtained. As shown in the priority of the flowchart, coverage should be the prime consideration, although often other considerations will dictate the use of a site that is an imperfect solution from a coverage standpoint. The flowchart is best utilized for analyzing a number of different site options. The tradeoffs associated with each can then be used to make the appropriate business decisions about the best location that has been analyzed.

16.3 System Selection Based Upon User Needs

The next considerations have to do with subscriber behavior. What form factor is acceptable for CPE? What average capacity requirements and usage characteristics will each user have? What security level will the user expect? And, finally, what is the acceptable initial and ongoing cost of the service?

These considerations are primarily driven by the needs of the customer. Meeting them will affect equipment and real estate decisions. For example, in a hotspot the equipment decision might be driven by the following logic: a hotspot serves customers that are using laptop computers or PDAs equipped with wireless access. The hotspot provides these users with short term Internet access. The system is primarily used for web browsing and email type activities. In this case the average use per customer is low, and the connection to the Internet

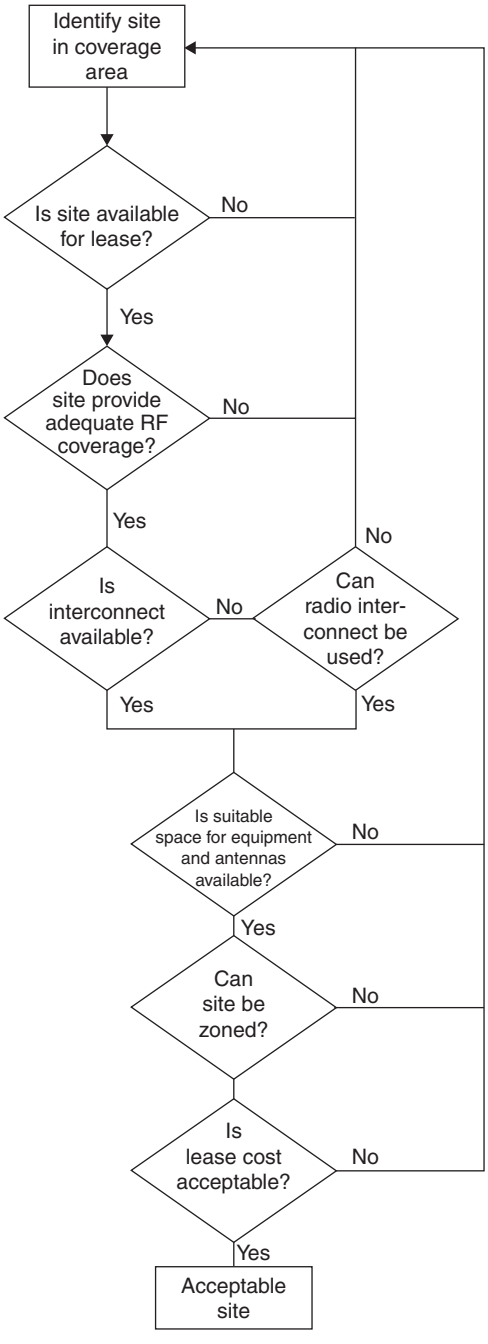

Identify site in coverage area
Is site available for lease?
No
Yes
Does site provide adequate RF coverage?
No
Yes
No
Is interconnect available?
No
Can radio inter-connect be used?
Yes
Yes
Is suitable space for equipment and antennas available?
No
Yes
Can site be zoned?
No
Is lease cost acceptable?
No
Yes
Acceptable site

Figure 16.2

is probably far slower than the speed of the radio interface, so the radio system is not the capacity bottleneck. In addition, the system must rely on the fact that the user is already equipped with a wireless interface in their computer. In late 2003 this situation calls for the use of 802.11b, since it is the most pervasive nomadic wireless standard available. This will probably not be the case in the future, as 802.11g, 802.11a, and 802.16 begin to permeate the marketplace. For the moment a cheap consumer quality 802.11b AP may be the perfect solution for the hotspot environment because it is inexpensive and includes network functions such as a low-end router, DHCP server, and NAT functionality. Thus it is a one-box solution for this particular environment.

In a large office LAN environment, the usage characteristics may be quite different. In addition to Internet and email access, the system will probably be used for transferring files and inter-network communications. In other words, the usage characteristics will be much higher; more like the usage characteristics of a wired LAN. Depending on the density of the users and the users' need for access to other systems (like a hotspot in a hotel or airport), 802.11a or 802.11g, with their greater throughput, may be better alternatives for serving this situation.

The WISP, on the other hand, is probably not best served by a traditional AP, or for that matter, 802.11. The customers will be geographically dispersed, leading to a requirement for large area coverage from a minimum of locations. The user will have usage expectations of this system that are similar to the expectations users have of wired equivalent services, like cable modem and DSL. The need for large area coverage means that the equipment will need to be tailored to high EIRP with high gain antennas, and an access sharing methodology much more robust than the CSMA/CA scheme associated with the 802.11 standard. As of the time this book is being written, 802.16 equipment does not yet exist, so the choices today are either a high gain 802.11 variant, like Vivato or YDI, or one of the purpose built systems from the likes of Motorola, Alvarion, Flarion, Proxim, or Navini. I anticipate that there will be continued development of new equipment to meet the needs of the WISP industry, and that we will see more manufacturers offer equipment tailored to this business.

Further expansion of the concept of matching equipment capabilities to the needs of the business and the user can be seen in Figure 16.3. This figure is a matrix that overlays typical user and system characteristics with equipment options. It can assist you in selecting the most appropriate technology to serve a particular user community.

16.4 Identification of Equipment Requirements

After gaining an understanding of the customer needs and expectations, the implementer should be able to determine what equipment meets those needs. As described above, the needs of the customer are a key driver of equipment selection, however they are not the only driver.

Subscriber service type	Designed for coverage		Equipment characteristics: CPE					Base station hardware						
	Indoor	Outdoor	Internal antenna	External antenna	Weather proof	AC power	Battery	External antenna	Weather proof	AC power	DC power	High EIRP	Standards-based	Proprietary
Fixed	N	Y	N	Y	Y	Y	N	Y	Y	Y	?	Y	Y	Y
Nomadic	?	?	Y	N	N	N	Y	Y	Y	Y	?	?	Y	N
Mobile	Y	Y	Y	N	N	N	Y	Y	Y	Y	Y	Y	Y	?
Hotspot	Y	N	Y	N	N	N	Y	?	N	Y	N	N	Y	N
Office LAN	Y	N	Y	N	N	Y	Y	?	N	Y	?	N	Y	N

Figure 16.3: User Needs Matrix

Size and environmental requirements, cost, manageability, reliability and availability all enter into the equipment decision matrix.

In the hotspot system, equipment should be selected that is compatible with the equipment that the users will have previously installed in their computers. This would drive equipment selection toward equipment operating on the prevailing standard adopted by the user base. Additionally, the equipment should integrate most of the network features that will be necessary to interface the equipment to the world, and to manage customers as they connect to the system.

The large LAN system will have its equipment selection driven by a more complex set of issues. The standard selected will have to be a balance between the now prevailing technical solution that is ubiquitous and low cost, and whatever standard is currently emerging as the next generation solution. This emergent solution will probably have a higher cost, since it has not yet reached mass appeal, and may also have some developmental wrinkles that still need to be fixed. On the other hand, it will also have greater capacity, greater spectral efficiency, and may offer better coverage and greater flexibility in deployment. Depending on your forecast of future usage demand and the need to allow your users to "roam" to other public or private locations and use their wireless access, one technology will be an appropriate, if not optimal, selection.

In this large LAN network, the additional network capabilities that were important to a hotspot are not necessary. Since this equipment will connect to an existing network, it is probable that all the advanced network features will be performed elsewhere on the network, and that the wireless equipment will only need to provide the wireless interface, and provide excellent remote management and fault isolation capabilities.

The WISP system has even more complex needs. Since standards like 802.11 were designed for wireless LAN type networks, they have not been optimized for serving a large area with dispersed users. While 802.11 has been made to work in these systems, other systems that were designed for use in this type of environment may offer a better technical solution. The trade-off here is cost. Because 802.11 is a mass-market product, equipment is very inexpensive and commonly available. Though a proprietary solution may be technically a better solution, it may be far more costly than using the 802.11 standard equipment.

Additionally, because of the nature of a WISP operation, the equipment must be located outside. This means that equipment must withstand the rigors of weather and an outdoor environment. Commonly available hardware designed for home or office use is not capable of surviving in the outdoor environment, so any equipment used in this environment will have to be located in a protected enclosure or will have to be designed for outdoor use. This leads to a set of long term maintenance issues, especially if equipment is mounted on towers: if the active electronics are located atop the tower, then the equipment will need to be removed and taken to the ground for service. Obviously, this is an additional ongoing cost to be considered in deciding the appropriate equipment design as well as the best location for the equipment.

Ultimately, any equipment solution has been designed with a set of expectations in mind, and has its appropriate place in the market. You need to understand the requirements of your system and the design intent of the equipment you are considering. Figure 16.4 provides an example of a comparison matrix that can assist with the identification of a solution that closely matches your needs. By completing such an analysis, you can be assured of selecting the manufacturer and solution that is best suited to serving your unique needs.

16.5 Identification of Equipment Locations

Now that you've narrowed the field of equipment options to a few that appear suitable for your system, we can begin to look at how and where to deploy sites to achieve the coverage and capacity desired in the system. Determining optimal antenna locations is the key to a successful deployment. An optimal location serves a multitude of needs: it provides optimal RF coverage, meaning it can be optimized to provide sufficient coverage of the area without leading to significant interference elsewhere in the system, it has easy access to power, it has easy access to network interconnect facilities, it can be easily installed and secured, and it has reasonable access for future service needs.

With the equipment selected, you have a baseline for the RF transmit power, receive sensitivity, and antenna options. These characteristics are used in conjunction with predictive modeling tools or survey tools to determine the area that can be covered with the selected equipment.

	Operating band		Suitability for serving environment					Costs	
	Licensed	Unlicensed	Office LAN	Hotspot	Fixed WISP	WISP	Mobile data	Base station	CPE
802.11b Traditional AP		Y	○	○	D	●	●	low	low
802.11b High power PtP		Y	●	●	D	D	●	high	medium
802.11g		Y	○	○	D	●	●	medium	medium
802.11a		Y	○	○	D	●	●	medium	medium
802.16	Y	Y	●	●	○	●	●	high	high
802.16e	Y		●	●	D	○	○	high	high
802.20	Y		●	●	●	D	○	high	high
CDMA2000	Y		●	●	●	D	○	high	high
Proprietary solutions	Y	Y	●	●	○	○	?	high	high

LEGEND

○ Optimal

D Suboptimal, but useful

● Not optimal

Figure 16.4

The first order of business is to evaluate the previously identified available locations for their suitability in providing coverage to the desired area, and so decide which tool is best suited to your need and determine the coverage potential of your sites.

By its definition the hotspot is a small open coverage area, so the simple path loss calculation is useable to determine if the system can cover the required area. In fact, in a hotspot system, finding a convenient mounting spot with available power and network connection is probably more important than finding an optimal RF location. Of course, just because I made this statement, your first attempt to build a hotspot will surely be in some unique and bizarre location where there are multiple unknown impediments to RF coverage. Even though the hotspot appears to be a simple deployment, it's still worth spending a little time validating your assumptions with a field survey.

Although the same estimation techniques used in a hotspot can be applied to a hotzone or LAN, determining optimal antenna locations becomes a little more complex. The size of the area to be covered, the user density within the space, and the layout of the space must be considered in order to optimize locations from a capacity and interference standpoint.

In the LAN environment, the primary usage of the system will be wired LAN replacement, thus the bandwidth requirements per user will be significantly higher than those of the casual user accessing the Internet. Depending on the user density in the covered area, you may find that a single base station may not have sufficient capacity to serve all the users in its coverage area. In this case it may be necessary to reduce power, relocate the base station, or change the antenna to provide a different coverage area that includes fewer users.

The first objective in designing such a system is to calculate the per user bandwidth requirements. Because there are so many opinions about how to accomplish this, I will not discuss it here. Use whatever method of calculation you are familiar and comfortable with. Once you know the average usage per user, you can determine the total users per base station by this simple calculation: I_t/BW_{user}, where I_t is the radio information throughput. Do not confuse this with raw device bandwidth. We need to use the achievable device throughput based on the types of traffic on the network. For example in 802.11b the channel is advertised to have 11 Mbps throughput. In reality this is the total channel throughput including all overhead. The information bandwidth of the channel is significantly less, more on the order of 4.5 to 6 Mbps depending on the types of traffic on the network. Also, because the 802.11b channel is a TDD channel, the total throughput is shared by both upstream and downstream traffic, meaning that another derating factor must be applied based on the mix of upstream and downstream capacity requirements. All this means that the real data throughput of an 802.11b system may be as low as 2 Mbps for bidirectional symmetrical usage.

BW_{user} is the average bandwidth requirement. This is not the peak requirement of the user, but either the average usage, or the predetermined lowest bandwidth level available to any user during peak usage periods.

For example, in a system with symmetrical uplink and downlink requirements, and an average bandwidth per user requirement of 200 Kbps, an 802.11 system will support only 10 to 15 users per AP (2 to 3 Mbps information throughput/200 Kbps per user). Remember, 802.11 has 4.5–6 MBPS total throughout. Since this example uses symmetrical traffic, the 4.5–6 MBPS is shared by the uplink and downlink traffic, thus leading to 2–3 MBPS available in either direction.

Assuming a normal office environment with cubicles and walkways, the average space allocation per employee is 250 square feet. In this environment an 802.11b AP will cover about a 50-foot radius, or about 8000 square feet of area with maximum bandwidth. This area may contain up to 32 users. In this situation the coverage defined area exceeds the capacity defined area.

There are several solutions for this. Antenna selection and power reduction can reduce the area covered to one that is more in line with the capacity needs of the users. Alternately, this

may be a situation where 802.11b is not an optimal technology selection. 802.11a or 802.11g may be more suitable protocols because of their higher throughput. Either 802.11g or 802.11a will provide about 5 times more bandwidth than 802.11b, however the area covered by this bandwidth will be smaller than the 802.11b maximum throughput coverage area. In this same environment, 802.11g will probably cover less than 3000 square feet with maximum bandwidth signal levels, while 802.11a may serve only 1000 square feet at maximum bandwidth due to the additional propagation losses associated with its higher operating frequency. Do remember that these technologies rate adapt to lower throughput speeds as the signal strength drops. It is entirely possible that the needs of users can be met over a coverage area associated with one of more of the data sub-rates.

This is a case where future growth and changes in usage characteristics should be considered as part of the selection criteria. If capacity needs are expected to increase over time, then a higher bandwidth standard like 802.11a or 802.11g should be considered instead of 802.11b.

Now that the capacity versus coverage issues have been addressed, you know how much area should be covered by each radio base station location. Now you can begin to plan the location of hardware to meet the needs of the deployment. Use the drawings, blueprints and other reference data you've collected in combination with a physical site review to identify locations where the equipment could be placed giving easy access to power and interconnect, easy access for maintenance, avoidance of utility walls and massive metal objects, as well as located centrally to the desired coverage area of each radio base station. Depending on the physical layout of the space to be covered and the availability of power and interconnect, a number of location options are viable: you could use an omni antenna located on the ceiling in a central location, or you could use a directional antenna located high up in a corner or along an outside wall and pointing toward the space to be covered. You might try drawing shapes consistent with the antenna pattern (circles for omni antennas, cardioids or teardrops for directional antennas) and scaled to an appropriate size to represent the desired coverage on the blueprints. Lay out the shapes on the blueprints, arranging them so as to provide coverage to all the desired locations. Then physically check the locations to assure that power and interconnect are easily available at the locations. If not, see where you need to move it to gain easy access to power and interconnect. As you move the locations to ease deployment, try to keep the spacing between the base stations as even as possible. This will make your frequency planning easier. You may also want to do a site survey on the selected locations. The reason is twofold: you can assure you get the expected coverage and, more importantly, you can determine the maximum coverage and therefore the interference area of each location. Knowing this will be useful when allocating channels to each location. Figure 16.5 shows one possibility for locating equipment to serve an office space. Depending on the unique needs and limitations of the space you are working with, such a solution may or may not be feasible for your deployment.

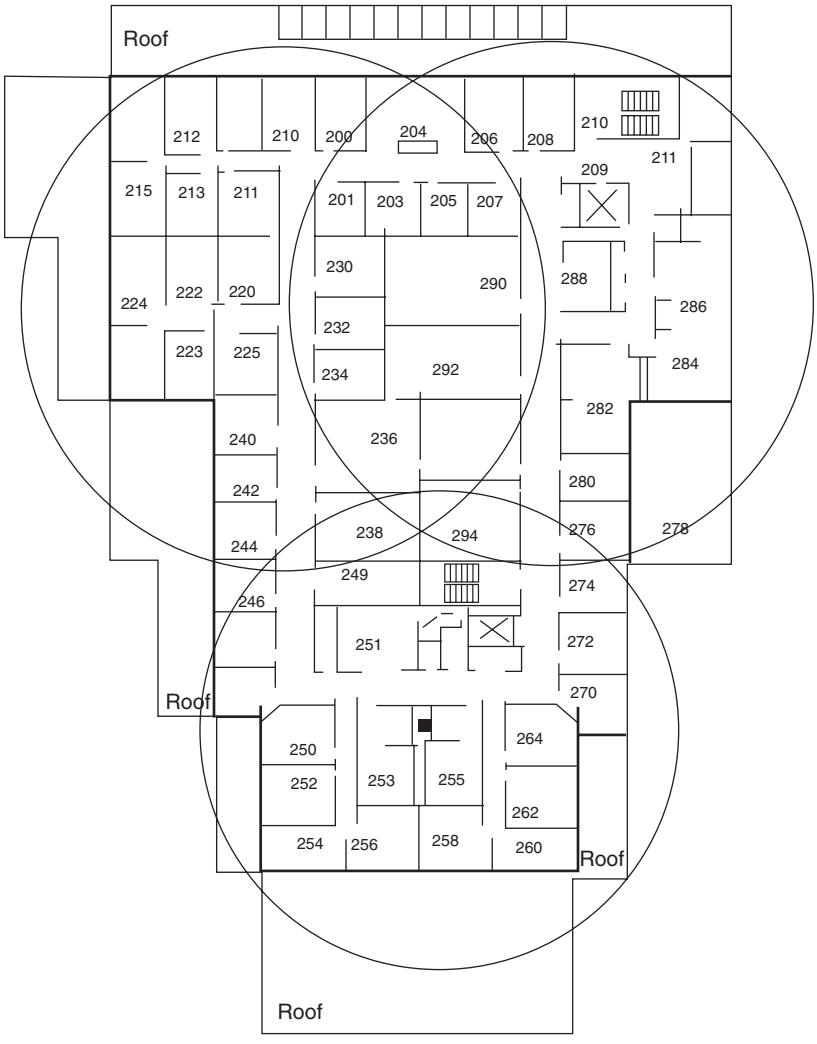


Figure 16.5

The WISP solution has much in common with the LAN system deployment. Coverage and capacity are both critical issues, multiple locations may be required to address coverage or capacity issues, and technology options will need to be considered according to system and cost requirements. The WISP system is designed to cover an extended area of potentially multiple square miles with a complex mix of terrain, morphology, and user locations. The system will be designed to provide Internet access service to residential and/or businesses within the coverage area. In addition, the location of the CPE is a consideration in determining the coverage of each site. If the CPE is located inside a structure, additional losses will be incurred, leading to a smaller reliable coverage area. Alternately, if the CPE can be located outdoors above the roofline, coverage distances will improve substantially.

Because of the size and complexity of the covered area, there is a very real possibility that users will be shielded from one another. This gives rise to problems using equipment that utilizes CSMA/CA, like 802.11-based hardware. Because CSMA/CA works by monitoring the channel prior to transmitting, the algorithm assumes that all users are close enough to each other to be able to hear each other. In the widely distributed WISP system, this is not true. This leads to a problem called the hidden transmitter problem, where users all have a clear path to communicate with the base station, but cannot hear each other. Since they cannot hear each other, they all assume the channel is clear, and try to transmit. The effect of this is that multiple stations to try to access the channel simultaneously, because they all think it's clear. In reality, the base station hears all the simultaneous users and cannot discriminate one from another. In essence the simultaneous users are all interfering with each other.

This behavior can be tamed using a feature in the 802.11 standard called RTS/CTS, or Request to send/Clear to send. This is simply an additional protocol where a user asks the base station for permission to transmit, and waits until it receives permission before beginning its transmission. There are three downsides to this scheme. First, RTS/CTS is not a perfect solution. Collisions can still happen. Multiple stations may be asking permission near simultaneously, then all transmitting based on hearing a CTS belonging to a different user. Second, RTS/CTS takes overhead, because each transaction has to be preceded by a request and acknowledgment, thus leading to a further decrease in real channel data throughput. Third, since closer users have a stronger signal, it is quite common for those close in users to get an unfair share of the available bandwidth. This is because the closer-in users tend to have more signal strength at the base station, and the additional signal is easier to hear than the weak signal from the further-out user. In some cases it is possible for the close-in station to override the weak signal completely, and be the only signal heard by the base station.

Conversly, 802.11 equipment is cheap and abundant, so from a cost standpoint it may be the best solution for a particular need. Just be aware that as with any system offering service for hire, a WISP needs to assure that the users have fair and equitable access to the service. Clearly, even though the equipment is inexpensive, the issues surrounding the access methodology need to be given considerable thought. While it may be coerced into working, it is entirely possible that the solution will never be as good as a purpose built solution.

There are other systems that have taken the unique needs of large area access into consideration. 802.16 and 802.20 were designed with the needs of wide area coverage and possible mobile access in mind. So have some of the proprietary system standards that have been created by equipment manufacturers. These systems have the benefit of being designed for the express purpose of providing wide area access to multiple geographically diverse users. The downside is cost and/or availability. At the time this book is being written (Sept 2003), there is not yet commercially available 802.16 or 802.20 equipment, and the proprietary equipment that is available costs 5 to 50 times more than 802.11b equipment.

The wireless data field is still growing. New standards and vendors continue to evolve. I suggest that you carefully review the technologies and vendors available to you at the time you are designing the system. Determine their suitability to your unique business plan, and make your selection based upon your unique mix of cost, capacity, coverage, and spectrum availability requirements.

Once a technology has been selected, you need to review the coverage and capacity needs of your system. Because you are covering a large outdoor environment, propagation modeling can be an effective way of estimating coverage area, and assessing the coverage available from specific sites. Since the users cannot be identified by desktop location as they could in the in building LAN system, another method of estimating user density is needed. This is where demographics data can come into play. Such data is available from many sources, and has varying resolution. You can find demographics as coarse as an entire county, or as fine as fractions of a square mile. Use this data to determine how many households and businesses are in the geographic area associated with coverage. Then by using your expected market penetration percentage, you can determine the number of users in the area. You should already have an idea of the expected per user traffic, so you can now use the same formula we used in the LAN example to determine the traffic capacity needed for the area. Once again, determine if the covered area has sufficient capacity to meet user demand. If not, reduce the coverage area and add more radio sites to the system as needed to serve the territory.

Once you have determined your site locations, one other consideration you will face is connecting the sites together or to the Internet. Since the sites will have significant distance between them, CAT5 cable, which can only support 300 feet of link distance, is not a viable methodology to connect them together. There are several alternatives for interconnection to the Internet: purchase an individual access facility from a telco, CLEC (Competitive Local Exchange Carrier), or other provider for each of your radio sites, or use point-to-point radio to connect your facilities together and bring all traffic to a common location for delivery to the Internet.

The former solution may lead to a more robust system, because a single facility outage isolated only one site and its associated coverage area. It will also be more expensive because you cannot aggregate traffic from multiple sites in order to most effectively use the capacity of the facility you're paying for. Additionally, there will be more equipment necessary because each site will have to have it's own network hardware to provide access control, DHCP, maintenance access, and other network and security functions.

The latter solution allows all the network equipment to reside at a single location, thus reducing the need for redundant equipment at each site. It also leads to the additional cost of a facility to connect the sites back to the designated central location. This could be accomplished with a leased Telco facility, although you'll probably be limited to the T-1 or E-1 facility speeds of 1.544 or 2.048 Mbps, or multiples of this. Given that this may be a fraction of the bandwidth of the radio, it may not be an appropriate choice. A better choice may be point-to-point radio facilities.

If the sites in question have clear line of sight to each other or to a central location, this becomes a feasible implementation. There is equipment available in both the licensed and unlicensed bands that can be used to provide this type of connection. Better yet, these radios can be had with an Ethernet output, so they can simply be plugged into your other network equipment without the need for specialized equipment to convert from Ethernet to some other standard like T-1 or E-1.

As discussed each of these solutions has tradeoffs. Use the flowchart in Figure 16.6 as a basis to assist in selection of the most appropriate solution based upon your unique environment and needs.

16.6 Channel Allocation, Signal-to-Interference, and Reuse Planning

The number of available channels in a system will be predicated on three things: the spectral allocation available to you, the spectral requirements of the equipment you've selected, and other users or interferers in the band.

When you did the site surveys at your locations, one of the things you noted was the noise floor, and any other users in the band that you noted. If possible, avoid channels with existing users or a high noise floor.

In the hotspot system this is simple: pick the quietest channel and implement it as the operating channel of your equipment. Since there is only one radio base station or AP, you're done.

In the office LAN or hotzone, you must consider not only outside interference but also the interference you will self generate when all the sites are active. This means that you must carefully allocate channels in order to minimize interference between locations. In general, get as much physical separation as possible between co-channel locations. As detailed in Chapter 8, the building layout and coverage information collected during the site survey will be invaluable in determining how to allocate the channels based on the actual coverage and overlap of each location. Ideal channel separation, i.e., sufficient so as to cause no interference, is rarely achievable. The best separation that can be designed for is separation sufficient to provide adequate C/I margin to the majority of the coverage area of each base station. The C/I requirements of the system will vary based upon the technology. If the manufacturer does not publish C/I requirements they can generally be thought of as identical to the S/N requirements of the technology, which by the way are normally the published receive sensitivity numbers. The receive sensitivity is nothing more than the amount of signal required for the equipment to perform at a certain threshold level. The reference against which this is measured is thermal noise in the channel.

What does this mean to the design? It means that the coverage and/or capacity performance of the network will not be as good as in would have been in an interference free environment. The required signal strength will increase by the number of dB the interference has raised the noise floor.

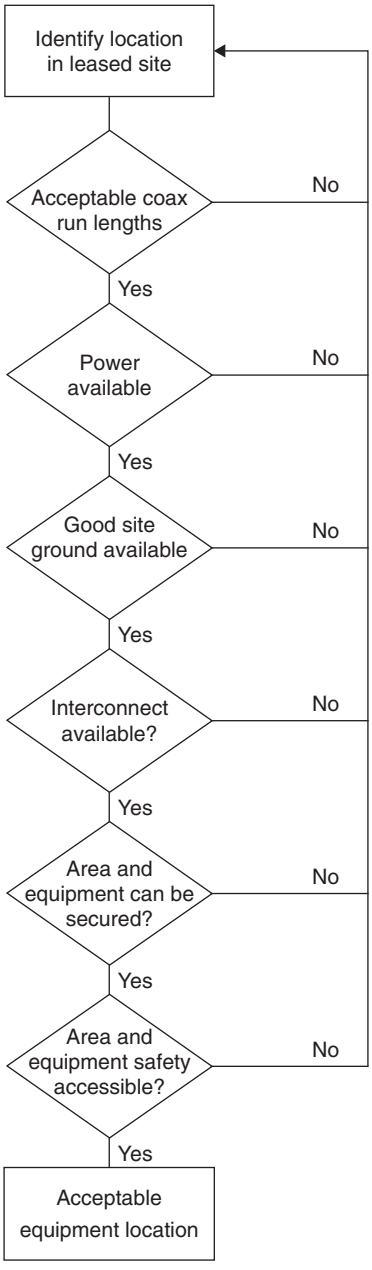


Figure 16.6

For example, the noise floor of a 20-MHz 802.11b channel in the 2.4-MHz band is −100.43 dBm as calculated by the thermal noise equation. The published receive sensitivity specifications are based on this noise floor. If interference adds undesired signal (man-made noise) to the coverage area, the noise floor increases above the level contributed by thermal

noise alone. The basic receive sensitivity does not change, but the system performance does. For every dB interference adds to the noise floor, the perceived receive sensitivity will be worsened by an equivalent amount. In the case of an 802.11b device with a published 1 Mbps sensitivity of −92 dBm, this sensitivity is based on an expected noise floor of −100.43 dBm. If interference raised the noise floor to −98.43 dBm, the device would no longer perform when the signal strength was −92 dBm. The interference has raised the noise floor by 2 dB, so the new signal requirement would be −90 dBm for the device to offer the same 1 Mbps performance level.

This is another reason why the site survey is a useful system-planning tool. By knowing the signal level contributed by other locations, you can assess how much interference adds to the noise floor. This allows you to estimate the real coverage of locations based on the additional noise level caused by co-channel users in the form of interference.

The WISP system requires the same diligence in allocating channels for the same reasons as those in the LAN environment. Interference, regardless of its source, will negatively impact either the coverage potential of a site or its ability to provide maximum throughput over its designated coverage area. Pick the channel with the lowest noise floor, and use antenna aperture, downtilt, and site separation distance to assure sufficient isolation of co-channel signals.

Antenna downtilt is another subject that the WISP operator should become familiar with. Because the WISP system is located outside, and is probably located at some elevation above ground, downtilt becomes an important factor in optimizing coverage and minimizing interference. Think about it this way: the antenna you will use has a vertical polarization, and has a main lobe with the highest energy density pointed perpendicular to the antenna orientation. This means that the main beam will be pointed horizontally, 90 degrees shifted from the mounting orientation of the antenna, in other words at the horizon.

As you can see in Figure 16.7, the main beam points at the horizon, and it is entirely possible to "miss" the intended service area with the main beam and to end up serving the desired area with side lobe and sub lobe energy. This situation leads to impaired service in both the desired coverage area and surrounding areas, because the main lobe energy misses the target user, and increases interference in adjacent areas because the main lobe energy is pointed toward the horizon, and other potential users.

Downtilt corrects both of these situations by pointing the antenna's main lobe toward the desired coverage area. This can be accomplished two ways: mechanically and electrically. Mechanical downtilt is used only with directional antennas, and is accomplished by physically mounting the antenna in such a way as to tilt it towards the ground by some number of degrees. Electrical downtilt is achieved by the design of the antenna, and can be applied to both directional and omnidirectional antennas. In fact, electrical downtilt is the only way to

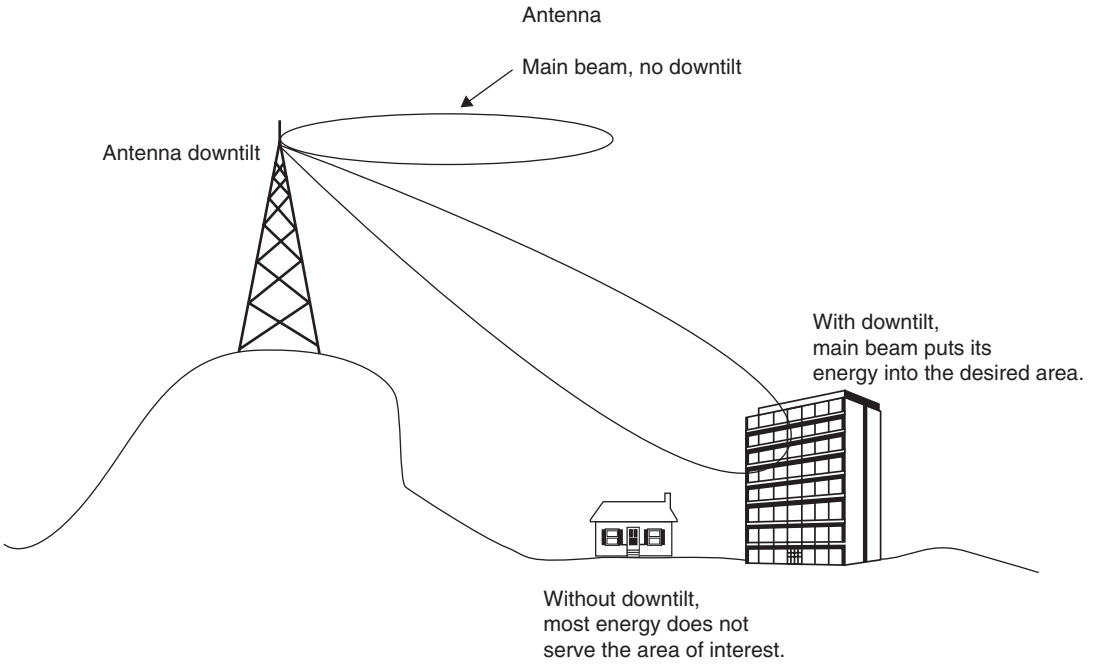


Figure 16.7

implement downtilt in an omni antenna. The number of degrees of downtilt is calculated with the simple formula: *arctangent (H/D)* where *H* is the effective height and *D* is the distance to the far edge of the desired coverage area.

To keep the calculation simple, the formula is based on calculating the value of the adjacent angle of a right triangle, therefore the effective height is determined as the difference in elevation between the antenna and the area to be covered, and the distance is the physical distance between the antenna and the far edge of coverage in the same units used for height measurement. For example, let's take a case where the antenna is mounted on a tower, which is on a hill overlooking the coverage area. The ground elevation at the edge of the coverage area is 540 feet, the top of the hill is 650 feet, and the antenna is mounted at the 100-foot level on the tower. In this case, the effective antenna height is (650 + 100) − 540, or 210 feet. If the distance to the far edge of coverage is ½ mile, then atan(210/2640) = 4.548 degrees of downtilt.

As shown in Figure 16.8, by downtilting the antenna by 4.5 degrees, the center of the main beam is aimed at the users furthest from the site, thus maximizing the energy density in the area furthest from the site. Areas closer to the site have less path loss due to distance, and are effectively served by the lessening energy density of the main lobe and sub lobes of the antenna. The other major benefit of downtilt is interference reduction. Without downtilt, the

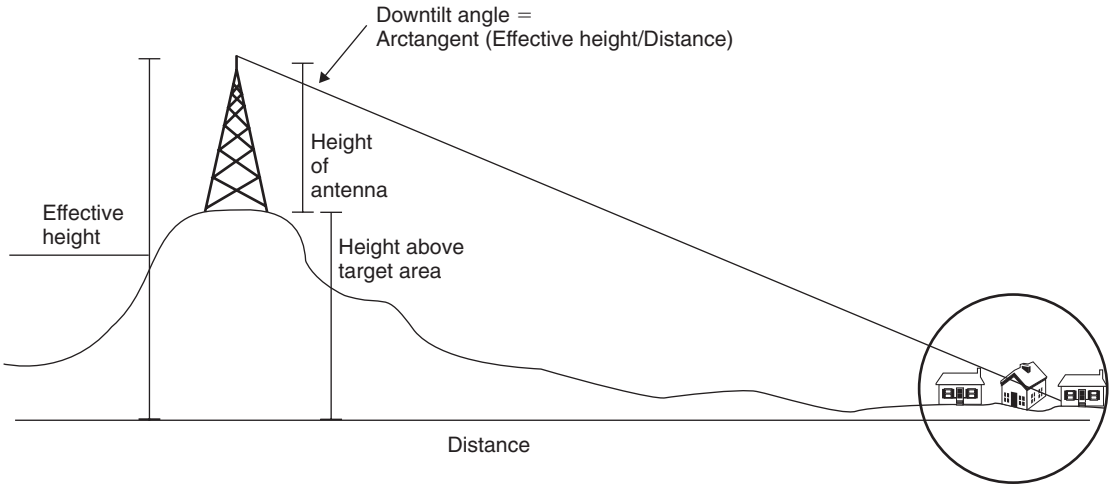


Figure 16.8

desired coverage area had less than ideal signal, while undesired areas were getting most of the site's energy.

By using downtilt to maximize energy density in the desired coverage area, a side benefit is that this energy no longer gets to places it does not belong, and therefore does not appear as interference in some undesired area.

16.7 Network Interconnect and Point-to-Point Radio Solutions

In small venues like a hotspot or small office, network connectivity should not be a significant issue. CAT 5 cable is an inexpensive and readily available solution for connectivity in these small area locations. Even large office spaces can have the radio solution effectively networked using CAT 5 cable. The distance limit for CAT 5 cable is 300 feet per run, which allows quite a large area to be served by cable alone, with no need for intermediate regenerators. If greater distances are needed, it is feasible to divide the network on a floor by floor basis, and run cables to a central point where they are connected using a switch or router, which is in turn connected to the wired network in the building. The decision of how to cable will be dependent on the existing wiring, existing network, and construction of the property.

But what about the campus environment, where the space to be served with wireless is in a number of disparate buildings, or the WISP system that covers a community from a number of different sites? In these cases, the distance limits associated with CAT 5 make it unusable. In the campus environment, Ethernet over fiber (10Base-FL and 100Base-FX) connections may be feasible, depending on the availability and cost of duct space between buildings. In the WISP system where such ducts between sites are rarely available, or in the distributed building office environment where no ducts are available, another interconnect option needs

to be considered. That option is of course radio. Not only can radio-based systems be used for connecting users, they can be used to connect together the disparate locations.

Radio facilities such as these are called point-to-point links, and can use a variety of licensed and unlicensed bands for operation. Unlicensed band links commonly offer limited bandwidth (1 to 50 Mbps), while licensed microwave bands can offer links with hundreds of Mbps throughput. When a single central facility is used as a distribution point for a number of remote locations the resulting network is known as a "hub and spoke" configuration. Figure 16.9 illustrates such a network.

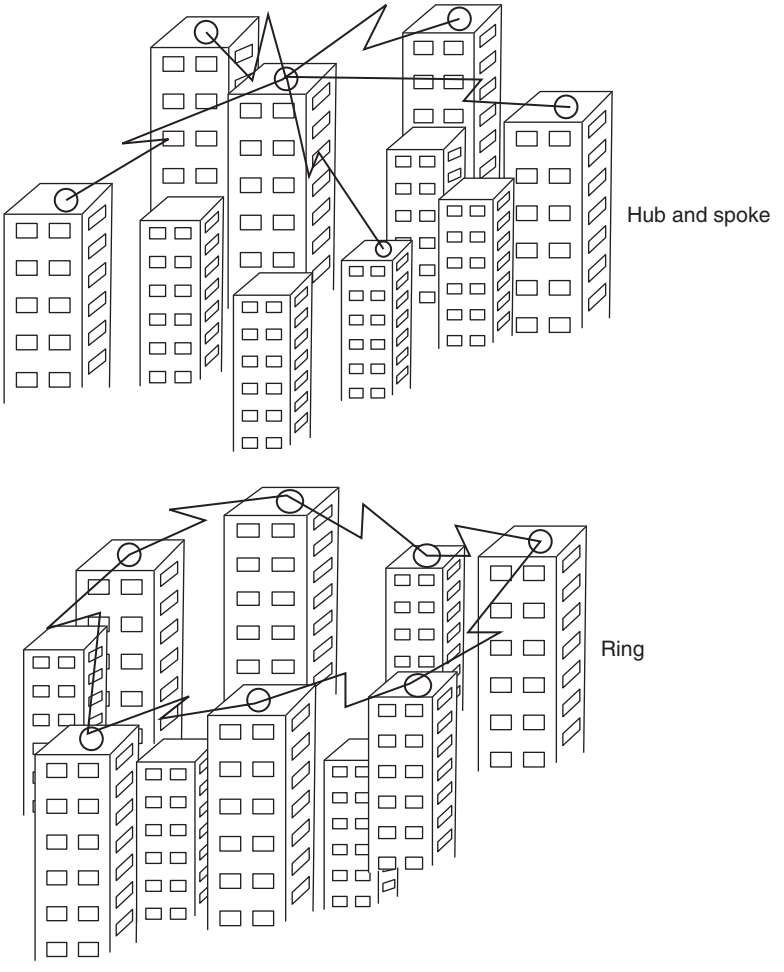


Figure 16.9: PTP Links Configured in Hub and Spoke and Ring

These links can be designed with the same tools you use for designing the radio networks for the users. The big difference is that there are only two points to connect, they are both known points that do not move, and they should be within line of sight of each other. This eases

the design because you need only consider the coverage at two points in space that can see each other. If the locations have clear Fresnel zones, Free Space losses can be applied; if the Fresnel zone is cluttered then line of sight loss characteristics should be used. Because there are only two locations, highly directional antennas can be used with the radios. This helps overcome the path loss as well as reduce unwanted interference by keeping the energy tightly focused on the other station. In fact, these antennas can have less than two degrees of aperture, depending on their size and frequency of operation. These small apertures mean that mounts must be sturdy, and the antennas must be carefully aimed at each other to assure the stations are initially, and remain, accurately pointed at each other.

A new concept to consider on these links is redundancy. Because these links are used to connect traffic-bearing locations together, a failure of one of these links will isolate those traffic-bearing locations from the rest of the network, thus leading to user communication failures. This single point failure can be remedied by making the interconnect links redundant. This can be accomplished in several ways.

The first method is the simplest but perhaps most expensive way to accomplish redundancy: use redundant radio equipment. This can be done in two ways, the first being to have two active independent radios each bearing half the total traffic. The advantage is that a failure will not isolate the end site, but may, depending on the total throughput capabilities of the radios, reduce the total traffic capacity that can be handled by the remote site. The disadvantage is that two independent channels are needed, and equipment costs are significantly increased due to the need for two radios. Another redundancy method is "Hot Standby." In this method, there is still the need for two independent radios, but only one is active at any given time.

If the equipment senses a fault, it automatically switches to the standby radio and allows traffic to flow unimpeded. There are two advantages to Hot Standby: only one RF channel is needed, and a failure does not impact traffic flow from the remote site. The disadvantage of cost increase over a single radio still exists in this scenario.

Full Redundancy and Hot Standby redundancy are commonly used when only a single remote site is connected, or there are multiple remote sites that cannot be connected to any other location. If more than one remote site needs to be connected and there are multiple choices for paths to connect them, another network topology called a ring should be considered.

In the ring network a single site connects to remote site one, remote site one connects to remote site two, and so on. The last remote site connects back to the origin site as shown in Figure 16.9. In this scenario, traffic can flow around the ring in either direction. A single failure cannot isolate a site, because there is an alternate path to route the traffic. There are several disadvantages to this topology: Backhaul facilities need to be large enough to handle the traffic presented by multiple sites, additional network hardware must be configured in each site in order to drop and insert traffic from the site, and to manage the traffic flow on the ring,

and finally, the locations need to be arranged in such a manner as to accommodate the ring topology. The primary advantage of the ring architecture is cost: one extra radio per ring is needed to achieve redundancy, instead of one extra radio per remote site. Another advantage of a ring architecture is that it can be used to extend the service area far beyond the distance limits of a single radio link. Considering each link as an independent starting point allows you to develop a network that continues to expand outward from each site on the ring. The only limitation is that at some point the network needs to loop back in order to close the ring.

From a frequency utilization standpoint, it is best to use different frequencies for the back-haul and network connections than you are using for connecting users to the system. Once again this has to do with potential interference generated by the additional facilities using common channels. Even though the backhaul facilities use directional antennas, there is still a possibility for mutual interference. In fact having backhaul in a completely different band sometimes makes sense because it allows all the channels available in your primary band to be used for providing service to users, thus allowing growth of capacity in the network. Also there may be other bands more suited to the needs of back haul because they have the ability to connect over greater distances or provide greater capacity.

16.8 Costs

The last discussion of this chapter might actually be your first consideration: what's it all going to cost, or looked at another way, how much can I afford to spend and what compromises will be necessary? Costs are largely broken into two categories CAPEX and OPEX. The CAPEX costs are one time costs associated with capital equipment, construction, design and planning, and so forth. OPEX are recurring costs associated with such things as interconnect, leases, utilities, maintenance, and other month-to-month costs.

Network design will have an impact on all these cost elements, and the final network design will have to include the trade-offs associated with costs as well as performance. As I've said before, there is no free lunch. For example, using cheap equipment to reduce CAPEX may have a serious negative impact on OPEX, because equipment reliability or maintainability suffers. There can be many consequences of CAPEX decisions that in the end have a significantly greater cost effect on OPEX. Be careful when making CAPEX financial decisions and make sure you consider the long-term ownership costs too. You may have to live with the aftermath of your capital decisions for a very long time.

16.9 The Five C's of System Planning

This and previous chapters have discussed elements of system selection, design and performance. They have also mentioned the fact that trade-offs are necessary when selecting a solution or planning a system. All of those discussions have led us to here: the five fundamental aspects of a real-world communication system. In its simplest form, there are five elements

that will affect your decision on what technology to select and implement. Those fundamental elements are **C**ost, **C**overage, **C**apacity, **C**omplexity, and **C**/I ratio. Cost includes both initial and ongoing costs of ownership, coverage can be either the total area to be covered or the area covered by a single base station, capacity can be either the necessary capacity or the capacity of an individual channel or base station, complexity can be defined as the overall size of the network (for example how many individual sites and discrete pieces of hardware are necessary to make the system function), and finally C/I or the amount of interference that needs to be tolerated. This interference can come from two sources: it can be internally generated through channel reuse in the system, or it can be generated by other systems over which you have no control.

These elements are inextricably interrelated and are almost mutually exclusive. For example, you cannot have a cheap but complex system, nor can you have a simple and inexpensive system that also has great coverage and capacity. The selection of a suitable system solution will be driven by an understanding of the business needs, the user expectations, and area to be served then balancing those needs against a prioritization of the five C's. This balance will be different in each situation. Always remember that a successful set of trade-offs in one situation may lead to a dismal failure in another situation.

Use these five elements when initially formulating your system requirements. Ranking these factors in order of importance to your business helps to organize your selection process by allowing you quickly to eliminate choices that obviously do not fit your hierarchy of need. For example, if low cost and maximum coverage are the most important criteria, a single site high power solution might be the best solution for your need, providing equipment and spectrum for such a system could be obtained. On the other hand, if high capacity and dealing with a limited coverage area and hostile interference environment are most important, then you may need to deal with the additional costs and complexity of a network requiring multiple sites in the coverage area. This of course gives rise to considering the number of channels needed to support the multiple sites and an analysis of whether the reuse required by such a plan can be managed.

As you see, the questions and considerations quickly multiply, and each time you make one decision, you may eliminate or severely modify another of your assumptions or requirements. Learning the relationships between the five C's as driven by your unique system requirements will allow you to make informed decisions and optimize the necessary trade-offs into a system that best suits the financial, user, and operational needs of your particular situation.

This page intentionally left blank

Index

G

H

I

K

L

M

S

T

U