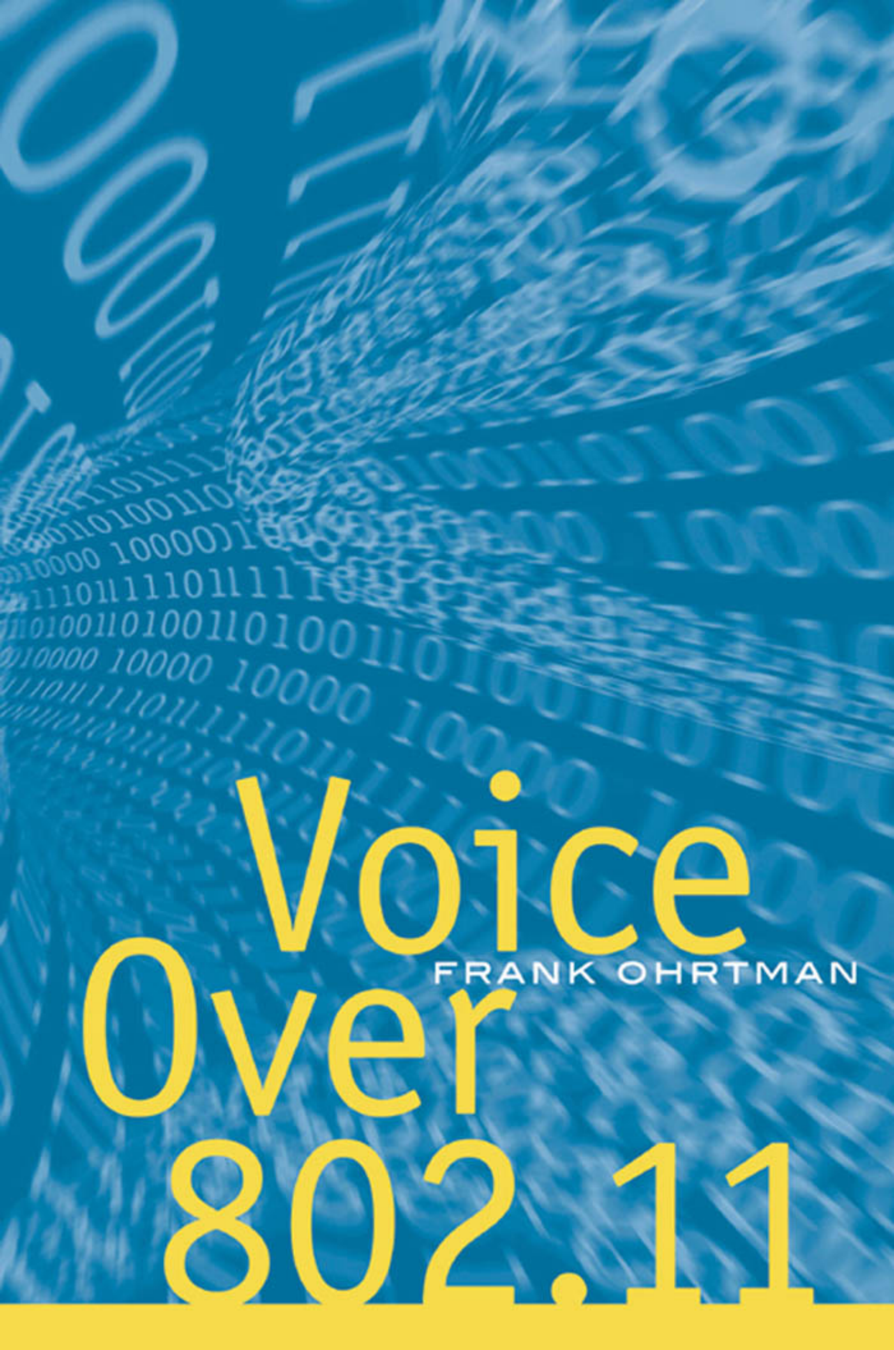


Voice Over 802.11

FRANK OHRTMAN

Voice over 802.11

For a listing of recent titles in the *Artech House Telecommunications Library*,
turn to the back of this book.

Voice over 802.11

Frank Ohrtman

Artech House
Boston • London
www.artechhouse.com

Library of Congress Cataloging-in-Publication Data
Ohrtman, Frank.
Voice over 802.11/Frank Ohrtman.
p. cm.
Includes bibliographical references and index.
ISBN 1-58053-677-8 (alk. paper)
1. IEEE 802.11 (Standard) 2. Internet telephony. I. Title.
TK5105.5668.O34 2004
641.382'12—dc22 2004041074

British Library Cataloguing in Publication Data
Ohrtman, Frank
Voice over 802.11.—(Artech House telecommunications library)
1. Internet telephony 2. Wireless communication systems 3. Broadband communications systems 4. IEEE 802.11 (Standard)
I. Title
621.3'845
ISBN 1-58053-677-8

Cover design by Gary Ragaglia

685 Canton Street
Norwood, MA 02062

International Standard Book Number: 1-58053-677-8

10 9 8 7 6 5 4 3 2 1

To Michelle Otte

Contents

1

Overview of Vo802.11

An understanding of the *public switched telephone network* (PSTN) and how it is potentially going to be replaced is best grasped by understanding its three major components: access, switching, and transport. *Access* pertains to how a user accesses the network. *Switching* refers to how a call is "switched" or routed through the network, and *transport* describes how a call travels or is "transported" over the network.

Access

As mentioned, *access* refers to how the user "accesses" the telephone network. For most users, access is gained to the network via a telephone handset. Transmission is a diaphragm in the mouthpiece that converts the air pressure of voice into an analog electromagnetic wave for transmission to the switch. The earpiece performs this process in reverse.

The most sophisticated aspect of the handset is its *dual-tone multifrequency* (DTMF) function, which signals the switch by tones. The handset is usually connected to the central office, where the switch is located, via copper wire known as *twisted pair* because, in most cases, it consists of a twisted pair of copper wire. The stretch of copper wire or, in newer installations, fiber-optic cable, connects the telephone handset to the central office. Everything that runs between the subscriber and the central office is known as *outside plant.* Telephone equipment at the subscriber end is called *customer-premises equipment* (CPE).

The emergence of wireless broadband Internet technologies such as 802.11a/b potentially allows the copper wires that have traditionally tethered residential and small business markets to telephone companies to be bypassed.

By not having to use copper wire to reach a residence or business, a competing service provider avoids the expense of the copper wire infrastructure as well as the legal entanglements of right of way and other issues to deploy a service that can compete with that of the incumbent service provider.

A market has sprung up in voice technologies for 802.11a/b networks. Major telecommunications equipment vendors such as Motorola, Cisco, and Avaya have products aimed at voice-over-wireless data networks. The focus of these industries is currently in the enterprise *local-area network* (LAN) market, however, it is not a stretch of the imagination to expect these technologies to, step by step, take market share from incumbent telephone service providers. The Telecommunications Act of 1996 was intended to open access to those copper wires for competing telephone companies (also known as *competitive local exchange carriers* or CLECs). It failed to do so to a meaningful degree. Competition will most likely come *to* the local loop, not *in* the local loop.

Switching

The PSTN is a star network, in which every subscriber is connected to another via at least one if not many hubs known as *offices.* In those offices are switches. Very simply, there are local offices for local service connections and tandem offices for long-distance service connections. Local offices, better known as *central offices* or COs, use Class 5 switches and tandem offices use Class 4 switches.

The late 1990s marked the emergence of the commercial *Voice over Internet Protocol* (VoIP). VoIP used a technology known as *softswitch* to replace Class 4 and Class 5 switches. A softswitch is simply software hosted on a server connected to an IP network. Instead of costing tens of millions of dollars and occupying vast CO space in expensive metro locations, a softswitch can be hosted almost anywhere on a server the size of a small refrigerator. Softswitch platforms cost a fraction of a Class 5 switch. By not having to route voice traffic through the incumbent service providers' Class 5 or Class 4 switches, a competing service provider could enjoy a greatly lowered barrier to entry to the voice market. The Telecommunications Act of 1996 was supposed to open the incumbent telephone companies' switching infrastructures to competitors, but it failed to do so. A softswitch allows a new market entrant to bypass the incumbent's Class 5 switch.

Transport

The *Memorandum of Final Judgment* (MFJ) of 1984 opened long-distance networks to competition. The emergence of the *Internet Protocol* (IP) as a transport technology sparked a boom in the construction of IP backbones, which led to bandwidth glut," that is, an overabundance of capacity on those networks. Contrary to traditional telephone networks, all a VoIP service provider needed to offer long-distance service was a connection to an IP backbone.

Vo802.11: Bypassing the Local Loop

The emergence of *voice over 802.11* (Vo802.11) was made possible by simply moving VoIP over 802.11 as an access mechanism, thereby replacing the copper wires of the PSTN. Once the VoIP stream reaches the wired part of such a network (the access point), it is transported on an IP network (LAN, IP backbone). By being based on the IP, VoIP can be managed (switched) by a VoIP-specific switch, the softswitch discussed in the preceding section. Although the conversation may originate and be switched on an IP network, it is still possible to originate and terminate calls on the PSTN. This is made possible with the interface of a VoIP gateway between the IP network and the PSTN. This gateway, depending on the direction of the flow of the traffic, packetizes or depacketizes the voice traffic traveling between the two dissimilar networks.

In summary, it is now possible to completely bypass the PSTN. By supplanting the elements of the PSTN with IP-based technologies, it is now possible to completely replicate the PSTN function for function. Not only does this represent a replacement of the PSTN, it is also makes possible a myriad of new elements for such a function. Application servers that operate with softswitches allow for the rapid creation of new features that were either not possible with the circuit-switched PSTN or would have cost the service provider too much to justify deployment.

This thesis is not without opposition. A number of objections to the deployment of Vo802.11 remain. Those objections are focused on concerns that the two chief elements of Vo802.11, that is, VoIP and 802.11, have perceived weaknesses that prevent them from delivering the same levels of service as the PSTN. After explaining the workings of the PSTN, 802.11, and VoIP, this book will overcome those objections.

In his book *The Innovator's Dilemma* [1], author Clayton Christensen describes what he terms *disruptive technology.* Initially, disruptive technology is "cheaper, simpler, smaller and more convenient to use." Eventually it matches the incumbent technology point for point and then ultimately triumphs, displacing the incumbent technology because the disruptive technology had a number of attributes of its own that the incumbent technology could not compete against. The following chapters will demonstrate how Vo802.11 is "cheaper, simpler, smaller and more convenient to use," while ultimately offering qualities that are superior to the incumbent technology.

Reference

[1] Christensen, C., *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, New York: HarperBusiness, 2000.

2

802.11: Alternative Access

What, technically speaking, is 802.11b and how does it relate to IEEE 802.11? This chapter covers the technology of transmitting data over the airwaves, the process of that transmission, and the topologies and components of wireless networks. Thousands of enterprises worldwide are "cutting the wires" to their LANs to enjoy greater productivity from their unwired workforce. The 802.11b technology also presents the potential to save money on infrastructure (wiring buildings for networks) and telecommunications services.

Because Vo802.11 is VoIP transmitted on 802.11, it is necessary to understand how this transmission medium functions. Just as voice has been transmitted over *asynchronous transfer mode* (ATM), frame relay, X.25, and the Internet Protocol, it can also be transmitted on 802.11. This chapter discusses how 802.11 works. From this, the reader will gain a better understanding of how 802.11 can be used to transmit voice.

How Does WiFi Work?

A networked desktop computer is connected to a larger network [LAN, *wide-area network* (WAN), Internet] via a network cable to a hub, router, or switch. The computer's network interface card sends zeros and ones down the cable by changing the voltage on the wires from +5V to –5V in a prearranged cadence. WiFi simply replaces these cables with small, low-powered two-way radios. Instead of changing voltage on a wire, it encodes the zeros and ones by laying an alternating radio signal over a constant existing signal, again in a prearranged cadence. The alternating signal encodes zeros and ones on the radio waves. The 802.11b specification allows for the wireless transmission of approximately

11 Mbps of raw data at distances up to a few hundred feet over the 2.4-GHz unlicensed band. The distance depends on impediments, materials, and line of sight.

The big "so what!?" of this technology is that it means PC users can install $40 PC cards in their laptops or PDAs and be connected just as well to the Internet or their corporate networks as if they were still tied to their desks and wall outlets by a physical wire. Enterprises have been quick to adopt this technology because (1) it is not constrained by the cost of wiring a building for voice and data, (2) it improves worker productivity by allowing mobility within a building or corporate campus, (3) it does not require right-of-way agreements to bring service to a business, (4) it is independent of distance to CO limitations, and (5) it is relatively free of federal, state, and local regulations.

A *wireless local-area network* (WLAN) installation usually uses one or more *access points* (AP), which are dedicated stand-alone hardware with typically more powerful antennas. Figure 2.1 illustrates a wireless LAN. In addition to servicing enterprise networks, 802.11b has become the most popular standard for public short-range networks, known as *hot spots*, which are 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 United States and internationally [1].

How Data Is Transmitted Via Wireless Technology

The 802.11 standard provides for two *radio-frequency* (RF) variations (as opposed to infrared) of the physical layer: *direct sequence spread spectrum* (DSSS) and *frequency hopping spread spectrum* (FHSS). Both of these were designed to

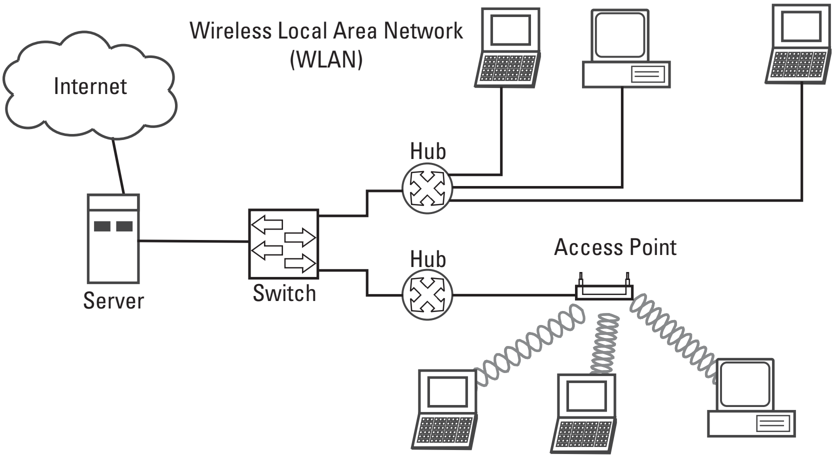


Figure 2.1 Wireless LAN on an enterprise network.

comply with FCC regulations (FCC 15.247) for operation in the 2.4-GHz band, which is an unlicensed spectrum. 802.11b uses DSSS.

DSSS systems use technology similar to that of *Global Positioning System* (GPS) satellites and some types of cell phones. Each information bit is combined with a longer *pseudorandom numerical* (PN) in the transmission process. The result is a high-speed digital stream, which is then modulated onto a carrier frequency using *differential phase-shift keying* (DPSK). DSSS works by taking a data stream of zeros and ones and modulating it with a second pattern, the chipping sequence. The sequence is also known as the *Barker code,* which is an 11-bit sequence (10110111000). The chipping or spreading code is used to generate a redundant bit pattern to be transmitted, and the resulting signal appears as wideband noise to the unintended receiver. One of the advantages of using spreading codes is that even if one or more of the bits in the chip are lost during transmission, statistical techniques embedded in the radio can recover the original data without the need for retransmission. The ratio between the data and width of the spreading code is called *processing gain.* It is 16 times the width of the spreading code and increases the number of possible patterns to 64,000 (2^{16}), thus reducing the chances of cracking the transmission.

The DSSS signaling technique divides the 2.4-GHz band into fourteen 22-MHz channels, of which 11 adjacent channels overlap partially and the remaining three do not overlap. Data are sent across one of these 22-MHz channels without hopping to other channels, causing noise on the given channel. To reduce the number of retransmissions and noise, chipping is used to convert each bit of user data into a series of redundant bit patterns called *chips.* The inherent redundancy of each chip, combined with spreading the signal across the 22-MHz channel, provides the error checking and correction functionality to recover the data. Spread spectrum products are often interoperable because many are based on the IEEE 802.11 standard for wireless networks. DSSS is used primarily in interbuilding LANs, because its properties are fast and far reaching [2].

At the receiver, a matched filter correlator is used to remove the PN sequence and recover the original data stream. At a data rate of 11 Mbps, DSSS receivers use different PN codes and a bank of correlators to recover the transmitted data stream. The high rate modulation method is called *complementary code keying* (CCK).

The PN sequence spreads the transmitted bandwidth of the resulting signal (hence, the term *spread spectrum*) and reduces peak power. Total power remains unchanged. On receipt, the signal is correlated with the same PN sequence to reject narrowband interference and recover the original binary data. Regardless of whether the data rate is 1, 2, or 5.5 of 11 Mbps, the channel bandwidth is about 20 MHz for DSSS systems.

The Significance of Spread Spectrum Radio

One of the basic technologies underlying the IEEE 802.11 series of standards is spread spectrum radio. The fundamental concept of spread spectrum radio is that it uses a wider frequency bandwidth than that needed by the information that is transmitted. Using extra bandwidth would seem to be wasteful, but it actually results in several benefits, including reduced vulnerability to jamming, less susceptibility to interference, and coexistence with narrowband transmissions. There are several spread spectrum techniques including time hopping, frequency modulation, FHSS, DSSS, and hybrids of these.

FHSS and DSSS are not modulation techniques, but simply methods of distributing a radio signal across bandwidth. In addition to spreading the signal across a frequency band, spread spectrum systems modulate the signal. *Modulation* is the variation of a radio signal to convey information. The base signal is called the *carrier*. The variation may be based on the strength (amplitude modulation), frequency, or phase (frequency offset) of the signal. The modulation technique directly affects the data rate. Higher data rate modulations are generally more complex and expensive to implement. Modulations resulting in higher data rates pack more information in the same bandwidth. Small disruptions in the signal cause the degradation of more data. This means that the signal must have a higher *signal-to-noise ratio* (SNR) at the receiver to be effectively processed. Because a radio signal is stronger the closer it is to the source, the SNR decreases with distance. This is why higher speed systems have less range. Examples of modulation techniques used in the IEEE 802.11 series of specifications are *binary phase-shift keying* (BPSK), *quadrature phase-shift keying* (QPSK), *Gaussian frequency-shift keying* (GFSK), and CCK.

802.11 Variants

In 1997 the *Institute of Electrical and Electronics Engineers* (IEEE) adopted IEEE Standard 802.11-1997, the first WLAN standard. This standard defines the *media access control* (MAC) and *physical* (PHY) layers for a LAN with wireless connectivity. It addresses local-area networking, in which the connected devices communicate over the air to other devices that are within proximity to each other. This is illustrated in Figure 2.2.

The *Wireless Ethernet Compatibility Alliance* (WECA) industry group certifies its members' equipment as conforming to the 802.11b standard and allows compliant hardware to be certified as WiFi compatible. This is an attempt at a guarantee of intercompatibility between hundreds of vendors and thousands of devices. Table 2.1 lists the variants of 802.11 and provides an overview of the relationship between 802.11b with other 802.11 variants.

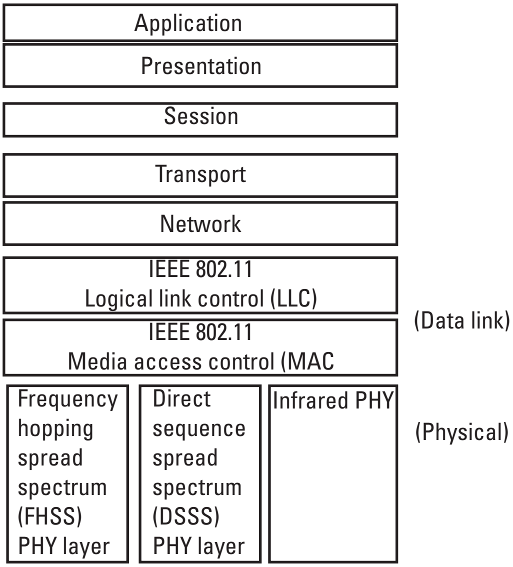

Application
Presentation
Session
Transport
Network
IEEE 802.11
Logical link control (LLC)
IEEE 802.11
Media access control (MAC
(Data link)
Frequency hopping spread spectrum (FHSS) PHY layer
Direct sequence spread spectrum (DSSS) PHY layer
Infrared PHY
(Physical)

Figure 2.2 IEEE 802.11 standards mapped to the *Open Systems Interconnect* (OSI) reference model.

FHSS (802.11a)

Spread spectrum radio techniques originated in the U.S. military in the 1940s. The unlikely copatent holders on spread spectrum technology are the actress Hedy Lamar and musician George Antheil. Lamar had been married to a German arms dealer and fled Germany as the Nazis came to power. One of Antheil's techniques involved the use of player pianos. These two facts came together to create one of the twentieth century's most influential radio technologies.

The military had started to use radio as a remote control mechanism for torpedoes, but this technique suffered from a vulnerability to jamming. Aware of this, Lamar suggested to Antheil that the radio signal should be distributed randomly over time across a series of frequencies. The transmission on each frequency would be brief and make the aggregate less susceptible to interruption or jamming. The problem was synchronizing the transmitter and receiver to the frequency being used at any point in time. Antheil used his musical expertise to design a synchronization mechanism using perforated paper rolls like those found in player in player pianos.

Lamar and Antheil were awarded U.S. patent number 2,292,387 and gave the rights to the Navy in support of the war effort. Although the Navy did not deploy the technology, engineers at Sylvania Electronic Systems applied electronic synchronization techniques to the concept in the late 1950s. The U.S.

Table 2.1
IEEE 802.11 Variants

802.11 Variant	Description
802.11a	Created a standard for WLAN operations in the 5-GHz band, with data rates of up to 54 Mbps. Published in 1999. Products based on this standard were released in 2003.
802.11b	Created a standard (also known as WiFi) for WLAN operations in the 2.4-GHz band, with data rates of up to 11 Mbps. Published in 1999. Products based on 802.11b include public space Internet kiosks, WLAN services such as Wayport, and wireless home networking products such as the Macintosh AirPort.
802.11c	Provided documentation of 802.11-specific MAC procedures to the *International Organization for Standardization/International Electrotechnical Commission* (ISO/IEC) 10038 (IEEE 802.1D) standard. Work completed.
802.11d	Publishing definitions and requirements to allow the 802.11 standard to operate in countries not currently served by the standard.
802.11e	Attempting to enhance the 802.11 MAC to increase the quality of service possible. Improvement in capabilities and efficiency are planned to allow applications such as voice, video, or audio transport over 802.11 wireless networks.
802.11f	Developing recommended practices for implementing the 802.11 concepts of access points and distribution systems. The purpose is to increase compatibility between AP devices from different vendors.
802.11g	Developing a higher speed PHY extension to the 802.11b standard, while maintaining backward compatibility with current 802.11b devices. The target data rate for the project is at least 20 Mbps.
802.11h	Enhancing the 802.11 MAC and 802.11a PHY to provide network management and control extensions for spectrum and transmitting power management in the 5-GHz band. This is will allow regulatory acceptance of the standard in some European countries.
802.11i	Enhancing the security and authentication mechanisms of the 802.11 standard.
802.1x	Also aimed at enhancing security of 802.11b.

Source: [3].

military began using these systems for secure communications in the early 1960s. The spread spectrum technique spawned from the work of Hedy Lamar and George Antheil is what we now call FHSS.

Local authorities also regulate the hopping rate. In North America, the hopping rate is set at 2.5 hops per second with each transmission occupying a channel for less than 400 milliseconds. Channel occupancy is also called *dwell time*. In 2001, the *Federal Communications Commission* (FCC) proposed to amend its Part 15 rules to allow adaptive hopping techniques to be used. This rulemaking is designed to reduce interference with other systems operating at the 2.4-GHz frequencies. Studies have shown that up to 13 IEEE 802.11 FHSS systems can be colocated before frequency channel collisions become an issue [4, pp. 124–126].

DSSS

DSSS systems mix high-speed bit patterns with the information being sent to spread the RF carrier. Each bit of information has a redundant bit pattern associated with it, effectively spreading the signal over a wider bandwidth. These bit patterns vary in length and the rate at which they are mixed into the RF carrier. They are called *chips* or *chipping codes* and vary in length from as small as 11 bits to extremely long sequences. The speed at which they are transmitted is called the *chipping rate.* To an observer, these sequences appear to be noise and are also called *pseudorandom noise codes* (Pncodes). Pncodes are usually introduced into the signal through the use of hardware-based shift registers, and the techniques used to introduce them are divided into several groups including Barker codes, Gold codes, M-sequences, and Kasami codes.

These spreading codes allow the use of statistical recovery methods to repair damaged transmissions. Another side effect of spreading the signal is lower spectral density—that is, the same amount of signal power is distributed over more bandwidth. The effect of a less spectrally dense signal is that it is less likely to interfere with spectrally dense narrowband signals. Narrowband signals are also less likely to interfere with a DSSS signal because the narrowband signal is spread as part of the correlation function at the receiver.

The frequency channel in IEEE 802.11 DSSS is 22-MHz wide. This means that it supports three nonoverlapping channels for operation. This is why only three IEEE 802.11b DSSS systems can be colocated.

In addition to spreading the signal, modulation techniques are used to encode the data signal through predictable variations of the radio signal. IEEE 802.11 specifies two types of DPSK modulation for DSSS systems. The first is BPSK and the second is QPSK. *Phase-shift keying* (PSK), as the name implies, detects the phase of the radio signal. BPSK detects 180-degree inversion of the signal, representing a binary 0 or 1. This method has an effective data rate of 1 Mbps. QPSK detects 90-degree phase shifts. This doubles the data rate to 2 Mbps. IEEE 802.11b adds CCK and *packet binary convolutional coding* (PBCC), which provide data rates up to 11 Mbps [4, pp. 126–128].

Orthogonal Frequency-Division Multiplexing

IEEE 802.11a (5 GHz) uses *orthogonal frequency-division multiplexing* (OFDM) as its frequency management technique and adds several versions of *quadrature amplitude modulation* (QAM) in support of data rates up to 54 Mbps. Bell

Laboratories patented OFDM in 1970 and it is based on a mathematical process called the *fast Fourier transform* (FFT). FFT enables 52 channels to overlap without losing their individuality or orthogonality. Overlapping channels is a more efficient use of the spectrum and enables them to be processed at the receiver more efficiently. IEEE 802.11a OFDM divides the carrier frequency into 52 low-speed subcarriers. Forty-eight of these carriers are used for data and four are used as pilot carriers. The pilot subcarriers allow frequency alignment at the receiver.

One of the biggest advantages of OFDM is its resistance to multipath interference and delay spread. Multipath is caused when radio waves reflect and pass through objects in the environment. Radio waves are attenuated or weakened in a wide range depending on the object's materials. Some materials (such as metal) are opaque to radio transmissions. You can imagine that a cluttered environment would be very different from an open warehouse environment for radio wave transmission and reception. This environmental variability is why it is so hard to estimate the range and data rate of an IEEE 802.11 system. Because of reflections and attenuation, a single transmission can be at different signal strengths and from different directions depending on the types of materials it encounters. This is the *multipath* aspect of interference. IEEE 802.11a supports data rates from 6 to 54 Mbps. It utilizes BPSK, QPSK, and QAM to achieve the various data rates.

Delay spread is associated with multipath. Because the signal is traveling over different paths to the receiver, the signal arrives at different times. This is delay spread. As the transmission rate increases, the likelihood of interference from previously transmitted signals increases. Multipath and delay spread are not much of an issue at data rates less than 3 or 4 Mbps, but some sort of mechanism is required as rates increase to mitigate the effect of multipath and delay spread. In IEEE 802.11b, it is CCK modulation. In 802.11a, it is OFDM. The IEEE 802.11g specification also uses OFDM as its frequency management mechanism [4, p. 131].

The adoption and refinement of advanced semiconductor materials and radio transmission technologies for IEEE 802.11 provides a solid basis for the implementation of higher-level functions. The next step up the protocol ladder is the definition of access functionality. Without structured access, the physical medium would be unusable [4, pp. 99, 129–131].

OFDM is not a new technique. Most of the fundamental work was done in the late 1960s, and U.S. patent number 3,488,445 was issued in January 1970. Recent *Digital Subscriber Line* (DSL) work [*high bit-rate DSL* (HDSL), *very high bit-rate DSL* (VDSL), and *asymmetric DSL* (ADSL)] and wireless data applications have rekindled interest in OFDM, especially now that better signal processing techniques make it more practical. OFDM does, however, differ from other emerging encoding techniques such as *code-division multiple access*

(CDMA) in its approach. CDMA uses complex mathematical transforms to put multiple transmissions onto a single carrier; OFDM encodes a single transmission into multiple subcarriers. The mathematics underlying the code division in CDMA is far more complicated than in OFDM. OFDM devices use one wide-frequency channel by breaking it up into several component subchannels. Each subchannel is used to transmit data. All the low subchannels are then multiplexed into one "ast" combined channel.

Carrier Multiplexing

When network managers solicit user input on network build-outs, one of the most common demands is for more speed. The hunger for increased data transmission has driven a host of researchers to search for ways to increase the speed of their technologies. OFDM takes a qualitatively similar approach to multilink PPP: When one link is not enough, use several in parallel.

OFDM is closely related to plain old frequency division multiplexing (FDM). Both "divide" the available bandwidth into slices called *carriers* or *subcarriers* and make those carriers available as distinct channels for data transmission. OFDM boosts throughput by using several subcarriers in parallel and multiplexing data over the set of subcarriers.

Traditional FDM was widely used by first-generation mobile telephones as a method for radio channel allocation. Each user was given an exclusive channel, and guard bands were used to ensure that spectral leakage from one user did not cause problems for users of adjacent channels [5, p. 199].

MAC Concepts and Architecture

The IEEE 802.11 MAC layer is common to all IEEE 802.11 PHY layers and specifies the functions and protocols required for control and access. The MAC layer is responsible for managing data transfer from higher level functions to the physical media. Figure 2.2, earlier in this chapter, illustrates this relationship to the OSI model.

MAC Layer Services

Devices using the IEEE 802.11 PHY and MAC as part of a WLAN are called *stations.* Stations can be endpoints or access points. APs are stations that act as part of the *distribution system* (DS) and facilitate the distribution of data between endpoints. The MAC provides nine logical services: authentication, deauthentication, association, disassociation, reassociation, distribution, integration, privacy, and data delivery. An AP uses all nine services. An endpoint uses authentication, deauthentication, privacy, and data delivery. Each service utilizes a set of messages with information elements that are pertinent to the services. These services are described in Table 2.2.

Power Management and Time Synchronization

In addition to *carrier-sense multiple-access /collision avoidance* (CSMA/CA) control frames (RTS, CTS, ACK, and contention polling), the MAC also provides control frames for power management and time synchronization. APs provide a time synchronization beacon to associated stations in an infrastructure *basic service set* (BSS). In an independent BSS, in which stations are operating as peers, an algorithm is defined that enables each station to reset its time when it receives a synchronization value greater than its current value. Stations entering a power-save mode may inform a PC through the frame control field of a message. The AP will then buffer transmissions to the station. A station is informed that it has buffered transmissions waiting when it wakes periodically to receive beacon frames. It can then request transmission. A station in active mode can receive frames at any time during a contention-free period. A station in power-save mode will periodically enter the active mode to receive beacons, broadcast, multicast, and buffered data frames [5, p. 128].

MAC Layer Architecture

As illustrated earlier in Figure 2.2, both the PHY and MAC layers are conceptually divided into management and data transfer capabilities. The PHY management capability is provided by the *PHY layer management entity* (PLME). The MAC management capability is provided by the *MAC layer management entity* (MLME). The PLME and the MLME exchange information about PHY medium capabilities through a *management information base* (MIB; see following paragraphs for more information). This is a database of physical characteristics such as possible transmission rates, power levels, and antenna types. Some of these characteristics are static and some can be changed by a management entity. These management functions support the main purpose of the MAC, which is to transfer data elements. These data elements originate in the *logical link control* (LLC) layer. Packages of data passed to the MAC from the LLC are called *MAC service data units* (Medusa). To transfer the Medusa to the PHY, the MAC uses messages (frames) containing functionality-related fields. There are three types of MAC frames: control, management, and data. One of these messages is called a *MAC protocol data unit* (MPDU). The MAC passes Medusa to the PHY layer through the *Physical Layer Convergence Protocol* (PLCP). The PLCP is responsible for translating Medusa into a format that is *physical medium dependent* (PMD). The PMD layer transfers the data onto the medium.

Table 2.2
IEEE 802.11 MAC Services and Agents

MAC Service	Definition	Station Type
Authentication	Because wireless LANs have limited physical security to prevent unauthorized access, 802.11 defines authentication services to control access to the WLAN. The goal of the authentication service is to provide access control equal to that of a wired LAN. The authentication service provides a mechanism for one station to identify another station. Without this proof of identity, the station is not allowed to use the WLAN for data delivery. All 802.11 stations, whether they are part of an independent BSS or *extended service set* (ESS) network, must use the authentication service prior to communicating with another station.	Endpoint and AP
Open system authentication	This is the default authentication method, which is a very simple, two-step process. First, the station wanting to authenticate with another station sends an authentication management frame containing the sending station's identity. The receiving station then sends back a frame alerting whether it recognizes the identity of the authenticating station.	
Shared key authentication	This type of authentication assumes that each station has received a secret shared key through a secure channel independent of the 802.11 network. Stations authenticate through shared knowledge of the secret key. Use of shared key authentication requires implementation of encryption via the *Wired Equivalent Privacy* (WEP) algorithm	
Deauthentication	Removes an existing authentication. The deauthentication service is used to eliminate a previously authorized user from any further use of the network. Once a station is deauthenticated, that station is no longer able to access the WLAN without performing the authentication function again. Deauthentication is a notification and cannot be refused. For example, when a station wishes to be removed from a BSS, it can send a deauthentication management frame to the associated access point to notify the AP of the removal from the network. An AP could also deauthenticate a station by sending a deauthentication frame to the station.	Endpoint and AP
Association	Maps a station to an access point and enables the AP to distribute data to and from the station The association service is used to make a logical connection between a mobile station and an AP. Each station must become associated with an AP before it is allowed to send data through the AP onto the distribution system. The connection is necessary in order for the distribution system to know where and how to deliver data to the mobile station. The mobile station invokes the association service once and only once, typically when the station enters the BSS. Each station can associate with only one AP, although an AP can associate with multiple stations.	AP

Table 2.2 (continued)

Mac Service	Description	Station Type
Disassociation	Breaks an existing association relationship. The disassociation service is used either to force a mobile station to eliminate an association with an access point or for a mobile station to inform an AP that it no longer requires the services of the DS. When a station becomes disassociated, it must begin a new association to communicate with an AP again. An AP may force a station or stations to disassociate because of resource restraints or if the access point is being shut down or removed from the network for a variety of reasons. When a mobile station is aware that it will no longer require the services of an AP, it may invoke the disassociation service to notify the access point that the logical connection to the services of the access point from this mobile station is no longer required. Stations should disassociate when they leave a network, although there is nothing in the architecture to ensure that this happens. Disassociation is a notification and can be invoked by either associated party. Neither party can refuse termination of the association.	AP
Reassociation	Transfers an association between APs. Reassociation enables a station to change its current association with an access point. The reassociation service is similar to the association service, with the exception that it includes information about the access point with which a mobile station has been previously associated. A mobile station will use the reassociation service repeatedly as it moves throughout the ESS, loses contact with the AP with which it is associated, and needs to become associated with a new access point. By using the reassociation service, a mobile station provides information to the AP with which it will be associated and information pertaining to the AP with which it will be disassociated. This allows the newly associated AP to contact the previously associated AP to obtain frames that may be waiting there for delivery to the mobile station as well as other information that may be relevant to the new association. The mobile station always initiates reassociation.	AP
Privacy	Prevents unauthorized viewing of data through use of the WEP algorithm. The privacy service of IEEE 802.11 is designed to provide an equivalent level of protection for data on the WLAN as that provided by a wired network with restricted physical access. This service protects that data only as they traverse the wireless medium. It is not designed to provide complete protection of data between applications running over a mixed network. With a wireless network, all stations and other devices can "hear" data traffic taking place within range on the network, seriously impacting the security level of a wireless link. IEEE 802.11 counters this problem by offering a privacy service option that raises the security of the	End-point and AP

Table 2.2 (continued)

MAC Service	Description	Station Type
	802.11 network to that of a wired network. The privacy service, applying to all data frames and some authentication management frames, is an encryption algorithm based on the 802.11.	
Distribution	Provides data transfer between stations through the DS. Distribution is the primary service used by an 802.11 station. A station uses the distribution service every time it sends MAC frames across the DS. The distribution service provides the distribution with only enough information to determine the proper destination BSS for the MAC frame. The three association services (association, reassociation, and disassociation) provide the necessary information for the distribution service to operate. Distribution within the DS does not necessarily involve any additional features outside of the association services, although a station must be associated with an access point for the distribution service to forward frames properly.	AP
Data delivery	Provides transfer of data between stations.	End-point and AP
Integration	Provides data transfer between the DS of an IEEE 802.11 LAN and a non-IEEE 802.11 LAN. The station providing this function is called a *portal.* The integration service connects the 802.11 WLAN to other LANs, including one or more wired LANs or 802.11 walls. A portal performs the integration service. The portal is an abstract architectural concept that typically resides in an AP, although it could be part of a separate network component entirely. The integration service translates 802.11 frames to frames that may traverse another network.	AP

Source: [4, 6].

MAC data transfer is controlled through two distinct coordination functions. The first is the *distributed coordination function* (DCF), which defines how users contend for the medium as peers. DCF data transfers are not time sensitive and delivery is asynchronous. The second is the *point coordination function* (PCF), which provides centralized traffic management for data transfers that are sensitive to delay and require contention-free access [4, pp. 134–135].

MIB

IEEE 802.11 contains extensive management functions to make the wireless connection appear much like a regular wired connection. The complexity of the additional management functions results in a complex management entity with dozens of variables. For ease of use, the variables have been organized into a management information base so that network managers can benefit from taking a structured view of the 802.11 parameters. The formal specification of the 802.11 MIB is Annex D of the 802.11 specification. The 802.11 MIB was designed by the 802.11 working group [5, p. 383].

DCF

The distributed coordination function defines how the medium is shared among members of the wireless network. It provides mechanisms for negotiating access to the wireless medium as well as mechanisms for reliable data delivery. One of the fundamental differences between wired and wireless media is that it is difficult to detect and manage data collisions on wireless media. The primary reason for this is that stations in a radio network are not guaranteed to hear every other station's transmissions. This is typically the case when an AP is used in IEEE 802.11's infrastructure BSS and is called the *hidden-node problem.*

PCF

The *point coordination function* (PCF) polls associated stations and manages frame transmissions on their behalf. A station performing PCF traffic management is called a *point coordinator* (PC). The PCF is an optional capability that provides connection-oriented services for delay-sensitive traffic. The PCF is more complex to implement, but it provides a moderate level of priority frame delivery for time-sensitive transmissions.

The PC uses beacon signals to broadcast duration for a contention-free period to all associated stations. This causes them to update their *network allocation vector* (NAV) and wait for the duration of the contention-free period. In addition, stations must await the *PCF interframe space* (PIFS) interval to further decrease the possibility of data collisions. The transmission of the additional polling and ACK messages required by the PCF is optimized through piggybacking multiple messages in a single transmission. For example, the PC may append both *acknowledgments* (Ax) of previous transmissions and polling messages for new traffic to a data frame. This enables the transmission to avoid waiting the interframe interval specified for individual frame transmissions [4, pp. 140–141].

The basic access method for 802.11 is the DCF, which uses CSMA/CA. This requires each station to listen for other users. If the channel is idle, the station may transmit. If the station is busy, it waits until transmission stops and then enters into a random back-off procedure (Figure 2.3). This prevents multiple stations from seizing the medium immediately after completion of the preceding transmission.

Packet reception in DCF requires acknowledgment as shown in Figure 2.3. The period between completion of packet transmission and start of the ACK frame is one *short interframe space* (SIFS). ACK frames have a higher priority than other traffic. Fast acknowledgment is one of the salient features of the 802.11 standard, because it requires ACKs to be handled at the MAC sublayer.

Transmissions other than ACKs must wait at least one *DCF interframe space* (DIFS) before transmitting data. If a transmitter senses a busy medium, it determines a random back-off period by setting an internal timer to an integer number of slot times. On expiration of a DIFS, the timer begins to decrement. If the time reaches zero, the station may begin transmission. If the channel is seized by another station before the timer reaches zero, the timer setting is retained at the decremented value for subsequent transmission. The method described above relies on the physical carrier sense. The underlying assumption is that every station can "hear" all other stations [7].

IEEE 802.11 Architecture

IEEE 802.11 supports three basic topologies for WLANs: the *independent basic service set* (IBSS), the BSS, and the ESS. All three configurations are supported by the MAC layer implementation.

The 802.11 standard defines two modes: *ad hoc/IBSS* and *infrastructure* mode. Logically, an ad hoc configuration (Figure 2.4) is analogous to a peer-to-peer office network in which no single node is required to function as a server. IBSS WLANs include a number of nodes or wireless stations that communicate directly with one another on an ad hoc, peer-to-peer basis, building a full-mesh or partial-mesh topology. Generally, ad hoc implementations cover a limited area and are not connected to a larger network.

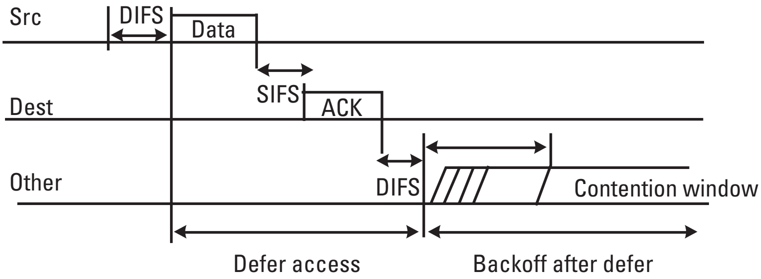


Figure 2.3 CSMA/CA back-off algorithm.

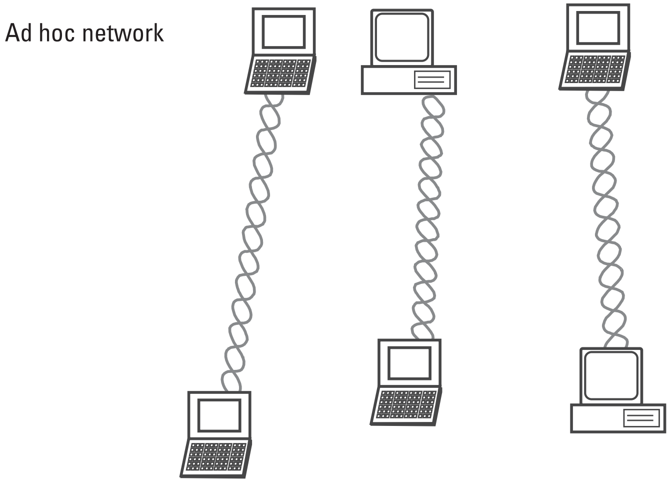


Figure 2.4 Wireless ad hoc network.

Using *infrastructure* mode, the wireless network consists of at least one AP connected to the wired network infrastructure and a set of wireless end stations. This configuration is called a *basic service set* (Figure 2.5). Because most corporate WLANs require access to the wired LAN for services (file servers, printers, Internet links), they will operate in infrastructure mode and rely on an AP that acts as the logical server for a single WLAN cell or channel. Communications between two nodes, A and B, actually flow from node A to the AP and then from the AP to node B. The AP is necessary to perform a bridging function and

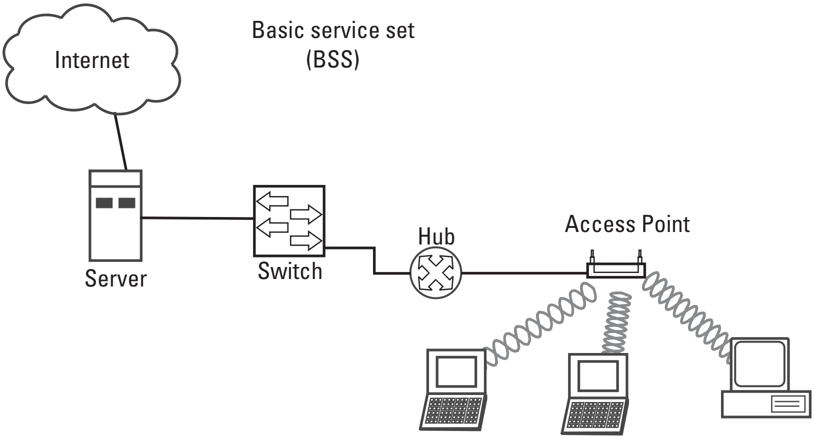


Figure 2.5 Wireless BSS.

connect multiple WLAN cells or channels, and to connect WLAN cells to a wired enterprise LAN.

An *extended service set* is a set of two or more BSSs forming a single subnetwork (Figure 2.6). ESS configurations consist of multiple BSS cells that can be linked by either wired or wireless backbones. IEEE 802.11 supports ESS configurations in which multiple cells use the same channel, and use different channels to boost aggregate throughput.

IEEE 802.11 Components

IEEE 802.11 defines two pieces of equipment, a wireless *station*, which is usually a PC equipped with a wireless *network interface card* (NIC) and an access point, which acts as a bridge between the wireless and wired networks. An AP usually consists of a radio, a wired network interface (802.3, for example), and bridging software conforming to the 802.11d bridging standard. The AP acts as the base station for the wireless network, aggregating access for multiple wireless stations onto the wired network. Wireless end stations can be 802.11 PC card, PCI, or ISA NICs, or embedded solutions in non-PC clients (such as an 802.11-based telephone handset).

An 802.11 WLAN is based on a cellular architecture. Each cell (BSS) is connected to the base station or AP. All APs are connected to a DS, which is similar to a backbone, usually Ethernet or wireless. All mentioned components appear as an 802 system for the upper layers of OSI and are known as the *ESS.*

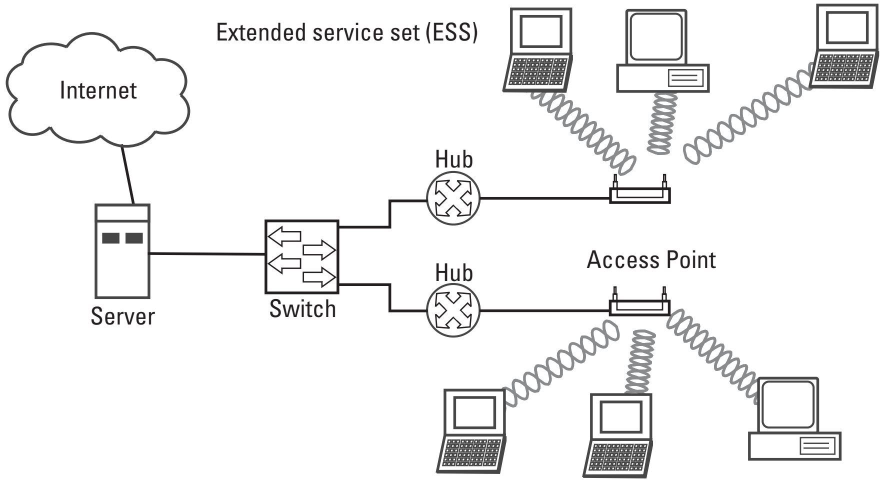


Figure 2.6 802.11 ESS.

The 802.11 standard does not constrain the composition of the DS; therefore, it may be 802 compliant or nonstandard. If data frames need transmission to and from a non-IEEE 802.11 LAN, then these frames, as defined by the 802.11 standard, enter and exit through a logical point called a *portal.* The portal provides logical integration between existing wired LANs and 802.11 LANs.

When the distribution system is constructed with 802-type components, such as 802.3 (Ethernet) or 802.5 (token ring), then the portal and the access point are the same, acting as a *translation bridge.* The 802.11 standard defines the distribution system as an element that interconnects BSSs within the ESS via access points. The distribution system supports the 802.11 mobility types by providing logical services necessary to handle address-to-destination mapping and seamless integration of multiple BSSs. An access point is an addressable station, providing an interface to the distribution system for stations located within various BSSs. The independent BSS and ESS networks are transparent to the LLC layer [2].

Mobility

Mobility of wireless stations may be the most important feature of a wireless LAN. The chief motivation of deploying a WLAN is to enable stations to move about freely from location to location either within a specific WLAN or between different WLAN "segments."

For compatibility purposes, the 802.11 MAC must appear to the upper layers of the network as a "standard" 802 LAN. The 802.11 MAC layer is forced to handle station mobility in a fashion that is transparent to the upper layers of the 802 LAN stack. This forces functionality into the 802.11 MAC layer that is typically handled by upper layers in the OSI model [6].

To understand this design restriction, it is important first to appreciate the difference between true mobility and mere portability. Portability certainly results in a net productivity gain because users can access information resources wherever it is convenient to do so. At the core, however, portability removes only the physical barriers to connectivity. It is easy to carry a laptop between several locations, so people do. But portability does not change the ritual of connecting to networks at each new location. It is still necessary to physically connect to the network and reestablish network connections, and network connections cannot be used while the device is being moved.

Mobility removes further barriers, most of which are based on the logical network architecture. Network connections stay active even while the device is in motion. This is critical for tasks requiring persistent, long-lived connections, which may be found in database applications.

IEEE 802.11 is implemented at the link layer and provides link-layer mobility. IP does not allow this. The 802.11 hosts can move within the last network freely, but IP, as it is currently deployed, provides no way to move across subnet boundaries. To the IP-based hosts of the outside world, the *virtual private network* (VPN) access control boxes are the last hop routers. To access an 802.11 wireless station with an IP address on the wireless network, it is possible to simply go through the IP router to the target network regardless of whether a wireless station is connected to the first or third AP. The target network is reachable through the last hop router. As far as the outside world can tell, the wireless station might as well be a workstation connected to an Ethernet.

A second requirement for mobility is that the IP address does not change when connecting to any of the access points. New IP addresses interrupt open connections. If a wireless station connects to the first AP, it must keep the same address when it connects to the third AP.

A corollary to the second requirement is that all of the wireless stations must be on the same IP subnet. As long as a station stays on the same IP subnet, it does not need to reinitialize its networking stack and can keep its TCP connections open. If it leaves the subnet, though, it needs to get a new IP address and reestablish any open connections. Multiple subnets are not forbidden, but if you have different IP subnets, seamless mobility between subnets is not possible.

The "single IP subnet backbone" restriction is a reflection on the technology deployed within most organizations. Mobile IP was standardized in late 1996 in RFC 2002, but it has yet to see widespread deployment. Until Mobile IP can be deployed, network designers must live within the limitations of IP and design networks based on fixed locations for IP addresses. A backbone network may be physically large, but it is fundamentally constrained by the requirement that all access points connect directly to the backbone router (and each other) at the link layer [5, pp. 295–296].

Conclusion

Some may argue that Morse code and the telegraph was the first technology that transmitted data via the airwaves (dots and dashes versus ones and zeros). The ability to transmit data over the airwaves presents some exciting opportunities for business networks. Businesses worldwide have made the switch from wired to wireless in order to save money and increase employee productivity.

IEEE 802.11b is a subvariant of 802.11, which is a standard that digresses slightly from the OSI model in that it provides a standard for wireless data transmission. To do this, the standard defines the MAC and PHY layers of the OSI model for use of DSSS (for 802.11b). The MAC layer is responsible for managing data transfer from higher level functions to the PHY media. This standard details how data are modulated for transmission and correlated at the receiving end. The topology of wireless networks is fairly simple. In a BSS, an AP is connected to an existing LAN from which wireless stations can access the network. An ESS extends this topology to expand the network. Using an ad hoc topology, stations (PCs) can communicate directly with one another. Mobility measures permit wireless users to access the wireless network from any point on the network and maintain their connection regardless of where they may roam on the network.

What is exciting about 802.11 is that it allows the transmission of voice over a unlicensed spectrum. That is, for the cost of a radio and antenna, a service provider can offer voice services similar to that of a telephone or cell phone company and avoid the expense of copper wires, right-of-way issues, and wireless cell phone spectrum. In short, 802.11 is an enabling technology that allows the local telephone companies to be bypassed.

References

[1] WiFi Consulting, "OK, What Is WiFi?" white paper, http://www.wificonsulting.com/WiFi101/wifi101.htm.

[2] Nedeltchev, P., "WLANS and the 802.11 Standard," Cisco Systems white paper, March 2001, p. 7, http://wwwin.cisco.com/cct/data/itm/wan/sdlc/wtsdllca.htm.

[3] WAVE Report, "IEEE 802.11 Standard Tutorial," http://www.wave-report.com/tutorials/ieee80211.htm, November 29, 2001.

[4] LaRocca, J., and R. LaRocca, *802.11 Demystified*, New York: McGraw-Hill, 2002.

[5] Gast, M., *802.11 Wireless Networks: The Definitive Guide,* Sebastopol, CA: O'Reilly & Associates, 2002.

[6] Intelligraphics, "Introduction to IEEE 802.11," white paper, http://www.intelligraphics.com/articles/80211_article.html.

[7] Zyren, J., and A. Petrick, "IEEE 80211 Tutorial," Wireless Ethernet white paper, p. 5, http://www.wirelessethernet.org/downloads/IEEE_80211_Primer.pdf.

3

Voice over Internet Protocol

What Is VoIP?

Voice over 802.11 is voice over IP used to transport voice over wireless Ethernet. Many books about VoIP are available, so this book does not attempt to cover it in detail. Instead, this chapter briefly outlines how VoIP functions. The intent of this chapter is to give the reader a basic understanding of VoIP before discussing Vo802.11.

Origins

In November 1988, Republic Telcom (yes, one "e") of Boulder, Colorado, received U.S. patent number 4,782,485 for a "multiplexed digital packet telephone system." The plaque from the Patent and Trademark Office describes it as follows:

> A method for communicating speech signals from a first location to a second location over a digital communication medium comprising the steps of: providing a speech signal of predetermined bandwidth in analog signal format at said first location; periodically sampling said speech signal at a predetermined sampling rate to provide a succession of analog signal samples; representing said analog signal samples in a digital format thereby providing a succession of binary digital samples; dividing said succession of binary digital samples into groups of binary digital samples arranged in a temporal sequence; transforming at least two of said groups of binary digital samples into corresponding frames of digital compression.

Republic and its acquiring company, Netrix Corporation, applied this voice-over-data technology to the data technologies of the times (X.25 and frame relay) until 1998 when Netrix and other competitors introduced VoIP onto their existing voice-over-data gateways. Although attempts at Internet telephony had been made from a software-only perspective, commercial applications were limited to using voice-over-data gateways that could interface the PSTN to data networks. Voice-over-data applications were popular in enterprise networks with offices spread across the globe (eliminated international interoffice long-distance bills), offices where no PSTN existed (installations for mining and oil companies), and for long-distance bypass (legitimate and illegitimate).

The popularity and applications of VoIP continued to grow. VoIP accounted for 6% of all international long-distance traffic in 2001 [1]. Six percent may not seem like an exciting number, but consider that a mere 3 years passed from when the technology was introduced to its capturing 6% of a trillion-dollar, 100-year-old industry—and it is clear that VoIP will continue to capture more market share. As VoIP migrates to 802.11 networks, it completes the bypass of the copper wires that tether residences and small business to incumbent telcos.

How Does VoIP Work?

Softswitch is increasingly considered to be almost synonymous with VoIP. However, it also works with TDM and ATM networks. The first process in an IP voice system is the digitization of the speaker's voice. The next step (and the first step when the user is on a handset connected to a gateway using a digital PSTN connection) is typically the suppression of unwanted signals and compression of the voice signal. This has two stages. First, the system examines the recently digitized information to determine if it contains a voice signal or only ambient noise and discards any packets that do not contain speech. Second, complex algorithms are employed to reduce the amount of information that must be sent to the other party. Sophisticated codecs enable noise suppression and compression of voice streams. Compression algorithms include G.723, G.728, and G.729.

Following compression, voice must be packetized and VoIP protocols added. Some storage of data occurs during the process of collecting voice data, since the transmitter must wait for a certain amount of voice data to be collected before it is combined to form a packet and transmitted via the network. Protocols are added to the packet to facilitate its transmission across the network. For example, each packet will need to contain the address of its destination, a sequencing number in case the packets do not arrive in the proper order, and additional data for error checking. Because IP is a protocol designed to interconnect networks of varying kinds, substantially more processing is required than in smaller networks. The network addressing system can often be very complex, requiring a process of encapsulating one packet inside another and, and as data move along, repackaging, readdressing, and reassembling the data.

When each packet arrives at the destination computer, its sequencing is checked to place the packets in the proper order. A decompression algorithm is used to restore the data to their original form, and clock synchronization and delay-handling techniques are used to ensure proper spacing. Because data packets are transported via the network by a variety of routes, they do not arrive at their destination in order. To correct this, incoming packets are stored for a time in a jitter buffer to wait for late-arriving packets. The length of time in which data are held in the jitter buffer varies depending on the characteristics of the network.

In IP networks, a percentage of the packets can be lost or delayed, especially during periods of congestion. Also, some packets are discarded due to errors that occurred during transmission. Lost, delayed, and damaged packets result in substantial deterioration of voice quality. In conventional error correction techniques used in other protocols, incoming blocks of data containing errors are discarded, and the receiving computer requests the retransmission of the packet, thus the message that is finally delivered to the user is exactly the same as the message that originated. Because VoIP systems are time sensitive and cannot wait for retransmission, more sophisticated error detection and correction systems are used to create sound to fill in the gaps. This process stores a portion of the incoming speaker's voice and then, using a complex algorithm to approximate the contents of the missing packets, new sound information is created to enhance the communication. Thus, the sound heard by the receiver is not exactly the sound transmitted, but rather portions of it that have been created by the system to enhance the delivered sound [2].

Protocols Related to VoIP

The softswitch revolution was made possible by the emergence of voice over data and, more specifically, VoIP. Note that there are softswitch solutions that use *time-division multiplexing* (TDM) and ATM. However, the consensus in the industry is that the future is going to be an IP network, ultimately dictating a VoIP solution. Before outlining softswitch solutions, it is first necessary to understand VoIP. VoIP is best understood as a collection of the protocols that make up its mechanics. Those protocols are loosely analogous to the PSTN, which is broken down into three categories: access, switching, and transport. Simply put, three categories of protocols are relevant to VoIP: signaling, routing, and transport.

Signaling, which is roughly analogous to the switching function described in the preceding chapter, protocols (H.323 and SIP) set up the route for the media stream or conversation. Gateway control protocols such as MGCP and MEGACO (also signaling protocols) establish control and status in media and signaling gateways.

Routing [using the *User Datagram Protocol* (UDP) and *Transmission Control Protocol* (TCP)] and transporting the media stream (conversation) once the route of the media stream has been established are the function of routing and transport protocols. Routing protocols such as UDP and TCP could be compared to the switching function described in Chapter 2.

RTP is analogous to the transport function. The signaling and routing functions establish what route the media stream will take and the routing protocols deliver the bits, that is, the conversation.

Signaling Protocols

The process of setting up a VoIP call is roughly similar to that of a circuit-switched call made on the PSTN. A media gateway must be loaded with the parameters to allow proper media encoding and the use of telephony features. Inside the media gateway is an intelligent entity known as an *endpoint.* When the calling and called parties agree on how to communicate and the signaling criteria have been established, the media stream over which the packetized voice conversation will flow is established. Signaling establishes the virtual circuit over the network for that media stream. Signaling is independent of the media flow. It determines the type of media to be used in a call. Signaling is concurrent throughout the call. Currently, two types of signaling are popular in VoIP: H.323 and SIP [3].

H.323

H.323 is the *International Telecommunication Union* (ITU-T) recommendation for packet-based multimedia communication. H.323 was developed before the emergence of VoIP. Because it was not specifically designed for VoIP, it has faced a good deal of competition from a competing protocol, the *Session Initiation Protocol* (SIP), which was designed specifically for VoIP. However, it has enjoyed a first-mover advantage and there now exists a considerable number of installed H.323 VoIP networks.

H.323 is comprised of a number of subprotocols. It uses protocol H.225.0 for registration, admission, status, call signaling, and control. It also uses protocol H.245 for media description and control, terminal capability exchange, and general control of the logical channel carrying the media stream(s). Other protocols make up the complete H.323 specification, which presents a protocol stack for H.323 signaling and media transport. H.323 also defines a set of call control, channel setup, and codec specifications for transmitting real-time video and voice over networks that do not offer guaranteed service or quality of service. As a transport, H.323 uses the *Real-Time Transport Protocol* (RTP), an *Internet Engineering Task Force* (IETF) standard designed to handle the requirements of streaming real-time audio and video via the Internet [3, p. 9]. H.323 was the first VoIP protocol for interoperability among the early VoIP gateway/gatekeeper vendors. Unfortunately, the promise of interoperability among diverse vendors platforms did not materialize with the adoption of H.323.

The H.323 standard is a cornerstone technology for the transmission of real-time audio, video, and data communications over packet-based networks. It specifies the components, protocols, and procedures that provide multimedia communication over packet-based networks. Packet-based networks include IP-based (including the Internet) or *Internet packet exchange* (IPX)-based LANs, *enterprise networks* (ENs), *metropolitan-area networks* (MANs), and WANs. H.323 can be applied in a variety of mechanisms—audio only (IP telephony); audio and video (videotelephony); audio and data; and audio, video, and data. H.323 can also be applied to multipoint-multimedia communications. H.323 provides myriad services and, therefore, can be applied in a wide variety of areas—consumer, business, and entertainment applications.

SIP

SIP is a signaling protocol. It uses a text-based syntax similar to that of the *Hypertext Transfer Protocol* (HTTP), which is used in Web addresses. Programs that are designed for parsing of HTTP can be adapted easily for use with SIP. SIP addresses, known as SIP *uniform resource locators* (URLs) take the form of Web addresses. A Web address can be the equivalent of a telephone number in a SIP network. In addition, PSTN phone numbers can be incorporated into a SIP address for interfacing with the PSTN. An e-mail address is portable. Using the proxy concept, one can check his or her e-mail from any Internet-connected terminal in the world. Telephone numbers, simply put, are not portable. They ring at only one physical location. SIP offers a mobility function that can follow a subscriber to whichever phone he or she is nearest at any given time.

Like H.323, SIP handles the setup, modification, and teardown of multimedia sessions including voice. Although it works with most transport protocols, its optimal transport protocol is RTP. Figure 3.1 shows how SIP functions as a signaling protocol, with RTP as the transport protocol for a voice conversation. SIP was designed as a part of the IETF multimedia data and control architecture. It is designed to interwork with other IETF protocols such as the *Session Description Protocol* (SDP), RTP, and *Session Announcement Protocol* (SAP). It is described in the IETF's *Request for Comments* (RFC) 2543. Many in the VoIP

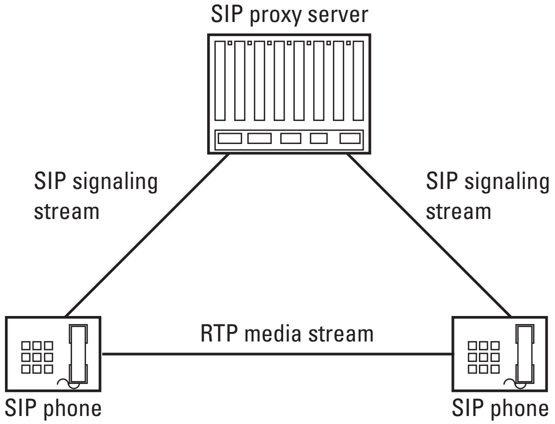


Figure 3.1 SIP is a signaling protocol and RTP transports the conversation. (*From:* [5]. © 2001 Artech House, Inc. Reprinted with permission.)

and softswitch industry believe that SIP will replace H.323 as the standard signaling protocol for VoIP.

SIP is part of the IETF standards process and is modeled on other Internet protocols such as *Simple Mail Transfer Protocol* (SMTP) and HTTP. It is used to establish, change, and tear down (end) calls between one or more users in an IP-based network. To provide telephony services, a number of different standards and protocols need to come together and work with each other specifically to ensure transport (RTP), provide signaling with the PSTN, guarantee voice quality (RSVP), provide directories (LDAP), authenticate users (RADIUS), and scale to meet anticipated growth curves.

What Is SIP?

SIP is focused on two classes of network entities: *clients* (also called *user agents*) and *servers.* VoIP calls on SIP originate at a client and terminate at a server. Types of clients in the technology currently available for SIP telephony would include a PC loaded with a telephony agent or a SIP telephone. Clients can also reside on the same platform as a server. For example, a PC on a corporate WAN might be the server for the SIP telephony application, but may also be used as a user's telephone (client) [4, p. 89].

SIP Architecture

SIP is a client/server architecture. The client in this architecture is the *user agent* (UA). The UA interacts with the user. It usually has an interface toward the user in the form of a personal computer or IP phone (SIP phone in this case). There are four types of SIP servers: (1) user agent server, (2) redirect servers, (3) proxy servers, and (4) a registrar. The type of SIP server used determines the architecture of the network.

SIP Calls Via UA Server

A UA server accepts SIP requests and contacts the user. A response from the user to the UA server results in a SIP response representing the user. A SIP device will function as both a user agent client and as a user agent server. As a user agent client the SIP device can initiate SIP requests. As a UA server, the device can receive and respond to SIP requests. As a stand-alone device, the UA can initiate and receive calls that empowers SIP to be used for peer-to-peer communications. Figure 3.2 describes the user agent server.

The function of SIP is best understood via the HTTP model on which it is based. SIP is a request/response protocol. A client is a SIP entity that generates a request. A server is a SIP entity that receives requests and returns responses. When a Web site is desired, the client generates a request by typing in the URL, for example, http://www.artechhouse.com. The server on which the Web site is hosted responds with Artech House's Web page. SIP uses the same procedure. The user agent sending the request is known as a *user agent client* (UAC) and the UA returning the response is the *user agent server* (UAS). This exchange is known as a *SIP transaction.*

Per Figure 3.2, a call initiates with the user agent of the calling party sending an INVITE command, caller@righthere.org, to the called party, callee@theotherside.com. The user agent for the caller has translated the name for

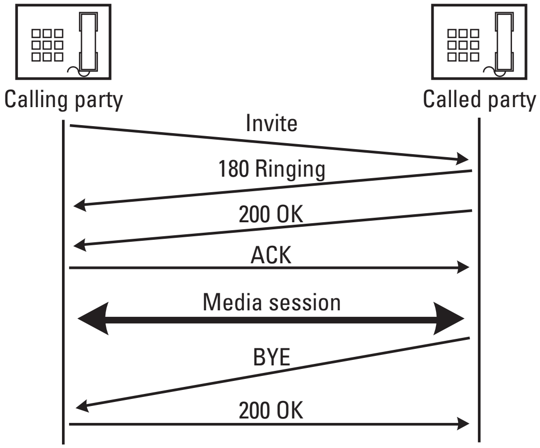


Figure 3.2 SIP UA server to UA server call. (*From:* [5]. © 2001 Artech House, Inc. Reprinted with permission.)

the called party into an IP address via a *domain name server* (DNS) query accessible via their own domain. The INVITE command is sent to a SIP UDP port and contains information such as media format and the From, To, and Via information. The TRYING informational response (180) from the calling party's call agent is analogous to the Q.931 CALL PROCEEDING message used in the PSTN, indicating the call is being routed. In the direct call model, a TRYING response is unlikely but for proxy and redirect models it is used to monitor call progress. SIP uses six methods of signaling as shown in Table 3.1. Additional methods are under consideration by the IETF at this time.

When the call arrives at the remote endpoint, the phone rings, and a new response is sent to that endpoint: RINGING (180). This is analogous to the Q.931 ALERTING message used in the PSTN. The time between when the user dials the last digit and the time RINGING is received by the caller is known as *post dial delay* (PDD) for SIP call setup. If a telephone number is involved in addressing the call, there is a need to translate the numbers into an IP address. Table 3.2 compares SIP and PSTN signals

When the called party answers the phone, a 200 response is sent to the calling party's user agent. The UA sends another request, ACK, acknowledging

Table 3.1
SIP Signaling Methods

SIP Method	Description
INVITE	First message sent by calling party to invite users to participate in a session. Contains information in the SIP header that identifies the calling party, call-ID, called party, and call and sequence number. Indicates a call is being initiated. When a multiple choice of SDP parameters is offered, the ones chosen are returned with the success (200) code in the response message.
ACK	Used to acknowledge the reception of a final response to an INVITE. A client originating an INVITE request issues an ACK request when it receives a final response for the INVITE, providing a three-way handshake.
OPTIONS	Query a server about its capabilities including which methods and which session description protocols it supports. This determines which media types a remote user supports before placing the call.
BYE	Used to abandon sessions. In two-party sessions, abandonment by one of the parties implies that the session is terminated. A return BYE from the other party is not necessary.
CANCEL	This method cancels pending transactions. The CANCEL method identifies the call via the call-ID, call sequence, and To and From values in the SIP header.
REGISTER	Users send REGISTER requests to inform a server about their current locations. SIP servers are colocated with SIP registrars. This enables the SIP server to find a user.

Source: [6].

Table 3.2
Comparison of SIP and PSTN Signals

SIP	PSTN
TRYING	Q.931 CALL PROCEEDING
RINGING	Q.931 ALERTING
ACK	Q.931 CONNECT
INVITE	Q.931 CONNECT

Source: [4].

the response to the INVITE request. At that moment, media begin to flow on the transport addresses of the two endpoints. The ACK may carry the final SDP parameters for media type and format provided by the receiving endpoint. The sequence of the INVITE and following ACK messages is similar to that of the Q.931 CONNECT message. ACKs do not require a response.

At this point in the call sequence, media flows over RTP, with RTCP providing the monitoring of the quality of the connection and its associated statistics. Next, as the name would imply, a BYE request from either party ends the call. Because all messages are sent via UDP, no further action is required.

SIP Calls Via Proxy Server

Proxy servers in the SIP sense are similar in function to proxy servers that "serve" a Web site (mail relay via SMTP) for a corporate LAN (Figure 3.3). A SIP client in this case would send a request to the proxy server, which would either handle it or pass it on to another proxy server that, after some translation, would take the call. The secondary servers would see the call as coming from the client. By virtue of receiving and sending requests, the proxy server is both a server and client. Proxy servers function well in call forwarding and follow-me services.

Proxy servers can be classified by the amount of state information they store in a session. SIP defines three types of proxy servers: call stateful, stateful, and stateless. *Call stateful proxies* need to be informed of all SIP transactions that occur during the session and are always in the path taken by SIP messages traveling between end users. These proxies store state information from the moment the session is established until the moment it ends.

A *stateful proxy* is sometimes called a *transaction stateful proxy* because the transaction is its sole concern. A stateful proxy stores the state related to a given transaction until the transaction concludes. It does not need to be in the path taken by the SIP messages for subsequent transactions. Forking proxies are good examples of stateful proxies. Forking proxies are used when a proxy server tries more than one location for the user, that is, it "forks" the invitation.

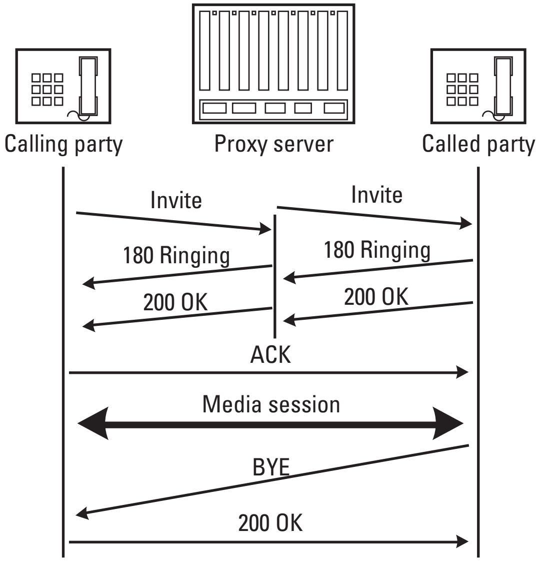


Figure 3.3 SIP call using a proxy server. (*From:* [5]. © 2001 Artech House, Inc. Reprinted with permission.)

Stateless proxies keep no state. They receive a request, forward it to the next hop, and immediately delete all states related to that request. When a stateless proxy receives a response, it determines routing based solely on the Via header and it does not maintain a state for it [6, pp. 126–129].

A SIP call using a proxy server is a little more complicated than the simple SIP call model described above. In this call, the caller is configured with the called party's SIP server. An INVITE is sent to the called party's SIP server with the called party's text address in the To field. The called party's server determines if the called party is registered in that server. If the called party is registered on that server, it then determines the called party's current location on the network. This is called the *mobility* feature.

Once the called party is located via the mobility feature, the proxy server generates an INVITE request with no alteration of the headers of the request except to add its own name in the Via field. Multiple servers may be involved in tracking down the called party.

Next the server must retain state information on the call. The server does this by correlating Cseq numbers, call-ID, and other elements of the headers as they pass through the proxy server. The server sends a TRYING message back to the calling party's agent. When the called party answers at the new location, a RINGING response is sent to the proxy server via the remote server (the called party's server). Both servers have Via entries in the response message to the calling party. Finally, ACK messages are exchanged, the call is established, and the media flow over RTP can begin. The call is terminated via a BYE request.

What makes the proxy server marketable is its user mobility feature. The called party can be logged in at multiple locations at once. This results in the proxy server generating the INVITE to all names on the list until the called party is found (preferably RINGING, but also TRYING or OK) [3, pp. 74–77].

SIP Calls Via Redirect Server

A redirect server accepts SIP requests, maps the destination address to zero or more new addresses, and returns the translated address to the originator of the request. After that, the originator of the request can send requests to the addresses returned by the redirect server. The redirect server originates no requests of its own. Redirect servers pose an alternative means of call forwarding and follow-me services. What differentiates the redirect server from a proxy server is that the originating client redirects the call. The redirect server provides the intelligence to enable the originating client to perform this task because the redirect server is no longer involved.

The redirect server call model is a mix of the two previously described call models. Here, a proxy model reverts to the direct call model once the called party is located. The redirect server returns a redirection response to the INVITE with code 301 or 302, indicating the called party is at the location listed in the Contact field of the message body. The calling party's media gateway controller closes its signaling with the redirect server and initiates another INVITE to the location returned in the redirect response. After that the call flow is that of the direct model. If the called party is registered at a number of locations, the redirect server will return a list of names (URI) to be contacted. The calling party can then contact those addresses directly.

Registrar

A registrar is a server that accepts SIP REGISTER requests. SIP includes the concept of user registration in which a user tells the network that she or he is available at a given address. This registration occurs by issuing a REGISTER request by the user to the registrar. A registrar is often combined with a proxy or redirect server. Practical implementations often combine the UAC and UAS with registrars with either proxy servers or redirection servers. This can result in a network having only user agents and redirection or proxy servers [7, pp. 167–168].

Location Servers

Location servers are not SIP entities, but are an important part of SIP architecture. A location server stores and returns possible locations for users. It can make use of information from registrars or from other databases. Most registrars upload location updates to a location server on receipt. SIP is not used between location servers and SIP servers. Some location servers use the *Lightweight Directory Access Protocol* (LDAP; see IETF RFC 1777) to communicate with SIP servers [7, p. 105].

Interworking with Other Multimedia Networks

The H.323 standard specifies four kinds of components, which, when networked together, provide the point-to-point and point-to-multipoint multimedia communication services: terminals, gateways, gatekeepers, and *multipoint control units* (MCUs)

Terminals

Used for real-time bidirectional multimedia communications, an H.323 terminal can either be a PC or a stand-alone device running H.323 and multimedia applications. It supports audio communications and can optionally support video or data communications. Because the basic service provided by an H.323 terminal is audio communications, an H.323 terminal plays a key role in IP-telephony services. The primary goal of H.323 is to interwork with other multimedia terminals. H.323 terminals are compatible with H.324 terminals on SCN and wireless networks, H.310 terminals on B-ISDN, H.320 terminals on ISDN, H.321 terminals on B-ISDN, and H.322 terminals on guaranteed *quality of service* (QoS) LANs. H.323 terminals may be used in multipoint conferences.

Gateways

A gateway connects two dissimilar networks. An H.323 gateway provides connectivity between an H.323 network and a non-H.323 network. For example, a gateway can connect and provide communication between an H.323 terminal and SCN networks (SCN networks include all switched telephony networks, for example, the PSTN). This connectivity of dissimilar networks is achieved by translating protocols for call setup and release, converting media formats between different networks, and transferring information between the networks connected by the gateway. A gateway is not required, however, for communication between two terminals on an H.323 network.

Gatekeepers

A gatekeeper can be considered the brain of the H.323 network. It is the focal point for all calls within the H.323 network. Although they are not required, gatekeepers provide important services such as addressing, authorization, and authentication of terminals and gateways; bandwidth management; accounting; billing; and charging. Gatekeepers may also provide call-routing services.

MCUs

MCUs provide support for conferences of three or more H.323 terminals. All terminals participating in the conference establish a connection with the MCU, which manages conference resources, negotiates among terminals to determine the audio or video *coder/decoder* (codec) to use, and may handle the media stream. Gatekeepers, gateways, and MCUs are logically separate components of the H.323 standard but can be implemented as a single physical device.

H.323 Zone

An H.323 zone is a collection of all terminals, gateways, and MCUs managed by a single gatekeeper. A zone includes at least one terminal and may include gateways or MCUs. A zone has only one gatekeeper. A zone may be independent of network topology and may be comprised of multiple network segments that are connected using routers or other devices.

Additional protocols specified by H.323 are listed next. H.323 is independent of the packet network and the transport protocols over which it runs and does not specify them. They are audio codecs; video codecs; H.225 *registration, admission, and status* (RAS); H.225 call signaling; H.245 control signaling; RTP; and *Real-Time Control Protocol* (RTCP).

Audio Codec

An audio codec encodes the audio signal from the microphone for transmission on the transmitting H.323 terminal and decodes the received audio code that is sent to the speaker on the receiving H.323 terminal. Because audio is the minimum service provided by the H.323 standard, all H.323 terminals must have at least one audio codec support, as specified in the ITU-T G.711 recommendation (audio coding at 64 Kbps). Additional audio codec recommendations such as G.722 (64, 56, and 48 Kbps), G.723.1 (5.3 and 6.3 Kbps), G.728 (16 Kbps), and G.729 (8 Kbps) may also be supported.

Video Codec

A video codec encodes video from the camera for transmission on the transmitting H.323 terminal and decodes the received video code that is sent to the video display on the receiving H.323 terminal. Because H.323 specifies support of video as optional, the support of video codecs is optional as well. However, any H.323 terminal providing video communications must support video encoding and decoding as specified in the ITU-T H.261 recommendation.

H.225 RAS

RAS is the protocol between endpoints (terminals and gateways) and gatekeepers. The RAS protocol is used to perform registration, admission control, bandwidth changes, status, and disengage procedures between endpoints and gatekeepers. An RAS channel is used to exchange RAS messages. This signaling channel is opened between an endpoint and a gatekeeper prior to the establishment of any other channels.

H.225 Call Signaling

H.225 call signaling is used to establish a connection between two H.323 endpoints. This is achieved by exchanging H.225 protocol messages on the call-signaling channel. The call-signaling channel is opened between two H.323 endpoints or between an endpoint and the gatekeeper.

H.245 Control Signaling

H.245 control signaling is used to exchange end-to-end control messages governing the operation of the H.323 endpoint. These control messages carry information related to the following: capabilities exchange, opening and closing of logical channels used to carry media streams, flow control messages, and general commands and indications.

RTP

RTP, a transport protocol, provides end-to-end delivery services of real-time audio and video. Whereas H.323 is used to transport data over IP-based networks, RTP is typically used to transport data via the UDP. RTP, together with UDP, provides transport-protocol functionality. RTP provides payload-type identification, sequence numbering, time stamping, and delivery monitoring. UDP provides multiplexing and checksum services. RTP can also be used with other transport protocols.

RTCP

RTCP is the counterpart of RTP that provides control services. The primary function of RTCP is to provide feedback on the quality of the data distribution. Other RTCP functions include carrying a transport-level identifier for an RTP source, called a *canonical name,* which is used by receivers to synchronize audio and video.

Gateway Control Protocols

The most immediate attraction of VoIP is that it saves money on long-distance transport. To date, it has been impractical to route VoIP "desktop to desktop," meaning that interworking between PSTN and IP networks must be facilitated. This is done with a gateway. The two most applied gateways are the media gateway and the signaling gateway. Media gateways interconnect dissimilar networks.

In this case, they connect the PSTN to IP networks. To do this successfully, they must mediate between both signaling and transport between the two dissimilar networks (PSTN and IP). Media gateways coordinate call control and status. Gateway control protocols are signaling protocols.

Media Gateway Control Protocol

The *Media Gateway Control Protocol* (MGCP) is the protocol used to mediate between the *media gateway controller* (MGC, also known as a *call agent*) and the media gateway. MGCP was developed by the IETF and details the commands and parameters that are passed between the MGC and the telephony gateway to be controlled.

MGCP assumes a call control architecture in which the call control "intelligence" is outside the gateways and handled by external call control elements. The MGCP assumes that these call control elements, or call agents, will synchronize with each other to send coherent commands to the gateways under their control. MGCP is a master/slave protocol, where the gateways are expected to execute commands sent by the call agents.

The purpose of MGCP is to send commands from the call agent to a media gateway. MGCP defines both endpoints and connections. Endpoints are sources or sinks of data and can be either physical (such as an interface terminating a digital trunk or analog line) or virtual (such as a designated audio source). An example of a virtual endpoint is an audio source in an audio-content server. Creation of physical endpoints requires hardware installation, while creation of virtual endpoints can be done by software. Endpoint identifiers have two components, the domain name of the gateway that is managing the endpoint and a local name within that gateway. Examples of physical endpoints include interfaces on gateways that terminate a trunk connected to a PSTN switch (Class 5 or Class 4) or an analog *plain old telephone system* (POTS) connection to a phone, key system, PBX, and so on. MGCP sends commands from the call agent to a media gateway. MGCP defines both endpoints and connections.

Connections can be either point to point or multipoint in nature. Further, connections are grouped into calls, where one or more connections can belong to one call. A point-to-point connection is an association between two endpoints with the purpose of transmitting data between these endpoints. Once this association is established for both endpoints, data transfer between these endpoints can take place. A multipoint connection is established by connecting the endpoint to a multipoint session. For point-to-point connections the endpoints of a connection could be in separate gateways or in the same gateway.

The connections and calls are established by the actions of one or more call agents. The information communicated between call agents and endpoints is either events or signals. An example of an event would be a telephone going off hook, while a signal may be the application of dial tone to an endpoint. These events and signals are grouped into what are called *packages,* which are supported by a particular type of endpoint. One package may support events and signals for analog line, whereas another package may support a group of events and signals for video lines.

As long as there are media gateways to interface analog or PSTN connections to IP networks, MGCP will be the controlling protocol. MGCP will continue to be an integral element in any softswitch architecture [8]. Figure 3.4 details the function of MGCP in softswitch architecture.

SS7-Related Protocols

In order for IP telephony networks to interoperate with the PSTN, they must interface with *Signaling System No. 7* (SS7). Softswitch solutions must include the *integrated services digital network user part* (ISUP) and *transactional capabilities application part* (TCAP). ISUP, defined by ITU-T Q.761 and Q.764, is the call control part of the SS7b protocol. ISUP is an SS7 protocol for signaling the parameters and procedures to set up and tear down circuit-switched voice calls between a softswitch/signaling gateway and an STP. ISUP determines the procedures for call setup and teardown on the SS7 network [10].

TCAP is a peer protocol to ISUP in the SS7 protocol hierarchy for end-to-end signaling not associated with call setup or specific trunks in the PSTN network. Some of its main uses are toll-free 800 number translations for routing across the network, and *local number portability* (LNP). It also provides the interface between databases and SCPs [3, p. 9]. TCAP provides services to any number

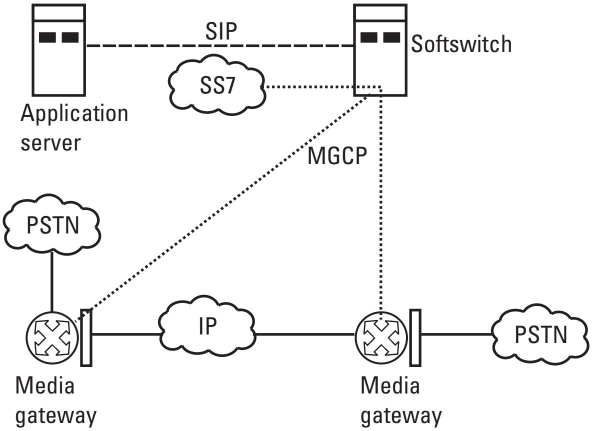


Figure 3.4 Relationship between signaling protocols and softswitch architecture components. (*After:* [9].)

of application parts. Common application parts include the *intelligent network application part* (INAP) and the *mobile application part* (MAP) [7, p. 311].

Routing Protocols

VoIP is routed over an IP network via routers. To deliver the best QoS, voice packets must be given priority over data packets. That means communicating to routers which packets have what priority. Router operations involve several processes. First, the router creates a routing table to gather information from other routers about the optimum path for each packet. This table may be static in that it is constructed by the router according to the current topology and conditions. Dynamic routing is considered a better technique because it adapts to changing network conditions. The router uses a metric of the shortest distance between two endpoints to help determine the optimum path. The router determines the least cost (most efficient) path from origin to destination.

Two algorithms are used to determine the least cost route: distance vector and link state. Protocols that make use of these algorithms are called *interior gateway protocols* (IGPs). The *Routing Information Protocol* (RIP) is an IGP based on the distance vector algorithm and the *Open Shortest Path First* (OSPF) Protocol is an IGP based on the link state algorithm. Where one network needs to communicate with another, it uses an *exterior gateway protocol* (EGP). One example of an EGP is the *Border Gateway Protocol* (BGP).

RIP

RIP is a distance vector protocol that uses hop count (the number of routers it passes through on its route to its destination) as its metric. RIP is widely used for routing traffic on the Internet and is an IGP, which means that it performs routing within a single autonomous system. Exterior gateway protocols, such as BGP, perform routing between different autonomous systems. RIP itself evolved as an Internet routing protocol, and other protocol suites use modified versions of RIP.

OSPF

OSPF is a routing protocol developed for IP networks by the IGP working group of IETF. The working group was formed in 1988 to design an IGP based on the *shortest path first* (SPF) algorithm for use in the Internet. Similar to the *Interior Gateway Routing Protocol* (IGRP), OSPF was created because in the mid-1980s, RIP was increasingly incapable of serving large, heterogeneous internetworks.

OSPF has two primary characteristics. The first is that the protocol is open, which means that its specification is in the public domain. The OSPF specification is published as the IETF's RFC 1247. The second principal characteristic is that OSPF is based on the SPF algorithm, which is sometimes referred to as the *Dijkstra algorithm,* named for the person credited with its creation.

OSPF is a link state routing protocol that calls for the sending of *link state advertisements* (LSAs) to all other routers within the same hierarchical area. Information on attached interfaces, metrics used, and other variables is included in OSPF LSAs. As OSPF routers accumulate link state information, they use the SPF algorithm to calculate the shortest path to each node.

SPF Algorithm

The SPF routing algorithm is the basis for OSPF operations. When an SPF router is powered up, it initializes its routing protocol data structures and then waits for indications from lower layer protocols that its interfaces are functional. After a router is assured that its interfaces are functioning, it uses the OSPF *Hello protocol* to acquire neighbors, which are routers with interfaces to a common network. The router sends hello packets to its neighbors and receives their hello packets. In addition to helping acquire neighbors, hello packets also act as keep-alives to let routers know that other routers are still functional.

Each router periodically sends an LSA to provide information on a router's adjacencies or to inform others when a router's state changes. By comparing established adjacencies to link states, failed routers can be detected quickly and the network's topology altered appropriately. From the topological database generated from LSAs, each router calculates a shortest path tree, with itself as the root. The shortest path tree, in turn, yields a routing table [11].

BGP

BGP performs interdomain routing in TCP/IP networks. BGP is an EGP, which means that it performs routing between multiple autonomous systems or domains and exchanges routing and reachability information with other BGP systems. BGP was developed to replace its predecessor, the now obsolete *Exterior Gateway Protocol,* as the standard exterior gateway-routing protocol used in the global Internet. BGP solves serious problems with EGP and scales to Internet growth more efficiently.

BGP performs three types of routing: interautonomous system routing, intra-autonomous system routing, and pass-through autonomous system routing. *Interautonomous system routing* occurs among two or more BGP routers in different autonomous systems. Peer routers in these systems use BGP to maintain a consistent view of the internetwork topology. BGP neighbors communicating between autonomous systems must reside on the same physical network. The Internet serves as an example of an entity that uses this type of routing because it is comprised of autonomous systems or administrative domains. Many of these domains represent the various institutions, corporations, and entities that make up the Internet. BGP is frequently used to provide path determination to provide optimal routing within the Internet.

Intra-autonomous system routing occurs between two or more BGP routers located within the same autonomous system. Peer routers within the same autonomous system use BGP to maintain a consistent view of the system topology. BGP also is used to determine which router will serve as the connection point for specific external autonomous systems. Once again, the Internet provides an example of interautonomous system routing. An organization, such as a university, could make use of BGP to provide optimal routing within its own administrative domain or autonomous system. The BGP protocol can provide both inter- and intra-autonomous system routing services.

Pass-through autonomous system routing occurs between two or more BGP peer routers that exchange traffic across an autonomous system that does not run BGP. In a pass-through autonomous system environment, the BGP traffic did not originate within the autonomous system in question and is not destined for a node in the autonomous system. BGP must interact with whatever intra-autonomous system routing protocol is being used to successfully transport BGP traffic through that autonomous system.

BGP Routing

As with any routing protocol, BGP maintains routing tables, transmits routing updates, and bases routing decisions on routing metrics. The primary function of a BGP system is to exchange network-reachability information, including information about the list of autonomous system paths, with other BGP systems. This information can be used to construct a graph of autonomous system connectivity from which routing loops can be pruned and with which autonomous system-level policy decisions can be enforced.

Each BGP router maintains a routing table that lists all feasible paths to a particular network. The router does not refresh the routing table, however. Instead, routing information received from peer routers is retained until an incremental update is received. BGP devices exchange routing information upon initial data exchange and after incremental updates. When a router first connects to the network, BGP routers exchange their entire BGP routing tables. Similarly, when the routing table changes, routers send the portion of their routing table that has changed. BGP routers do not send regularly scheduled routing updates, and BGP routing updates advertise only the optimal path to a network.

BGP uses a single routing metric to determine the best path to a given network. This metric consists of an arbitrary unit number that specifies the degree of preference of a particular link. The BGP metric typically is assigned to each link by the network administrator. The value assigned to a link can be based on any number of criteria, including the number of autonomous systems through which the path passes, stability, speed, delay, or cost [12].

Resource Reservation Protocol

The *Resource Reservation Protocol* (RSVP) is a network control protocol that enables Internet applications to obtain special QoSs for their data flows. RSVP is not a routing protocol; instead, it works in conjunction with routing protocols and installs the equivalent of dynamic access lists along the routes that routing protocols calculate. RSVP occupies the place of a transport protocol in the OSI model seven-layer protocol stack. The IETF is now working toward standardization through an RSVP working group. RSVP operational topics discussed in this chapter include data flows, quality of service, session startup, reservation style, and soft state implementation.

In RSVP, a data flow is a sequence of messages that have the same source, destination (one or more), and quality of service. QoS requirements are communicated through a network via a *flow specification*, which is a data structure used by internetwork hosts to request special services from the internetwork. A flow specification often guarantees how the internetwork will handle some of its host traffic.

RSVP supports three traffic types: best effort, rate sensitive, and the delay sensitive. The type of data flow service used to support these traffic types depends on the QoS implemented. The following paragraphs address these traffic types and associated services.

Best effort traffic is traditional IP traffic. Applications include file transfer, such as mail transmissions, disk mounts, interactive logins, and transaction traffic. The service supporting best effort traffic is called *best effort service.*

Rate-sensitive traffic is willing to give up timeliness for guaranteed rate. Rate-sensitive traffic, for example, might request 100 Kbps of bandwidth. If it actually sends 200 Kbps for an extended period, a router can delay traffic. Rate-sensitive traffic is not intended to run over a circuit-switched network; however, it usually is associated with an application that has been ported from a circuit-switched network (such as ISDN) and is running on a datagram network (IP). An example of such an application is H.323 videoconferencing, which is designed to run on ISDN (H.320) or ATM (H.310) but is found on the Internet. H.323 encoding is constant rate or nearly constant rate, and it requires a constant transport rate. The RSVP service supporting rate-sensitive traffic is called *guaranteed bit-rate service.*

Delay-sensitive traffic is traffic that requires timeliness of delivery and varies its rate accordingly. MPEG-II video, for example, averages about 3 to 7 Mbps depending on the amount of change in the picture. As an example, 3 Mbps might be a picture of a painted wall, although 7 Mbps would be required for a picture of waves on the ocean. MPEG-II video sources send key and delta frames. Typically, 1 or 2 key frames per second describe the whole picture, and 13 or 28 frames describe the change from the key frame. Delta frames are usually substantially smaller than key frames. As a result, rates vary quite a bit from frame to frame. A single frame, however, requires delivery within a frame time or the codec is unable to do its job. A specific priority must be negotiated for delta-frame traffic. RSVP services supporting delay-sensitive traffic are referred to as *controlled-delay service* (nonreal-time service) and *predictive service* (real-time service).

In the context of RSVP, QoS is an attribute specified in flow specifications that is used to determine the way in which data interchanges are handled by participating entities (routers, receivers, and senders). RSVP is used to specify the QoS by both hosts and routers. Hosts use RSVP to request a QoS level from the network on behalf of an application data stream. Routers use RSVP to deliver QoS requests to other routers along the path(s) of the data stream. In doing so, RSVP maintains the router and host state to provide the requested service [13].

Transport Protocols

RTP

RTP is the most popular of the VoIP transport protocols. It is specified in RFC 1889 under the title of “RTP: A Transport Protocol for Real-Time Applications.” This RFC describes both RTP and RTCP. As the names would suggest, these two protocols are necessary to support real-time applications like voice and video. RTP operates on the layer above UDP, which does not avoid packet loss or guarantee the correct order for the delivery of packets. RTP packets overcome those shortcomings by including sequence numbers that help applications using RTP to detect lost packets and ensure packet delivery in the correct order. RTP packets include a time stamp that gives the time when the packet is sampled from its source media stream. This time stamp assists the destination application to determine the synchronized playout to the destination user and to calculate delay and jitter—two very important detractors of voice quality. RTP does not have the capacity to correct delay and jitter, but does provide additional information to a higher layer application so that the application can make determinations as to how a packet of voice or data is best handled.

RTCP provides a number of messages that are exchanged between session users and that provide feedback regarding the quality of the session. The type of information includes details such as the numbers of lost RTP packets, delays, and interarrival jitter. As voice packets are transported in RTP packets, RTCP packets transfer quality feedback. Whenever an RTP session opens, an RTCP session is also opened. That is, when a UDP port number is assigned to an RTP session for transfer of media packets, another port number is assigned for RTCP messages.

RTP Payloads

RTP carries the digitally encoded voice by taking one or more digitally encoded voice samples and attaching an RTP header to provide RTP packets, which are made up of an RTP header and a payload of the voice samples. These RTP packets are sent to UDP, where a UDP header is attached. This combination then goes to IP where an IP header is attached and the resulting IP datagram is routed to the destination. At the destination, the headers are used to pass the packet up the stack to the appropriate application.

RTP Headers

RTP carries the carried voice in a packet. The RTP payload is comprised of digitally coded samples. The RTP header is attached to this payload and the packet is sent to the UDP layer. The RTP header contains the necessary information for the destination to reconstruct the original voice sample.

RTCP

RTCP enables exchanges of control information between session participants with the goal of providing quality-related feedback. This feedback is used to detect and correct distribution issues. The combination of RTCP and IP multicast allows a network operator to monitor session quality. RTCP provides information on the quality of an RTP session. RTCP empowers network operators to obtain information about delay, jitter, and packet loss and to take corrective action where possible to improve quality.

Internet Protocol Version 6

The previous discussion assumed the use of *Internet Protocol version 4* (IPv4), the predominant version of IP in use today. A new version, *Internet Protocol version 6* (IPv6) is now coming on the market. The explosion of Internet addresses necessitates the deployment of IPv6. IPv6 makes possible infinitely more addresses than IPv4. Enhancements offered by IPv6 over IPv4 include the following:

- *Expanded address space:* each address is allocated 128 bits instead of 32 bits in IPv4;
- *Simplified header format:* enables easier processing of IP datagrams;
- *Improved support for headers and extensions:* enables greater flexibility for the introduction of new options;
- *Flow-labeling capability:* enables the identification of traffic flows for real-time applications;
- *Authentication and privacy:* support for authentication, data integrity, and data confidentiality are supported at the IP level rather than through separate protocols or mechanisms above IP.

Conclusion

This chapter addressed the building blocks of VoIP. It will be necessary in future chapters to understand many of the concepts contained in this chapter. Just as the PSTN and softswitch networks can be broken down into the three elements of access, switching, and transport, VoIP can be broken down into a study of three types of protocols: signaling, routing, and transport. The proper selection of VoIP signaling protocols for a network is an essential issue. Although protocols will continue to evolve and new protocols will emerge, those addressed in this chapter will constitute the predominant structure of Vo802.11 [4, p. 49]. By understanding the elements of VoIP and mating it to 802.11 (wireless Ethernet), it is possible to bypass the "last mile" of the PSTN.

References

[1] TeleGeography, *TeleGeography 2002—Global Traffic Statistics and Commentary*, http://www.TeleGeography.com, 2001.

[2] *Report to Congress on Universal Service*, CC Docket No. 96-45, white paper on IP voice services, March 18, 1998, http://www.von.org/docs/whitepap.pdf.

[3] Douskalis, B., *IP Telephony: The Integration of Robust VoIP Services,* Upper Saddle River, NJ: Prentice Hall, 2000.

[4] Ohrtman, F. D., *Softswitch: Architecture for VoIP,* New York: McGraw-Hill, 2002.

[5] Johnston, A. B., *SIP: Understanding the Session Initiation Protocol,* Norwood, MA: Artech House, 2001.

[6] Camarillo, G., *SIP Demystified,* New York: McGraw-Hill, 2002.

[7] Collins, D., *Carrier Grade Voice over IP,* 2nd ed., New York: McGraw-Hill, 2002.

[8] Internet Engineering Task Force, “Media Gateway Control Protocol,” RFC 2705, October 1999.

[9] http://www.nuera.com/products/gxseries_diag.cfm and http://www.nuera.com/products/ssc_diag.cfm.

[10] Newton, H., *Newton’s Telecom Dictionary,* 16th ed., Gilroy, CA: CMP Books, 2000, p. 486.

[11] Cisco Systems, “Open Shortest Path First,” white paper, http://www.cisco.com.

[12] Cisco Systems, “Border Gateway Protocol,” white paper, June 1999, http://www.cisco.com.

[13] Cisco Systems, “Resource Reservation Protocol,” white paper, June 1999, http://www.ciscosystems.com.

4

Switching TDM and VoIP Networks

As explained earlier in this book, a telephone network is comprised of three elements: access, switching, and transport. This chapter will introduce the reader to switching both for legacy TDM technologies and for VoIP. Vo802.11 replaces PSTN Class 4 and Class 5 switches with softswitch platforms. This potentially spares an enterprise high fees for Centrex services. In addition, a service provider utilizing softswitch technologies need not purchase multimillion-dollar Class 4 or Class 5 switches. In bypassing Class 4 and Class 5 switches, a Vo802.11 service provider can greatly lower the barriers to entry to the telecommunications market.

TDM Switching

To fully understand VoIP switching and subsequently Vo802.11, we must first understand the physics of voice switching. Much of this technology has evolved during the century of telephony. Many aspects of TDM voice technology have been incorporated by VoIP.

Multiplexing

The earliest approach to enabling multiple conversations over one circuit was *frequency-division multiplexing* (FDM). FDM was made possible by the vacuum tube, in which the range of frequencies was divided into parcels that were distributed among subscribers. In the first FDM architectures, the overall system bandwidth was 96 kHz. This 96 kHz could be divided among a number of subscribers into, for example, 5 kHz per subscriber, meaning almost 20 subscribers could use this circuit.

FDM, however, is an analog technology and suffers from a number of shortcomings. It is susceptible to picking up noise along the transmission path. This FDM signal loses its power over the length of the transmission path. FDM requires amplifiers to strengthen the signal over that path. However, the amplifiers cannot separate the noise from the signal and the end result is an amplified noisy signal.

The improvement over FDM was time-division multiplexing. TDM was made possible by the transistor, which arrived on the market in the 1950s and 1960s. As the name implies, TDM divides the *time* rather than the frequency of a signal over a given circuit. Where FDM was typified by "some of the frequency all of the time," TDM is "all of the frequency some of the time." TDM is a digital transmission scheme that uses a small number of discrete signal states. Digital carrier systems have only three valid signal values: one positive, one negative, and zero. Everything else is registered as noise. A repeater, known as a *regenerator,* can receive a weak and noisy digital signal, remove the noise, reconstruct the original signal, and amplify if before transmitting the signal onto the next segment of the transmission facility. Digitization brings with it the advantages of better maintenance and troubleshooting capability resulting in better reliability. Also, a digital system allows improved configuration flexibility.

TDM has made the *multiplexer,* also known as the *channel bank,* possible. In the United States the multiplexer or *mux* enables 24 channels per single four-wire facility. This is called a T1, DS1, or T-Carrier. Outside North America and Japan, it is 30 channels per facility. These systems came on the market in the early 1960s as a means to transport multiple channels of voice over expensive transmission facilities.

Voice Digitization

One of the first processes in the transmission of a telephone call is the conversion of an analog signal into a digital signal. This process is called *pulse code modulation* (PCM). This is a four-step process consisting of *pulse amplitude modulation* (PAM) sampling, companding, quantization, and encoding.

PCM

PAM Sampling

The first stage in PCM is known as *pulse amplitude modulation.* In order for an analog signal to be represented as a digitally encoded bit stream, the analog signal must be sampled at a rate that is equal to twice the bandwidth of the channel over which the signal is to be transmitted. As each analog voice channel is allocated 4 kHz of bandwidth, each voice signal is sampled at twice that rate, or

8,000 samples per second. In a T-Carrier, the standard in North America and Japan, each channel is sampled every one-eight-thousandth of a second in rotation, resulting in the generation of 8,000 pulse amplitude samples from each channel every second. If the sampling rate is too high, too much information is transmitted and bandwidth is wasted. If the sampling rate is too low, aliasing may result. *Aliasing* is the interpretation of the sample points as a false waveform due to the lack of samples.

Companding

The second stage in PCM is *companding.* Companding is the process of compressing the values of the PAM samples to fit the nonlinear quantizing scale that results in bandwidth savings of more than 30%. It is called *companding* because the sample is compressed for transmission and expanded for reception [1].

Quantization

The third stage in PCM is *quantization.* In quantization, values are assigned to each sample within a constrained range. In using a limited number of bits to represent each sample, the signal is quantized. The difference between the actual level of the input analog signal and the digitized representation is known as *quantization noise.* Noise is a detraction to voice quality and it is necessary to minimize noise. The way to do this is to use more bits, thus providing better granularity. There is, in this case, the inevitable trade-off in bandwidth versus quality. More bandwidth usually improves signal quality. Bandwidth costs money. Service providers, whether using TDM or VoIP for voice transmission, will always have to choose between quality and bandwidth.

A process known as *nonuniform quantization* involves the use of smaller quantization steps at smaller signal levels and larger quantization steps for larger signal levels. This gives the signal greater granularity or quality at low signal levels and less granularity (quality) at high signal levels. The result is to spread the signal-to-noise ratio more evenly across the range of different signals and to enable fewer bits to be used compared to uniform quantization. This process results in less bandwidth being consumed than for uniform quantization [2, pp. 95–96].

Encoding

The final process in PCM is *encoding* the signal. This is performed by a codec. There are three types of codecs: waveform codecs, source codecs, and hybrid codecs.

Waveform codecs sample and code incoming analog signals without regard to how the signals were generated. Quantized values of the samples are then transmitted to the destination where the original signal is reconstructed at least to a certain approximation of the original. Waveform codecs are known for simplicity with high-quality output. The disadvantage of waveform codecs is that they consume considerably more bandwidth than the other codecs. When waveform codecs are used at low bandwidth, speech quality degrades markedly.

Source codecs, also known as *vocoders,* match an incoming signal to a mathematical model of how speech is produced. They use the linear predictive filter model of the vocal tract, with a voiced/unvoiced flag to represent the excitation that is applied to the filter. The filter represents the vocal tract and the voice/unvoiced flag represents whether a voiced or unvoiced input is received from the vocal chords. The information transmitted is a set of model parameters as opposed to the signal itself. The receiver, using the same modeling technique in reverse, reconstructs the values received into an analog signal.

Source codecs operate at low bit rates and reproduce a synthetically sounding voice. Using higher bit rates does not result in improved voice quality. Vocoders are most widely used in private and military applications.

Hybrid codecs are deployed in an attempt to derive the benefits from both technologies. They perform some degree of waveform matching while mimicking the architecture of human speech. Hybrid codecs provide better voice quality at low bandwidth than waveform codecs.

Popular Speech Codecs

Tables 4.1 and 4.2 list *International Telecommunication Union* (ITU) voice codecs, their descriptions, and data rates. The most popular of the speech codecs are discussed next.

G.711

G.711 is the best known coding technique in use today. It is a waveform codec and is the coding technique used circuit-switched telephone networks all over the word. G.711 has a sampling rate of 8,000 Hz. If uniform quantization were to be used, the signal levels commonly found in speech would be such that at least 12 bits per sample would be needed, giving it a bit rate of 96 Kbps. Nonuniform quantization is used with 8 bits representing each sample. This quantization leads to the well-known 64-Kbps DS0 rate. G.711 is often referred to as PCM.

G.711 has two variants: A-law and μ-law. μ-law is used in North America and Japan where T-Carrier systems prevail. A-law is used everywhere else in the world. The difference between the two lies in the way nonuniform quantization is performed. Both are symmetrical at approximately zero. Both A-law and μ-law offer good voice quality with a *mean opinion score* (MOS), a means of rating relative voice quality with 5 being the best and 1 being the worst), of 4.3. Despite being the predominant codec in the industry, G.711 suffers one significant drawback: It consumes 64 Kbps in bandwidth. Carriers seek to deliver like voice quality using less bandwidth.

Table 4.1
Brief Descriptions of ITU Voice Codecs

ITU Standard	Description
P.800	Subjective rating system to determine mean opinion score or quality of telephone connections
G.114	Maximum one-way delay end to end for VoIP call (150 ms)
G.165	Echo cancellers
G.168	Digital network echo cancellers
G.711	PCM of voice frequencies
G.722	7-kHz audio coding within 64 Kbps
G.723.1	Dual-rate speech coder for multimedia communications transmitting at 5.3 and 6.3 Kbps
G.729	Coding for speech at 8 Kbps using conjugate-structure algebraic code-excited linear prediction
G.729A	Annex A reduced complexity 8-Kbps CS-ACELP speech codec
H.323	Packet-based multimedia communications system
P.861	Specifies a model to map actual audio signals to their representations inside the human head
Q.931	Digital Subscriber Signaling System No. 1 ISDN User-Network Interface Layer 3 Specification for Basic Call Control

Table 4.2
ITU Voice Codecs and Their Performance

Standard	Data Rate (Kbps)	Delay (ms)	MOS	Codec
G.711	64	0.125	4.8	Waveform
G.721	16, 24, 32, 40	0.125	4.2	
G.723				
G.726				
G.728	16	2.5	4.2	
G.729	8	10	4.2	
G.723.1	5.3, 6.3	30	3.5, 3.98	

G.728 LD-CELP

Code-excited linear predictor (LD-CELP) codecs implement a filter and contain a codebook of acoustic vectors. Each vector contains a set of elements in which the elements represent various characteristics of the excitation signal. CELP coders transmit to the receiving end a set of information determining filter coefficients, gain, and a pointer to the chosen excitation vector. The receiving end contains the same codebook and filter capabilities so that it reconstructs the original signal.

G.728 is a backward adaptive coder as it uses previous speech samples to determine the applicable filter coefficients. G.728 operates on five samples at one time. That is, five samples at 8,000 Hz are needed to determine a codebook vector and filter coefficients based on previous and current samples. Given a coder operating on five samples at a time, a delay of less than 1 ms is the result. Low delay equals better voice quality.

The G.728 codebook contains 1,024 vectors, which requires a 10-bit index value for transmission. It also uses five samples at a time taken at a rate of 8,000 per second. For each of those five samples, G.728 results in a transmitted bit rate of 16 Kbps. Hence, G.728 has a transmitted bit rate of 16 Kbps. Another advantage here is that this coder introduces a delay of 0.625 ms with a MOS of 3.9, the difference from G.711's MOS of 4.3 is imperceptible to the human ear. The bandwidth savings between G.728's 16 Kbps per conversation and G.711's 64 Kbps per conversation make G.728 very attractive to carriers given the savings in bandwidth.

G.723.1 ACELP

G.723.1 *algebraic code-excited linear prediction* (ACELP) can operate at either 6.3 or 5.3 Kbps with the 6.3-Kbps mode providing higher voice quality. Bit rates are contained in the coder and decoder and the transition between the two can be made during a conversation. The coder takes a bank-limited input speech signal that is sampled a 8,000 Hz and undergoes uniform PCM quantization resulting in a 16-bit PCM signal. The encoder then operates on blocks or frames of 240 samples at a time. Each frame corresponds to 30 ms of speech, which means that the coder causes a delay of 30 ms. Including a look-ahead delay of 7.5 ms gives a total algorithmic delay of 37.5 ms.

G.723.1 has a MOS of 3.8, which is highly advantageous with regard to the bandwidth used. The delay of 37.5 ms one way does present an impediment to good quality, but the round-trip delay over varying aspects of a network determine final delay and not necessarily the codec used.

G.729

G.729 is a speech coder that operates at 8 Kbps. This coder uses input frames of 10 ms, corresponding to 80 samples at a sampling rate of 8,000 Hz. This coder includes a 5-ms look-ahead, resulting in an algorithmic delay of 15 ms (considerably better than G.723.1). G.729 uses an 80-bit frame. The transmitted bit rate is 8 Kbps. Given that it turns in a MOS of 4.0, G.729 is perhaps the best trade-off in bandwidth for voice quality.

The previous sections provide and overview of the multiple means of maximizing the efficiency of transport via the PSTN. We find today that TDM is almost synonymous with circuit switching. Telecommunications engineers use the term *TDM* to describe a circuit-switched solution. The 64-Kbps G.711 codec is the standard in use for PSTN. The codecs described in the previous pages apply to voice over IP as well. VoIP engineers seeking to squeeze more conversations over valuable bandwidth have found these codecs very valuable in compressing voice over IP conversations over an IP circuit [3, pp. 19–21].

Signaling

For much of the history of circuit-switched networks, signaling followed the same path as conversation. This is called *channel-associated signaling* (CAS) and is still in wide use today. R1 *Multifrequency* (MF) used in North American markets and R2 *Multifrequency Compelled* (RFC) used elsewhere in the world are the best examples of this. Another name for this is *in-channel signaling.*

The newer technology for signaling is called *common channel signaling* (CCS), also known as out-of-band signaling. CCS uses a separate transmission path for call signaling and not the bearer path for the call. This separation enables the signaling to be handled in a different manner than is the call. This allows signaling to be managed by a network independent of the transport network.

SS7

SS7 is common channel signaling and is the standard for CCS with many national variants throughout the world (for example, Mexico's NOM-112). It routes control messages through the network to perform call management (setup, maintenance, termination) and network management functions. Although the network being controlled is circuit switched, the control signaling is implemented using packet-switching technology. In effect, a packet-switched network is overlaid on a circuit-switched network in order to operate and control the circuit-switched network. SS7 defines the functions that are performed in the packet switched network but does not dictate any particular hardware implementation.

The SS7 network and protocol are used for the following tasks:

- Basic call setup, management, and tear down;
- Wireless services such as *personal communications services* (PCS), wireless roaming, and mobile subscriber authentication;
- Local number portability;
- Toll-free (800/888) and toll (900) wire-line services;

- Enhanced call features such as call forwarding, calling party name/number display, and three-way calling;
- Efficient and secure worldwide telecommunications.

Signaling Links

SS7 messages are exchanged between network elements over 56- or 64-Kbps bidirectional channels called *signaling links.* Signaling occurs out of band on dedicated channels rather than in band on voice channels. Compared to in-band signaling, out-of-band signaling provides the following:

- Faster call setup times, compared to in-band signaling using MF signaling tones;
- More efficient use of voice circuits;
- Support for *intelligent network* (IN) services, which require signaling to network elements without voice trunks (e.g., database systems);
- Improved control over fraudulent network usage.

Signaling Points

Each signaling point in the SS7 network is uniquely identified by a numeric point code. Point codes are carried in signaling messages exchanged between signaling points to identify the source and destination of each message. Each signaling point uses a routing table to select the appropriate signaling path for each message. Per Figure 4.1, there are three kinds of signaling points in the SS7 network: a *service switching point* (SSP), *signal transfer point* (STP), and *service control point* (SCP).

SSPs are switches that originate, terminate, or tandem calls. An SSP sends signaling messages to other SSPs to set up, manage, and release the voice circuits required to complete a call. An SSP may also send a query message to a centralized database (an SCP) to determine how to route a call (e.g., a toll-free 1-800/888 call in North America). An SCP sends a response to the originating SSP containing the routing number(s) associated with the dialed number. An

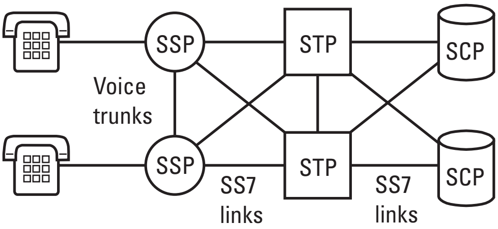


Figure 4.1 SS7 signaling points. (*From:* [4]. © 2003 Performance Technologies, Inc. Reprinted with permission.)

alternate routing number may be used by the SSP if the primary number is busy or the call is unanswered within a specified time. Actual call features vary from network to network and from service to service.

Network traffic between signaling points may be routed via a packet switch called an STP. An STP routes each incoming message to an outgoing signaling link based on routing information contained in the SS7 message. Because it acts as a network hub, an STP provides improved utilization of the SS7 network by eliminating the need for direct links between signaling points. An STP may perform global title translation, a procedure by which the destination signaling point is determined from digits present in the signaling message (e.g., the dialed 800 number, calling card number, or mobile subscriber identification number). An STP can also act as a "firewall" to screen SS7 messages exchanged with other networks.

Because the SS7 network is critical to call processing, SCPs and STPs are usually deployed in mated pair configurations in separate physical locations to ensure network-wide service in the event of an isolated failure. Links between signaling points are also provisioned in pairs. Traffic is shared across all links in the linkset. If one of the links fails, the signaling traffic is rerouted over another link in the linkset. The SS7 protocol provides both error correction and retransmission capabilities to allow continued service in the event of signaling point or link failures.

SS7 Protocol Stack

The hardware and software functions of the SS7 protocol are divided into functional abstractions called *levels.* These levels map loosely to the OSI seven-layer model defined by the ISO, as shown in Figure 4.2.

Features

Custom local-area signaling service (CLASS) features are basic services available in each *local access and transport area* (LATA). Features and the services they enable are a function of Class 5 switches and SS7 networks. The Class 4 switch offers no features of its own. It transmits the features of the Class 5 switch. With almost 3 decades of development, the Class 4 switch has a well-established history of seamless interoperability with the features offered by the Class 5 and SS7 networks. Features often allow service provider systems to generate high margins that, of course, equate to stronger revenue streams.

Examples of features offered through the DMS-250 system can be grouped under two major portfolios: basic and enhanced services. The basic services include 1+, 800/900 service, travel cards, account codes, pin numbers, operator access, speed dialing, hotline service, *automatic number identification* (ANI) screening, *virtual private networks* (VPNs), calling cards, call detail recording. Enhanced services include information database services [*number*

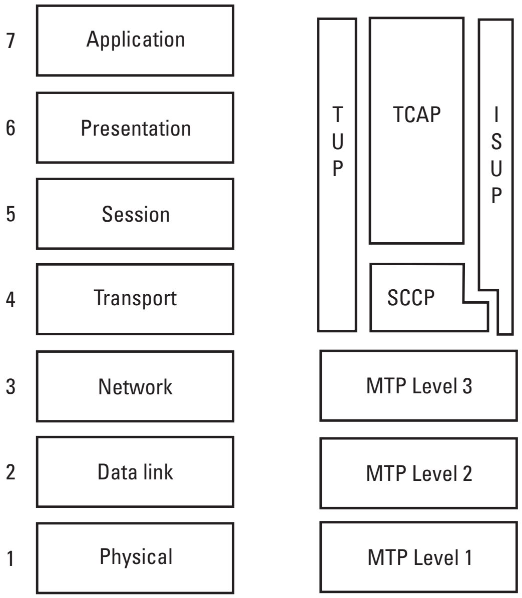


Figure 4.2 Comparison of OSI reference model and SS7 protocol stacks. (*From:* [4]. © 2003 Performance Technologies, Inc. Reprinted with permission.)

identification (NXX) number service, authorization codes, calling card authorization, debit/prepaid card], and routing and screening [includes *circuit identification code* (CIC) routing, time-of-day screening, ANI screening, class-of-service screening]. Enhanced features include enterprise networks, data and video services (dedicated access lines, ISDN PRI services, dialable wideband services, switched 56 Kbps), and multiple dialing plans (full 10-digit routing, 70-digit VPN routing, 15-digit international dialing, speed dialing, hotline dialing). Most of these features have been standard on the DMS-250 and other Class 4 switches for many years.

The long list of features above is evidence of the importance of features in the legacy market in which they were developed. Service providers are reluctant to give up these features and the higher margins they generate. In the converging market, features are equally important to reliability because service providers do not want to offer fewer features to their customers and they will want to continue to offer high margin features.

Transport

The PSTN was built over a period of a century plus at great expense. Developers have been obsessed over the years with getting the maximum number of conversations transported at the least cost in infrastructure possible. Imagine an early telephone circuit running from New York to Los Angeles. The network of copper wire, repeaters, and other mechanisms involved in transporting a conversation this distance was immense for its time. Hence, the early telephone engineers and scientists had to find ways to transport the maximum number of conversations over this network. Through much research, different means were developed to wring the maximum efficiency from the copper wire infrastructure. Many of those discoveries translated into technologies that worked equally well when fiber-optic cable came on the market [5].

Softswitch and Distributed Architecture: A “Stupid” Network

Figure 4.3 illustrates the distributed architecture that is generally agreed on as the model for softswitched networks. This model decouples the underlying packet-switching hardware from the call control, service logic, and new service creation. This distribution allows for flexibility in hardware choices as well as innovation of new services without requiring changes in the switching fabric or structure, and opens up the opportunities for third-party developers. The bottom layer is considered the bearer or transport plane, which physically transports

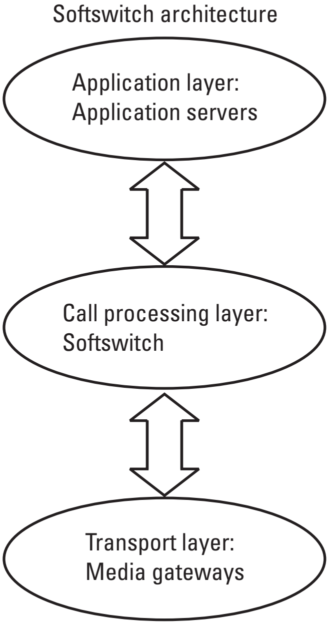


Figure 4.3 Softswitch architecture. (*From:* [4]. © 2003 Performance Technologies, Inc. Reprinted with permission.)

both voice and data traffic. This plane consists of the media gateways in the softswitch solution [6]. What makes this possible is the client/server architecture of softswitch as opposed to the mainframe architecture of the Classes 4 and 5 switches (Figure 4.4). One advantage is that it allows a service provider to start small and grow with demand as opposed to making a large up-front investment in a Class 4 switch.

In 1997 *Computer Telephony* magazine printed a white paper entitled "Rise of the Stupid Network" authored by David Isenberg, a scientist at Bell Labs [5]. In the paper, Isenberg pointed out that the Internet is the inverse of the PSTN in that the intelligence of the Internet resides at the periphery of the network as opposed to residing at the core of the network as it does in the PSTN. Softswitch architecture reflects a "stupid" network. Softswitch is a sum of its parts distributed across an IP network, as opposed to the PSTN where a few, large, highly centralized Classes 4 and 5 switches operate. In the following sections, we discuss the components and ideology of a "stupid" network [3, p. 52]. Softswitch can be considered a "stupid" solution because it utilizes a distributed architecture (intelligence at the periphery) as opposed to the "smart" or centralized architecture of the Classes 4 and 5 switches.

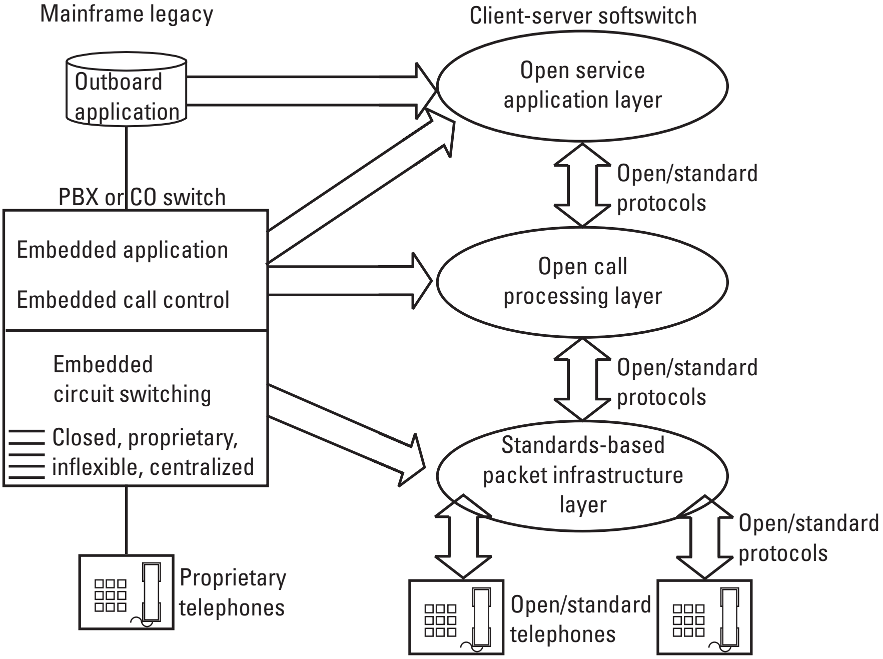


Figure 4.4 Mainframe versus softswitch client/server architecture. Note fluidity of softswitch architecture. (*From:* [7]. © 1999 Cisco Systems, Inc. Reprinted with permission.)

Access

The softswitch architecture, like the PSTN, can be described as having three elements: (1) access—how a subscriber gains access to the network, (2) switching—how a call is controlled across the network, and (3) transport—how a call is transported across the network. In the case of accessing a VoIP network, access can be gained either from an IP source (PC or IP phone) or from a legacy, analog handset via a media gateway.

PC-to-PC and PC-to-Phone Applications

The first VoIP applications used personal computers equipped with speakers and microphones as terminals for access to a VoIP network. Initially, the QoS left much to be desired and, as a result, this form of access did not immediately catch on in the market. This service is often referred to as *PC-to-PC service.* It is also possible to complete phone calls using *PC-to-phone service* (Figure 4.5). PC-to-PC and PC-to-phone applications are now used most widely by consumers for long-distance bypass. The market driver for this form of access has been saving money on long-distance calls, most specifically on international long distance. While often touted as an enterprise telephony solution, the use of PC as a telephony terminal has not seized any significant market share. Even where the QoS was acceptable for the task, there remained anthropological issues. PCs do not resemble telephones in appearance, feel, or function. This presents a psychological barrier to the user for using a PC as readily as a telephone handset [8].

IP Phones (IP Handsets) Phone-to-Phone VoIP

It did not take industry long to realize the benefits of a PC-in-a-handset for use in VoIP. Thus was born the *IP phone.* Early pioneers of this technology included

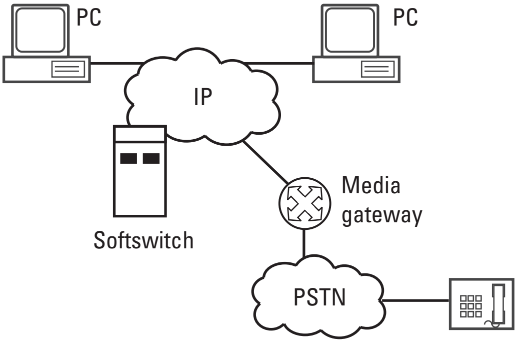


Figure 4.5 PC-to-PC and PC-to-phone applications.

e-tel and PingTel. The IP handset incorporates all computer hardware necessary to make an IP phone call possible. Another strong advantage of the IP handset is that it removes anthropological objections to VoIP calls. The IP handset looks and functions like a telephone as opposed to a PC. IP handsets are stand-alone devices an present an IP desktop-to-desktop solution.

IP handsets offer further benefits in that they do not require a gateway and its incumbent investment and management responsibilities. The chief advantage of an IP phone to an enterprise is that the phone requires a minimum of network configuration and management. Each employee equipped with an IP phone can take that phone anywhere on a network with no reconfiguration of the phone or the network. IP phone-equipped employees are potentially more productive because the *graphic user interface* (GUI) on the IP phone makes using features much easier than with a 12-button conventional telephone handset and its list of star codes.

The IP handset has its own IP address, which is recognized wherever it is connected on an IP network. In an enterprise setting, a worker can disconnect his or her IP phone and move to another cubicle, building, or state and the phone will function with no reprogramming necessary. In a legacy enterprise setting with a circuit-switched PBX, tools for managing moves/adds/changes tend to be difficult to use and, consequently, administrators learn only the basic management skills. This makes it very expensive to administer the switch. According to some estimates it can cost as much as $500 per PBX move/add/change. For a Centrex line, it can take weeks for a change to be implemented by the telephone company [9].

IP phones are available in two “flavors” (or VoIP protocols, to be covered in following chapters), H.323 and SIP, and require no gateway. Physically, the IP phone connects to the network via an Ethernet connection (RJ-45). In a business environment, an Ethernet hub serves to concentrate VoIP phone lines, whereas in a legacy network an expensive PBX would be required. The advantage to VoIP service providers is that they need not maintain a Class 4 or Class 5 switch or VoIP gateway. In a legacy voice network, service providers must purchase and maintain large Class 4 and Class 5 switches and measure their investment in a cost per DS0 (single phone line) on a switch, which is further extrapolated into a cost per subscriber for the switch. In a VoIP network, it is generally assumed that the subscriber will purchase his or her own CPE. Hence, in such a next-generation network, the cost per DS0 for the service provider is $0.

SIP and Java programs also enable a whole new generation of applications that are impossible with circuit-switched telephony architectures. These applications can generally be divided into three categories: (1) personal productivity applications, (2) occupation-specific and industry-specific applications, and (3) *Web-telephony integration* (WTI) applications.

Most IP phones have LCD screens with GUIs, which allow expanded functions over a 12-button analog handset. With a conventional handset, the user must memorize long reams of number codes to perform functions such as conferencing, voice mail retrieval, call forwarding, and so on. For many users this presents a psychological barrier that limits them to using only a handful of the features available on a PBX, thus preventing them from being as efficient in their communications as they possibly could be. An IP phone with a GUI overcomes a number of these shortcomings by presenting the user with graphic choices to access their features.

A disadvantage to the IP phone was that the IP phones on the market were very expensive relative to a conventional handset. IP phones from Cisco, Nortel, or PingTel originally cost at least $500 as opposed to a conventional PBX-connected handset at about $150 per handset. That high cost makes this technology unattractive to the residential market. However, an IP handset that is competitively priced compared to feature-rich analog or digital handsets would probably be very popular and would further the growth of IP telephony. Price competition had driven the price of IP phones to about $100 by mid-2003.

In summary, the chief advantage of IP phones to a service provider is that they do not require the service provider to invest in a switch or gateway. In theory, the subscriber has covered that investment by buying the IP phone. Furthermore, service providers enjoy high margins on offering features, especially features not possible via circuit-switched telephones and networks. GUI interfaces on IP phones make these services easier to use, which can result in greater marketability of those services for the service provider. In short, the voice network of the not-so-distant future will consist of IP phones that connect to IP networks where the intelligence for that service is provided by a softswitch located anywhere on the network [3, p. 57].

Media Gateways (VoIP Gateway Switch)

The most successful commercial form of access has been the use of a VoIP gateway. The gateway provides a connection between an endpoint on a data network and the PSTN or switched circuit network. The gateway translates between transmission formats and the communication procedures that are used on each side. Gateways can be provided as stand-alone devices or integrated into other systems. In this form of access, an existing telephone handset interfaces a gateway either via a direct connection, through a PBX, or through a Class 5 switch. The gateway packetizes the voice and routes it over the IP network.

A media gateway can serve as both an access and limited switching device. The media gateway resides on the edge of a network and simply interfaces between TDM and IP networks. It is here that analog or digital signals from a handset or PBX Class 4 or Class 5 switch are digitized (if analog), packetized, and compressed for transmission over an IP network. In the reverse situation, incoming calls are translated and decompressed from an IP network for reception on digital or analog telephone devices [9]. The media gateway interfaces directly with a TDM switch (PBX, Class 4, or Class 5).

The design of a gateway includes three key elements: (1) an interface for the TDM side of the network (described in terms of DS0s or T1s), (2) an interface for the packet side of the network (usually an Ethernet connection), and (3) the necessary signal processing between these two sides. Signal processing is done on *digital signal processors* (DSPs) on circuit boards designed to support voice. Signal processing functions include echo cancellation, coding/decoding of the analog signal with algorithms discussed in Chapter 2 such as G.711 or G.723.1, and adapting the digitally encoded information into a series of IP datagrams and transmitting those datagrams via a network to their ultimate destination.

On softswitch architecture, a media gateway can also be part of the switching equation, depending on the amount of intelligence contained in the gateway. The trend is toward less intelligence in the media gateway and more intelligence in the softswitch. In the early days of the VoIP industry, the gateway had to contain a good deal of intelligence to make calls possible. However, the evolution of gatekeeper technology into carrier grade softswitch has drawn the intelligence out of the gateway and on to the softswitch.

Perhaps the most important issue for media gateways is their scalability. The density (number of DS0s or ports in one chassis) determines its classification. Depending on its density, a media gateway falls into one of the three following classifications: (1) residential or *small office/home office* (SOHO), (2) enterprise, and (3) carrier grade.

Residential or SOHO Gateways

Per Table 4.3, residential or SOHO gateways have two to eight ports for low-density operations.

Enterprise Gateways

Enterprise gateways aggregate the legacy telephone infrastructure for interfacing with VoIP networks. This is usually done by connecting a gateway to the trunk side of a legacy PBX. Users retain their existing handsets. This has the effect of making VoIP indistinguishable to the end user. The business need not train its staff to use new hardware or software because they are simply using their existing telephone handsets. This option also offers investment protection in that the business retains its expensive PBX and PBX-associated telephone handsets. The only thing that has changed is that the company reduces or eliminates interoffice phone bills.

Enterprise gateways are usually configured in multiples of T1 or E1 cards in a single chassis that interface with the trunk side of the PBX. The T1 trunks

Table 4.3
Gateways and Their Markets Based on Scalability

Gateway	Scale
Residential/SOHO	Two to eight ports (DS0s) linking telephone handsets via an RJ-11 connector
Enterprise	From two ports to multiple T1/E1s connecting to the PBX or directly to the handset
Carrier grade	High-density, multiple T1/E1 line side and OC-3 plus trunk side, 3,000 DS0s plus in one 7-foot rack

connect to the line side of the gateway. The trunk side of the gateway is its Ethernet connection to a router if a router is not built into the gateway.

Carrier Grade Gateways

The early applications for VoIP gateways were for international long-distance bypass and enterprise interoffice long distance. Success in these applications led to a demand for expanded density gateways for carrier operations. These gateways needed to be densely populated (have enough DS0s or ports) enough to interface with Class 4 and Class 5 switches (up to 100,000 DS0s in one node with OC-3 trunk side interface). These switches also had to offer the reliability of being able to interface with a circuit switch that boasted the "five 9s" of reliability (see Chapter 5).

Another requirement was that they be certified as being "NEBS 3" compliant, a requirement for any platform to be installed in a central office. The *Network Equipment Building Standards* (NEBS) publication [10] addresses the physical reliability of a switch. It is contained in Telcordia specification SR 3580, an extensive set of rigid performance, quality, safety, and environmental requirements applicable to network equipment installed in a carrier's central office. Nearly all major carriers in North America require that equipment in their central offices or switching locations undergo rigorous NEBS testing. Tests include electrical safety, immunity from electromagnetic emissions, lightning and power faulting, and bonding and grounding evaluations. Equipment must meet physical standards including successful temperature, humidity, and altitude testing, fire resistance (usually by destructive burning), earthquake vibration resistance, and a battery of other rigid tests. NEBS compliance also means having backup and disaster recovery strategies in place, including ensuring access to mirror sites, fire and waterproof storage facilities for critical databases, configuration backup information, and an *uninterruptible power supply* (UPS) to prevent network outages due to power failures [3, p. 49].

Switching in IP Networks

Softswitch (Gatekeeper, Media Gateway Controller)

A *softswitch* is the intelligence in a network that coordinates call control, signaling, and features that make a call across a network or multiple networks possible. Primarily a softswitch performs call control. Call control performs call setups and teardowns. Once a call is set up, connection control ensures that the call stays up until it is released by the originating or terminating user. Call control and service logic refer to the functions that process a call and offer telephone features. Examples of call control and service logic functions include recognizing that a party has gone off hook and that a dial tone should be provided, interpreting the dialed digits to determine where the call is to be terminated, determining if the called party is available or busy and, finally, recognizing when the called party answers the phone and when either party subsequently hangs up, and recording these actions for billing.

A softswitch coordinates the routing of signaling messages between networks. Signaling coordinates actions associated with a connection to the entity at the other end of the connection. To set up a call, a common protocol must be used that defines the information in the messages and which is intelligible at each end of the network and across dissimilar networks. The main types of signaling a softswitch performs are peer-to-peer signaling for call control and softswitch-to-gateway signaling for media control. For signaling, the predominant protocols are SIP, SS7, and H.323. For media control, the predominant signaling protocol is MGCP.

As a point of introduction to softswitch, it is necessary to clarify the evolution to softswitch and define the terms *media gateway controller* and *gatekeeper,* which were the precursors to softswitch. Media gateway controllers and gatekeepers (essentially synonymous terms for the earliest forms of softswitch) were designed to manage low-density (relative to a carrier grade solution) voice networks. MGCs communicate with both the signaling gateway and the media gateway to provide the necessary call processing functions. MGCs use either the MGCP or MEGACO/H.248 (described in a later chapter) for intergateway communications.

Gatekeeper technology evolved out of H.323 technology (a VoIP signaling protocol described in the next chapter). Because H.323 was designed for LANs, an H.323 gatekeeper can only manage activities in a zone. A zone is a collection of one or more gateways managed by a single gatekeeper. A gatekeeper should be thought of as a logical function and not a physical entity. The functions of a gatekeeper are address translation (i.e., a name or e-mail address for a terminal or gateway and a transport address) and admissions control (authorizes access to the network).

As VoIP networks got larger and more complex, management solutions with far greater intelligence became necessary. Greater call processing power became necessary as did the ability to interface signaling between IP networks with the PSTN (VoIP signaling protocols to SS7). Other drivers included the need to integrate features on the network and interface disparate VoIP protocols. Thus was born the softswitch.

A significant market driver for softswitch is the protocol intermediation necessary to interface, for example, H.323 and SIP networks. Another market driver for softswitch is the need to interface between the PSTN (SS7) and IP networks (SIP and H.323). Another function for softswitch is the intermediation between media gateways of dissimilar vendors. Despite emphasis on standards such as H.323, interoperability remains elusive. A softswitch application can overcome intermediation issues between media gateways. More information on VoIP protocols and signaling IP to PSTN is provided in later chapters.

The softswitch provides usage statistics to coordinate billing and to track operations and administrative functions of the platform while interfacing with an application server to deliver value-added subscriber services. The softswitch controls the number and type of features provided. It interfaces with the application server to coordinate features (conferencing, call forwarding, and so on) for a call.

Physically, a softswitch is software hosted on a server chassis filled with IP boards and includes the call control applications and drivers [3, p. 49]˙ Very simply, the more powerful the server, the more capable the softswitch. That server need not be colocated with other components of the softswitch architecture.

Signaling Gateway

Signaling gateways are used to terminate signaling links from the PSTN or other signaling points. The SS7 signaling gateway serves as a protocol mediator (translator) between the PSTN and IP networks. That is, when a call originates in an IP network using H.323 as a VoIP protocol and must terminate in the PSTN, a translation from the H.323 signaling protocol to SS7 is necessary in order to complete the call. Physically, the signaling function can be embedded directly into the media gateway controller or housed within a stand-alone gateway.

Application Server

The application server accommodates the service and feature applications made available to the service provider's customers. Examples include call forwarding, conferencing, voice mail, forward-on-busy, and so on. Physically an application server is a server loaded with a software suite that offers the application programs. The softswitch accesses these and enables and applies them to the appropriate subscribers as needed (Figure 4.6).

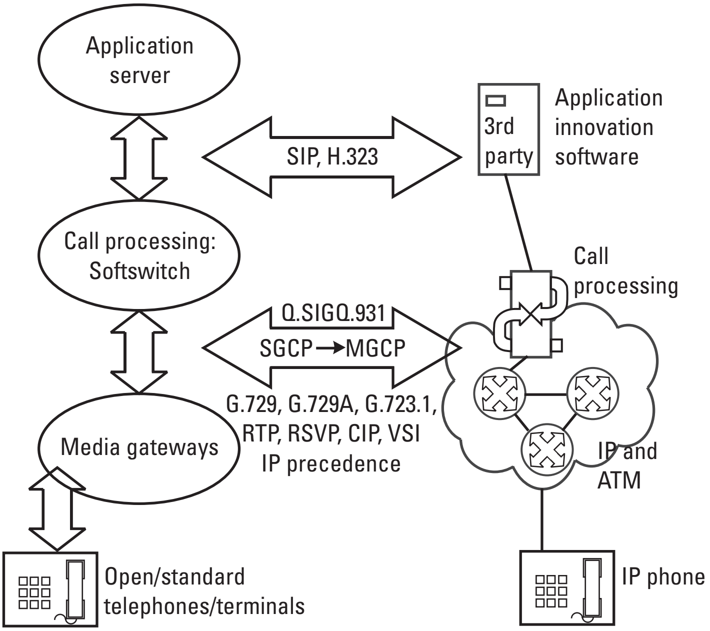

Application server
SIP, H.323
3rd party
Application innovation software
Call processing: Softswitch
Call processing
Q.SIGQ.931
SGCP→MGCP
G.729, G.729A, G.723.1, RTP, RSVP, CIP, VSI IP precedence
Media gateways
IP and ATM
Open/standard telephones/terminals
IP phone

Figure 4.6 Relationship of softswitch with other components. (*From:* [7]. © 1999 Cisco Systems, Inc. Reprinted with permission.)

A softswitch solution emphasizes open standards as opposed to the Class 4 or Class 5 switch, which historically offered a proprietary and closed environment. A carrier was a "Nortel shop" or a "Lucent shop." No components (hardware or software) from one vendor were compatible with products from another vendor. Any application or feature on a DMS-250, for example, had to be a Nortel product or specifically approved by Nortel. This usually translates into less than competitive pricing for those components. Softswitch open standards are aimed at freeing service providers from vendor dependence and the long and expensive service development cycles of legacy switch manufacturers.

Concern as to whether a softswitch solution can transmit a robust feature list identical to those found on a Nortel DMS-250, Class 4 switch, for example, is an objection among service providers. Softswitch offers the advantage of allowing a service provider to integrate third-party applications or even write their own while interoperating with the features of the PSTN via SS7. This is potentially the greatest advantage to a service provider presented by softswitch technology.

Features reside at the application layer in the softswitch architecture. The interface between the call control layer and specific applications is the *application program interface* (API). Writing and interfacing an application with the rest of the softswitch architecture occurs in the service creation environment. This is covered in greater detail in a following chapter.

Applications for Softswitch

IP PBX

Perhaps the earliest and most popular application for enterprise VoIP giving rise to the softswitch was the installation of a VoIP gateway on the trunk side of a PBX. This gateway packetized the voice stream and routed it over an IP network, which saved the business a lot of money in long-distance transport costs. This solution used the existing PBX's set of features (conferencing, call forwarding, and so on). It also provided "investment protection" to the user by leveraging the legacy PBX into a VoIP solution. The intelligence in this solution was contained in software known as the *gatekeeper*. The gatekeeper was the precursor to the softswitch.

Eventually, software developers devised a "soft" PBX, which could replace legacy PBXs. These "soft PBXs" (Figure 4.7) were considerably less expensive than a hardware PBX. They then came to be known as IP PBXs. An IP PBX can be thought of as an enterprise grade softswitch.

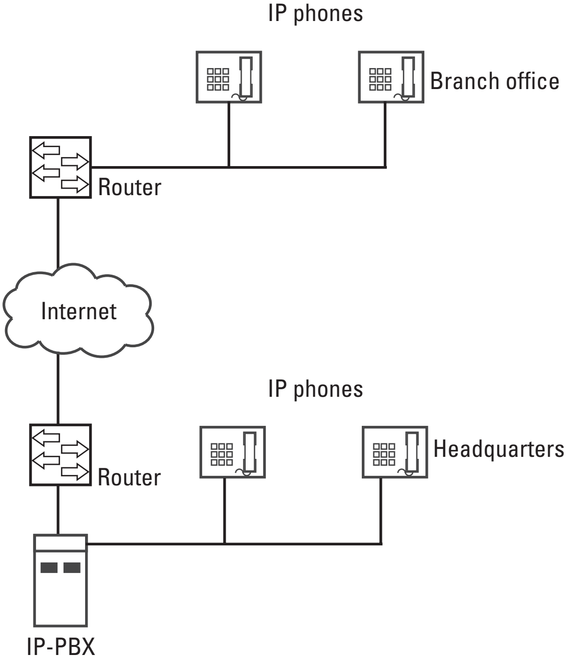


Figure 4.7 IP PBX, also known as "soft" PBX.

IP Centrex

Just as the Centrex model followed the PBX in circuit switching, it does the same in packet switching. Shortly after IP PBXs began to catch on in the market, the *regional Bell operating companies* (RBOCs) began to realize a threat to their circuit-switched Centrex services from VoIP applications. Centrex accounted for about 15% of all business lines and many subscribers were locked into 5-year contracts with the RBOCs. As these contracts began to expire in the late 1990s, many customers were actively evaluating less expensive alternatives to Centrex. If large companies could route their interoffice voice traffic over a corporate WAN using an IP PBX, what would be the demand for their circuit-switched Centrex services? With this threat in mind, IP Centrex services arrived on the market.

Centrex is a set of specialized business solutions (primarily, but not exclusively, for voice service) where the equipment providing the call control and service logic functions is owned and operated by the service provider and hence is located on the service provider's premises. Because Centrex frees the customer from the costs and responsibilities of major equipment ownership, Centrex can be thought of as an outsourcing solution.

In traditional Centrex service (i.e., analog Centrex and ISDN Centrex), call control and service logic reside in a Class 5 switch located in the CO. The Class 5 switch is also responsible for transporting and switching the electrical signals that carry the callers' speech or other information (e.g., faxes).

IP Centrex refers to IP telephony solutions where Centrex service is offered to a customer that transmits its voice calls to the network as packetized streams across an IP network. One benefit is increased utilization of access capacity. In IP Centrex, a single broadband access facility is used to carry the packetized voice streams for many simultaneous calls. In analog Centrex, one pair of copper wires is need to serve each analog telephone station, regardless of whether the phone has an active call; once the phone is not engaged in a call, the bandwidth capacity of those wires is unused. An ISDN BRI can support two simultaneous calls (i.e., 128 Kbps), but similar to analog lines, an idle BRI's bandwidth capacity cannot be used to increase the corporate LAN's interconnection speed.

IP Centrex Using Class 5 Switch Architecture

In this platform, existing Class 5 switches support IP Centrex service in addition to traditional POTS and ISDN lines. This is accomplished through the use of a media gateway (as described earlier in this chapter) at the CPE and a GR-303 gateway colocated with the Class 5 switch (Figure 4.8). The media gateway can be of any size from an IP phone to a carrier grade media gateway. The media gateway connects to the switch as if it were a digital loop carrier system. (Digital loop carriers use protocols such as GR-303 to deliver POTS and ISDN signaling information to switches for longer than average loops.) The GR-303 gateway

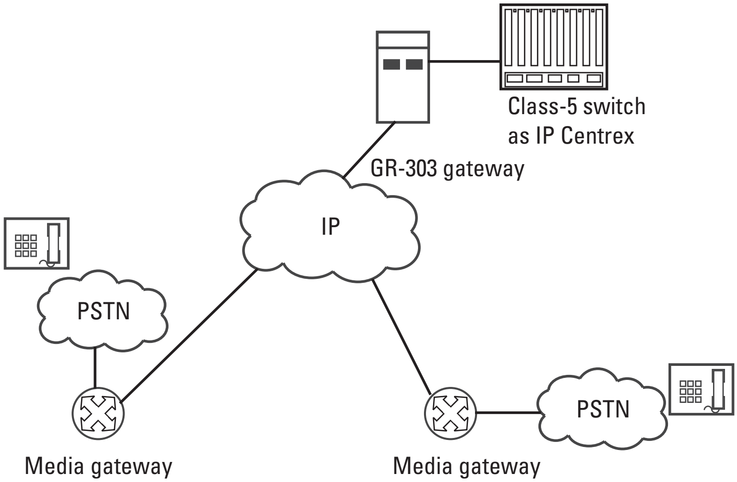


Figure 4.8 IP Centrex using a Class 5 switch with GR-303 interface.

translates any signaling information it receives from the customer's media gateway and depacketizes the voice stream for delivery to the switch. Similarly, it translates signaling messages from the switch into the IP telephony protocol (H.323, SIP, or MGCP) and packetizes the voice stream for transmission to the customer's media gateway. The customer's media gateway performs comparable functions for the standard telephone sets that it supports. As a result, the GR-303 gateway, customer's media gateway, and IP network connecting them appear to the Class 5 switch as an ordinary *digital loop carrier* (DLC) system, and the telephone sets connected to the customer gateway appear to the switch as ordinary phone lines. Because the IP Centrex solution is treated as a DLC system by the Class 5 switch, the switch is able to deliver the same features to IP Centrex users that it delivers to analog and ISDN Centrex users. Consequently, an extensive set of features is immediately available to IP Centrex users without needing to upgrade the Class 5 switch.

IP Centrex Using Softswitch Architecture

In a different approach to IP Centrex, the Class 5 switch is replaced by a softswitch (Figure 4.9). A softswitch is a telephony application running on a large, high-availability server in the network. Like the Class 5 switch, the softswitch provides call control and service logic. Unlike the Class 5 switch, the softswitch is not involved in transport or switching of the packetized voice stream. The softswitch and the IP Centrex CPE (customer media gateways and IP phones) signal one another over a packet network using an IP telephony protocol, such as H.323 or SIP.

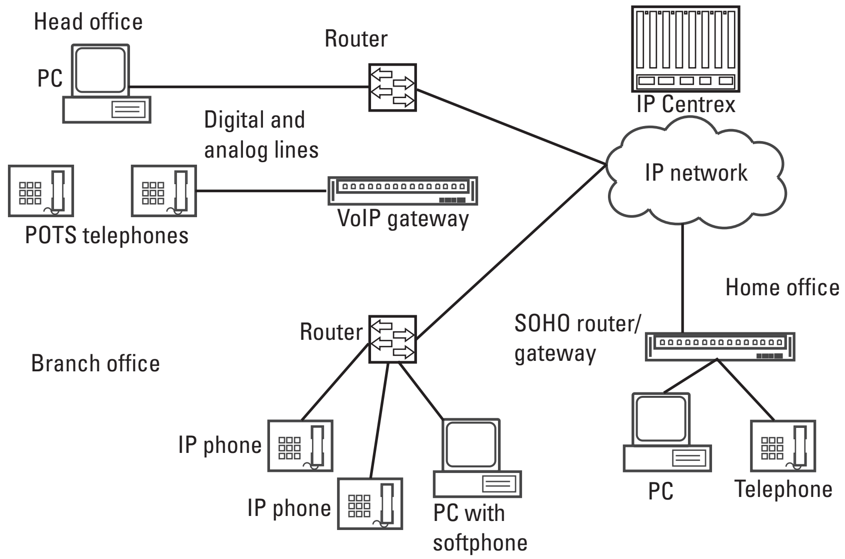


Figure 4.9 IP Centrex with softswitch. (*From:* [11]. © 2003 Artech House, Inc. Reprinted with permission.)

After it receives call setup information, the softswitch determines where the called party resides. If the called party is a member of the Centrex group, then the softswitch instructs the originating media gateway (or IP phone) and terminating media gateway (or IP phone) to route the packetized voice streams directly to one another; consequently, the voice stream never leaves the corporate LAN/WAN. If the called party is served by the PSTN, then the softswitch instructs the originating media gateway (or IP phone) to route the packetized voice stream to a trunking gateway. The trunking gateway has traditional interoffice facilities for Class 4 or Class 5 switches in the PSTN. The trunking gateway packetizes/depacketizes the voice stream so that it can be transmitted over these circuit-switched facilities. The trunking gateway works in conjunction with a signaling gateway. The signaling gateway is used to exchange SS7 messages with the PSTN. Both the trunking and signaling gateways receive their instructions from the softswitch [12].

Class 4 Replacement Softswitch

The next step in scale for the VoIP industry and tangentially the softswitch industry was Class 4 replacement. The origins of Class 4 replacement softswitch solutions lay in the long-distance bypass industry. Long-distance bypass operators used VoIP gateways for international transport. This technology allowed them to be very competitive relative to the "Big Three" long-distance companies. Part of that success was due to the fact that they were able to avoid paying into international settlements (described later in this book). Initially, these service providers used enterprise grade media gateways that interfaced with TDM switches in the PSTN. Technical challenges for these operators arose as their businesses flourished and demand grew. First, the media gateways were not dense enough for the levels of traffic they were handling. Second, the gateways that controlled these gateways were also limited in their ability to handle ever-increasing levels of traffic over these networks. Third, international traffic called for interfacing different national variants of SS7 signaling (each nation has its own variant).

In short, market demand dictated that a more scalable and intelligent solution be offered in the long-distance bypass industry. That solution came in the form of what is known as a Class 4 replacement softswitch solution comprised of more densely populated gateways managed with greater intelligence than a media gateway controller (Figure 4.10). The first applications involved installing a dense gateway on the trunk side of a Class 4 switch such as a Nortel DMS-250. As in the PBX scenario, the media gateway packetized the voice stream coming out of the Class 4 switch and routed it over an IP network, saving the service provider money on long-distance transport. The next step in the evolution of a Class 4 replacement softswitch was the removal of the circuit-switched Class 4 switch from that architecture. That is, the Class 5 switch connected directly to a media gateway, which routed the call over an IP network. The call control, signaling, and other features were controlled by a softswitch and the Class 4 switch was replaced in its entirety.

For the purposes of this book it is assumed that the arena of competition is similar to a scenario where Class 4 switches (DSM-250s from Nortel) are

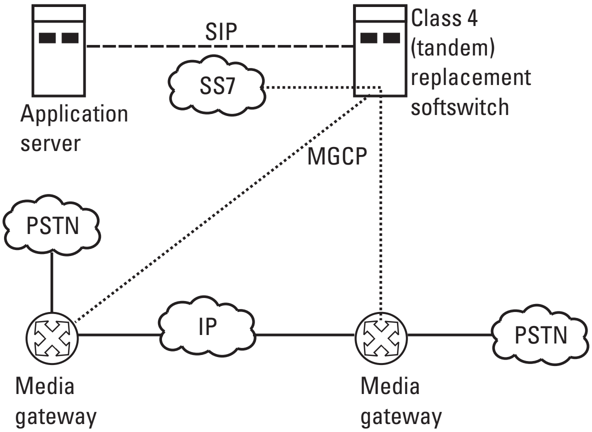


Figure 4.10 Class 4 replacement softswitch solution. Note absence of Class 4 TDM switches. (*After:* [13].)

connected to an IP backbone and long-distance traffic is transported via that IP backbone [3, p. 57]. At this service provider, softswitch, as a Class 4 replacement switch, competes directly with the Class 4 switch.

Class 5 Replacement Softswitch

The next level of progression in the development of softswitch technologies was the Class 5 replacement. This is the most exciting debate over softswitch. The ability of the softswitch industry to replace the Class 5 switch marks the final disruption of the legacy telecommunication infrastructure. A Class 5 switch can cost tens of millions of dollars and require at least one-half of a city block in real estate. The evolution of a successful Class 5 replacement softswitch has staggering implications for the world's local telephone service providers.

From the early days of the telephone industry, it was assumed that the cost of deploying local phone service with its copper pair access and local phone switches (most recently, a Class 5) would be so expensive that only a monopoly could effect this economy of scale and scope. Enter a Class 5 replacement softswitch (Figure 4.11) that does not cost tens of millions of dollars nor require a centrally located and very expensive CO and the barriers to entry and exit crumble. The result is that new market entrants may be able to effectively compete with quasimonopolistic incumbent service providers. This is potentially disruptive to incumbent local service providers and their Class 5 switch vendors.

Objections to a Class 5 replacement softswitch solution include the need for E911 and CALEA. This will be addressed in a later chapter. Another objection is the perception that softswitch cannot match Class 5 in features. A 5ESS Class 5 switch from Lucent Technologies is reported to have some 3,500 features that have been developed over a 25-year time frame. This features debate will be addressed in a later chapter. At the time of this writing, a number of successful Class 5 replacement softswitch installations have taken place and this segment of the industry is growing rapidly.

In summary, the softswitches that replace PBXs and Classes 4 and 5 switches (including Centrex) are differentiated in their scale, that is, by their processing power as measured by the number of busy hour call attempts or calls per second they can handle. Other differentiating factors include their ability to handle features from a feature server and to interface disparate signaling protocols. Softswitch is software that rides on a server. The limitations are the complexity of the software and the processing power of the server.

Conclusion

VoIP solutions replace their counterparts in the PSTN, enabling the PSTN to be bypassed in delivering voice services to subscribers. Many concepts deployed

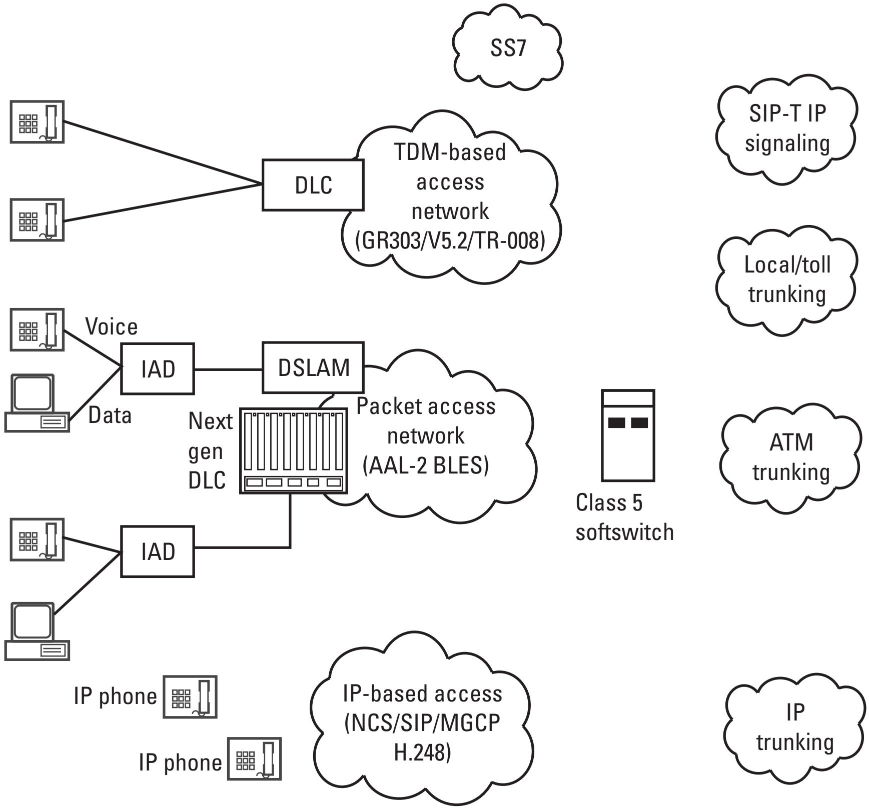


Figure 4.11 Class 5 replacement softswitch solution. (*After:* [14].)

in the PSTN have been translated into Vo802.11 networks including signaling and voice codecs. This chapter covered switching in both the PSTN as well as in VoIP networks. As this technology is replicated by startup technology providers and implemented by competitive service providers, competition to the local loop becomes possible. By avoiding the expense of millions of dollars for one Class 5 switch (an average city would require dozens of such switches), alternative service providers can enjoy lower barriers to entry in order to compete with incumbent service providers. Softswitches make bypass of the central office possible.

References

[1] Shepard, S., *Sonet/SDH Demystified,* New York: McGraw-Hill, 2001, pp. 15–21.

[2] Collins, D., *Carrier Grade Voice over IP,* 2nd ed., New York: McGraw-Hill, 2002.

[3] Ohrtman, F., *Softswitch: Architecture for VoIP,* New York: McGraw-Hill, 2002.

[4] "SS7 Tutorial," Performance Technologies, 2003, http://www.pt.com/tutorials/SS7.

[5] Isenberg, D., "Rise of the Stupid Network," *Computer Telephony,* August 1997, pp. 16–26; see also http://www.isen.com.

[6] Flynn, C., "Softswitches: The Brains Behind the Brawn," *Yankee Group,* May 2000, p. 3.

[7] Cisco Systems, "Cisco Multiservice Networking: Date, Voice, and Video Integration Strategy," presentation 0781_03F9_c1, 1999.

[8] PingTel, "Next-Gen VoIP Services and Applications Using SIP and Java," white paper, 2001, http://www.pingtel.com.

[9] International Softswitch Consortium, "Enhanced Service Framework," Applications Working Group, 2001, http://www.Softswitch.org.

[10] *Network Equipment Building Standards Requirements: Physical Protection,* Telecordia, GR-63-CORE, Piscataway, NJ, December 2002.

[11] Abrahams, J. R., and M. Lollo, *Centrex or PBX: The Impact of IP,* Norwood, MA: Artech House, 2003.

[12] "Softswitch Architecture," IP-Centrex.org, http://www.ip-centrex.org/how/index#softswitch, 2001.

[13] http://www.nuera.com/products/gxseries_diag.cfm and http://www.nuera.com/products/ssc_diag.cfm.

[14] http://www.santera.com/apps/class5.html and MetaSwitch, "NGN Migration Strategies," white paper, 2002, http://www.metaswitch.com/news/whitepapers.htm.

5

Objections to Vo802.11

To properly analyze the prospects for the use of Vo802.11 (given the variants of 802.11, this book will refer to 802.11 and not specify the many variants), it is necessary to categorize where potential weaknesses or objections may occur in such a network. Would potential degradations occur in the 802.11 segment of the network or in technologies related to VoIP? If so, where and how can those degradations be minimized or eliminated? Objections would focus on those related to 802.11 and VoIP.

Objections Related to 802.11

Detractors to IEEE 802.11 state that the technology will not achieve popular acceptance because it is limited in range, security, and QoS. As with any other technology, the market constantly strives to overcome these objections with improvements in 802.11.

The position that wireless technologies will replace the PSTN meets with a number of objections. Primarily, these objections are focused on the QoS issues, security of the wireless network, and limitations in the range of the delivery of the service.

QoS

One of the primary concerns about wireless data delivery is that, like the Internet over wired services, the QoS is inadequate. Contention with other wireless services, lost packets, and atmospheric interference are recurring objections to

802.11b and associated wireless protocols as an alternative to the PSTN (Figure 5.1). QoS is also related to the ability of a service provider to accommodate voice on its network. The PSTN cannot be replaced until there is an alternative, competent replacement for voice over copper wire.

Security

The press has been quick to report on weaknesses found in wireless networks. The 802.11b network has two basic security mechanisms built into it. They are *Service Set ID* (SSID) and *Wireless Equivalency Protocol* (WEP). These measures may be adequate for residences and small businesses but inadequate for enterprises that require stronger security. A number of measures can be added, however, to those wireless networks that will provide the necessary level of security for the subscriber.

Range

In most omnidirectional applications, 802.11 offers a range of about 100m. So how, one might ask, will that technology offer the range to compete with the PSTN? Range is a function of antenna design and power, but mostly antenna design. With the right antenna and power, the range of 802.11 is extended to tens of miles [1].

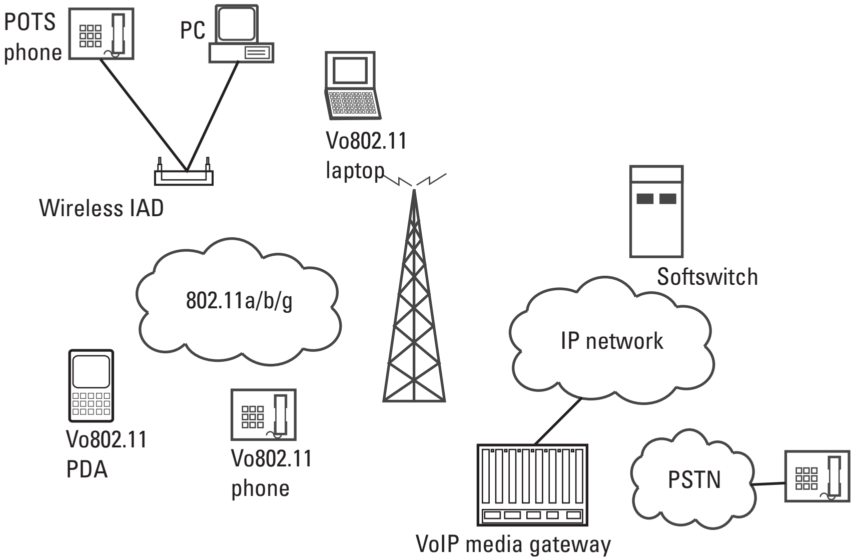


Figure 5.1 Overview of a broadband wireless alternative to the PSTN.

Objections Related to Voice over IP

Reliability

The chief concern service providers have when comparing competitive technology to the PSTN's Class 4 and Class 5 switches is reliability. Class 4 and Class 5 switches have a reputation for the "five 9s" of reliability. That is, they will be out of service only 5 minutes in 1 year. Engineering a voice switching solution to achieve "five 9s" is neither black magic nor a mandate from heaven on golden tablets. It is a matter of meticulously engineering into the solution the elements of redundancy, no single point of failure, and NEBS to a point where, when figuring in planned downtime, the solution has 5 minutes or less of downtime per year. Many softswitch solutions now offer "five 9s" or better reliability.

Scalability

Of secondary importance to service providers is the scalability of a softswitch relative to a Class 4 or Class 5 switch. To compete with a Class 4 or Class 5 switch, a softswitch solution must scale up to tens of thousands (phone lines or ports) in one location. Softswitch solutions, by virtue of new, high-density media gateways, now match or exceed 24,000 DS0s in one 7-foot rack as opposed to the nine racks it takes a Class 4 or Class 5 switch to make 24,000 DS0s. In addition, softswitch platforms now offer call processing power in terms of *busy hour call attempts* (BHCAs) in the millions as opposed to the hundreds of thousands offered by legacy switching platforms. One significant advantage of softswitch solutions over Class 4 and Class 5 switches with regard to scalability is that they can scale down to as little as two port media gateways or even one port in the case of IP handsets, allowing unlimited flexibility in deployment. The minimum configuration for a Class 4 switch, for example, is 480 DS0s.

QoS

Early VOIP applications garnered a reputation for poor quality of service. First available in 1995, these applications were often characterized by using PCs with microphones and speakers over the public Internet. The calls were often dropped and the voice quality was questionable. Vast improvements in IP networks during the last 7 years, coupled with advances in media gateway technologies, now deliver voice quality that matches or exceeds that delivered via Class 4 and Class 5 switches over the PSTN.

Signaling

An element of the PSTN that was designed to deliver good QoS and thousands of features is SS7. The interfacing of SS7 and IP networks necessary to deliver calls that travel over both the PSTN and an IP network is a significant challenge. Much progress has been made, including the emergence of a new technology that is roughly the equivalent of SS7 designed to operate with IP networks known as *SigTran.* In addition, the VoIP industry has new protocols such as SIP that match or exceed SS7 in signaling capabilities.

Features and Applications

Many proponents of the PSTN dismiss VoIP and softswitch solutions with the interrogatory "Where are the 3,500 5ESS features?" referring to Lucent Technologies' 5ESS Class 5 switch, which is reported to have approximately 3,500 calling features. An interrogatory to Lucent Technologies did not produce a list of what each of those 3,500 features is or does. It is doubtful that each and every one of those 3,500 features is necessary to the successful operation of a competitive voice service. Telcos that require new features must contract with the switch vendor (in North America that is Lucent Technologies in 90% of the Class 5 market) to obtain new features. Obtaining those new features from the switch vendor requires months if not years of development and hundreds of thousands of dollars.

Softswitch solutions are often based on open standards and use software applications such as *Voice XML* (VXML) to write new features. Service providers using softswitch solutions can often write their own features in house in a matter of days. Service providers can also obtain new features from third-party software vendors. Given this ease and economy of developing new features, the question arises: Why limit yourself to a mere 3,500 features? Why not 35,000 or more features?

This ease and flexibility in deploying new features in a softswitch solution offer a service provider the ability to quickly deploy high-margin features that generate revenues not possible with Class 4 or Class 5 switches. In a net present value calculation, a softswitch solution, given its lower cost of acquisition and operation coupled with an ability to generate greater revenues, will win over a Class 4 or Class 5 solution [2].

Conclusion

In order for Vo802.11 applications to reach widespread commercial acceptance, it will have to be clear to decision makers that the technology is sound and that objections to the technology are easily overcome.

References

[1] Ohrtman, F., *Softswitch: Architecture for VoIP,* New York: McGraw-Hill, 2002, pp. 6–7.

[2] Ohrtman, F., *Wi-Fi Handbook: Building 802.11b Wireless Networks,* New York: McGraw-Hill, 2003, pp. 8–9.

6

Vo802.11: Range Is a Matter of Engineering

One of the major misperceptions regarding 802.11b and other wireless protocols is that the range is limited to 100m and thus proves impractical as a last mile solution. The truth is that with proper engineering, 802.11b can reach beyond 20 miles from point to point. In the quest for PSTN bypass, this is one of the most exciting developments. By steering an antenna in the direction of the subscriber's home, the service provider can bring broadband wireless to masses of homes without so much as stringing a single strand of copper wire, digging up a single street, or engaging in a single legal battle for right of way.

How can a service provider cover a residential market with access points that have a maximum range of 100m? Such a scenario would result in a service that is not economically viable due to the limited range of the 802.11 infrastructure. If a Vo802.11 service is to be economically viable, its infrastructure must have the access points (radios and antennas) that have a range far greater than 100m. Such products are coming on the market; in fact, some products now cover several square miles. By utilizing the infrastructure to get greater coverage, more subscribers can be serviced per access point, making any Vo802.11 service more economically viable if not profitable.

Furthermore, new wireless protocols for MANs provide for the construction of wireless networks that can cover whole cities. Ad hoc peer-to-peer networks stretch the range of a wireless network with a minimum of investment. Some power line communications solutions use power lines to deliver Wi-Fi.

This chapter first covers the science of antennas and how proper engineering can stretch the most modest resources to deliver essential services to the home. This chapter then explains how 802.11b antenna systems can be used to stretch the range of delivery out to a number of miles so as to blanket large metropolitan areas and even reach out to rural subscribers. Most important in designing a broadband wireless network is the inclusion of a new protocol, 802.16, in the deployment of wireless MANs to feed suburban 802.11b networks. Other technologies such as mesh networks also extend the range of broadband wireless networks.

In data networking, the success of 802.11 has inexorably linked it with RF engineering. Where a wired network requires little or no knowledge on the part of the installer about how data travel via an Ethernet cable, wireless requires a strong knowledge of radios and antennas.

RF systems complement wired networks by extending them. Different components may be used depending on the frequency and the distance that signals are required to reach, but all systems are fundamentally the same and made from a relatively small number of components. Three RF components of particular interest to 802.11 users are antennas, sensitive receivers, and amplifiers. The following paragraphs provide a basic overview of wireless transmission systems, with antennas being of particular interest because they are the most tangible feature of an RF system.

Antennas

Antennas are the most critical component of any RF system because they convert electrical signals on wires into radio waves and vice versa. To function at all, an antenna must be made of conducting material. Radio waves hitting an antenna cause electrons to flow in the conductor and create a current. Equally, applying a current to an antenna creates an electric field around the antenna. As the current to the antenna changes, so does the electric field. A changing electric field causes a magnetic field, and the wave is off.

The size of the antenna you need depends on the frequency: the higher the frequency, the smaller the antenna. The shortest simple antenna possible at any frequency is one-half wavelength long. This rule of thumb accounts for the huge size of radio broadcast antennas and the small size of mobile phones. An AM station broadcasting at 830 kHz at a wavelength of about 360m and has a correspondingly large antenna, but an 802.11b network interface operating in the 2.4-GHz band has a wavelength of just 12.5 cm. With some engineering tricks, an antenna can be incorporated into a PC card or the top of a laptop computer.

Antennas can also be designed with directional preference. Many antennas are omnidirectional, which means they send and receive signals from any direction. Some applications may benefit from directional antennas, which radiate and receive on a narrower portion of the field. Figure 6.1 compares the radiated power of omnidirectional and directional antennas.

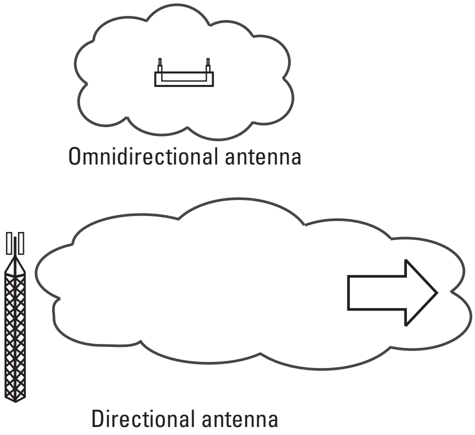


Figure 6.1 Radiated power and reach of antennas: omnidirectional and directional.

For a given amount of input power, a directional antenna can reach farther with a clearer signal. The antenna must also have much a higher sensitivity to radio signals in the dominant direction. When wireless links are used to replace wire-line networks, directional antennas are often used. Mobile telephone network operators also use directional antennas when cells are subdivided. The 802.11 networks typically use omnidirectional antennas for both ends of the connection.

Antennas are the most likely to be separated from the rest of the electronics. A transmission line (some kind of cable) between the antenna and the transceiver is also necessary. Transmission lines usually have an impedance of 50 ohms. In terms of practical antennas for 802.11 devices in the 2.4-GHz band, the typical wireless PC card has an antenna built in. The antenna plugs into the card.

Wireless cards all have built-in antennas, but these antennas are, at best, minimally adequate. If you were planning to cover an office or an even larger area, such as a campus you will almost certainly want to use external antennas for your access points [1, pp. 42–43].

Factors Affecting Range

It is tempting to think that you can put up a high-gain antenna and a power amplifier and cover a huge amount of territory, thus economizing on access points and serving a large number of users at once. This is not, however, a particularly good idea. The larger the area you cover, and the more users located in that area, the more users your access points must serve. Twenty to 30 users per access points is a good upper bound. A single access point covering a large territory may look like a good idea, and it may even work well while the number of users remains small. But if a network is successful, the number of users will grow quickly, and the network will soon exceed the access point's capacity [2, pp. 316–322]. Once this happens, it is necessary to install more access points and divide the original cell into several smaller ones and lower the power output at all of the cells.

Sensitive Receivers

Besides antennas, the most critical item in a Wi-Fi system is the receiver. In particular, it is important to look for receiver sensitivity. The receiver sensitivity is the lowest level signal that can be decoded by the receiver. The lower the receive sensitivity, the longer the range.

Amplifiers make signals bigger. Signal boost, or gain, is measured in decibels (dB). Amplifiers can be broadly classified into three categories: low noise, high power, and everything else. *Low-noise amplifiers* (LNAs) are usually connected to an antenna to boost the received signal to a level that is recognizable by the electronics to which the RF system is connected. LNAs are also rated with a noise figure, which is the measure of how much extraneous information the amplifier introduces to the signal-to-noise ratio. A smaller noise figure allows the receiver to hear smaller signals and thus allow for a greater range.

Amplifiers

High-power amplifiers (HPAs) are used to boost a signal to the maximum power possible before transmission. Output power is measured in dBm, which are related to watts. Amplifiers are subject to the laws of thermodynamics: They give off heat in addition to amplifying the signal. The transmitter in an 802.11 PC card is necessarily low power because it needs to run off a battery if it is installed in a laptop, but it is possible to install an external amplifier at feed access points, which can be connected to the power grid where power is more plentiful. This is where things can get tricky with respect to compliance with regulations. The 802.11 devices are limited to 1W of power output and 4W *effective radiated power* (ERP). ERP multiplies the transmitter's power output by the gain of the antenna minus the loss in the transmission line. With a 1W amplifier, an antenna that gives you 8 dB of gain, and 2 dB of transmission line loss, the result is an ERP of 4W; the total system gain is 6 dB, which multiplies the transmitter's power by a factor of 4 [1, pp. 44–46].

The 802.11b Network at 20 to 72 Miles

Point-to-multipoint links in excess of 1,500 feet with ordinary equipment at the client side are very possible. Using high-gain antennas, sensitive receivers, and amplifiers if necessary, it is possible to achieve Ethernet-like speeds over 20+-mile point-to-point links (Figure 6.2). An experiment proved that it is theoretically possible to drive 802.11b signals well over 20 miles, using stock equipment [3]. In fact, a 72-mile link from San Diego to San Clemente Island has been established by Hans Werner-Braun [4] with some specialized 802.11 equipment on the 2.4-GHz band.

In summary, 802.11b, by itself, is *not* limited to a range of 100m. Its maximum range is in excess of 20 miles. A comparison is that of the telephone central office, where the maximum range of the signal over copper wire is 18,000 feet or 3 miles without a repeater. It could be argued that the maximum, unboosted range of 802.11b exceeds that of the PSTN.

Architecture: The Large Network Solution

While a point-to-point 802.11b connection may have a range of 20 miles, and a point-to-multipoint connection that is somewhat shorter, building a wireless network to compete with the PSTN is considerably more complicated. Issues revolving around bandwidth sharing and frequency contention require a multitiered strategy for building a *wireless metropolitan-area network* (WMAN) to replace the PSTN in a given municipality.

Overcoming limitations of range can be achieved through proper architectural planning of a wireless network. Four elements of network architecture can be employed to extend the maximum range of 802.11b and its associated wireless protocols to cover an entire metropolitan area. First, a WMAN is fed from an IP backbone at a high bandwidth, say, 100 Mbps. This WMAN would operate at a licensed frequency to ensure a high quality of transmission devoid of interference. The chief subscribers of the WMAN would be *wireless Internet service providers* (WISPs). The WMAN would then feed lesser networks, the

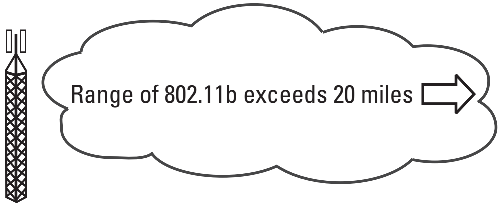


Figure 6.2 The range of 802.11b exceeds 20 miles. (*From:* [5]. © 2001 O'Reilly & Associates, Inc. Reprinted with permission.)

wireless wide-area networks (WWANs). The WWANs could operate at the 802.11a bandwidth (54 Mbps) at a frequency in the 5.8-GHz range. Subscribers of the WWAN would include large enterprises and smaller WISPs. The WWAN would, in turn, feed WLANs. WLANs would feed residences and small businesses. *Wireless personal area networks* (WPANs) would feed off WLANs to serve components within a given residence (Figure 6.3). Finally, an ad hoc peer-to-peer network, consisting of subscriber devices, intelligent access points, and wireless routers can extend the network even further with little infrastructure cost.

MANs

The WMAN encompasses a range of radio- and laser-based technologies targeted at providing wireless networking over distances of a few hundred meters to several miles. Wireless broadband, *broadband wireless access* (BWA), *wireless local loop* (WLL), fixed wireless, and wireless cable all refer to technologies for delivering telecommunications services over the last few miles of the network. *Wireless broadband* and *BWA* are general terms referring to high-speed wireless networking systems. WLL is derived from the wired telephony term *local loop,* which refers to the connection between a local telephone switch and a subscriber. WLL

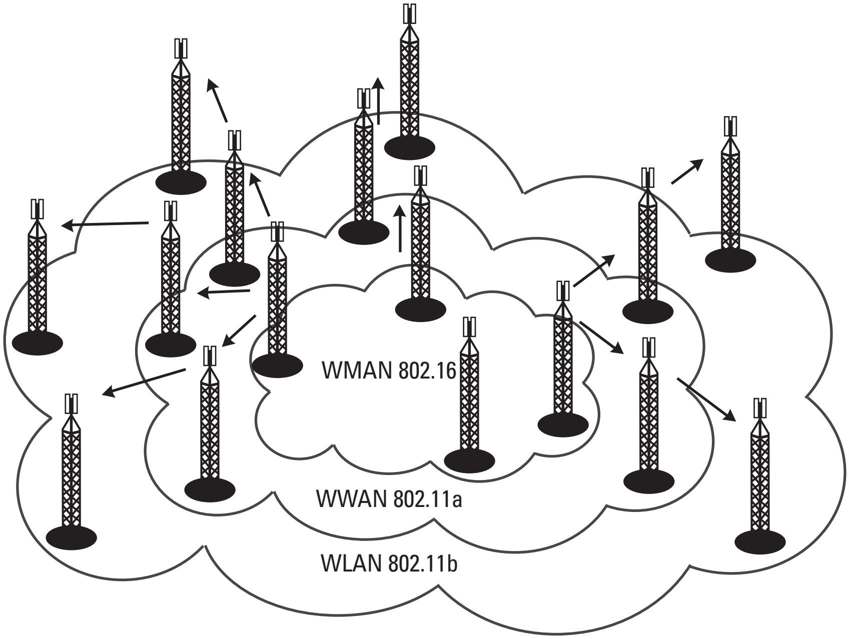


Figure 6.3 Covering a metropolitan area with WMANs, WWANs, WLANs, and WPANs.

and fixed wireless generally refer to the delivery of voice and data services between fixed locations over a high-speed wireless medium. Some new market entrants offer mobile applications of this technology. Fixed wireless includes *local multipoint distribution service* (LMDS), *multichannel multipoint distribution service* (MMDS), U-NII systems, and similar networks. Wireless cable usually refers to MMDS systems used to deliver television signals such as the *instructional television fixed service* (ITFS).

Two basic network topologies are supported by these systems. The simplest is a point-to-point system providing a high-speed wireless connection between two fixed locations. Bandwidth is not shared, but links typically require line of sight between the two antennas. The second topology is a point-to-multipoint network in which a signal is broadcast over an area (called a *cell*) and communicates with fixed subscriber antennas in the cell. Because bandwidth in the cell is finite and is shared among all users, performance may be a concern in high-density cells. Systems of different frequencies may be combined to cover an area where terrain or other obstructions prevent full coverage.

Other than frequency, the main difference between fixed wireless systems and cellular, WLAN, and WPAN networks is the mobility subscriber equipment. There has been some discussion about adding support for mobile subscriber equipment to fixed wireless systems. The addition of mobility support would enable these BWA systems to potentially function as *fourth generation* (4G) cellular networks, delivering subscriber speeds of several megabits. Several technical, regulatory, and commercial hurdles remain to be overcome before this could become a reality, but companies such as Wi-Fi have already started examining products targeted at this potential application.

LMDS

LMDS is a fixed wireless, radio-based technology. In North America, LMDS operates in the 28- to 31-GHz frequency range, but may operate anywhere from 2 to 40 GHz in other regions. In 1998, the FCC held an auction for this spectrum, dividing each geographic A Block and B Block. The A Block had a bandwidth of 1.5 GHz and the B Block had a bandwidth of 150 MHz. The intent was for the auction winners to deploy high-speed voice and data communications services in the last mile. The realities of deployment have not yet lived up to that vision.

The network topology of LMDS uses a central transmitter sending its signal over a cell with a radius of 5 km or less. Antennas are usually placed on rooftops for line of sight to the central transmitter. This is because *first generation* (1G) LMDS equipment uses radio technology that is affected by hills, walls, trees, and other physical barriers. This limitation may be reduced as equipment starts to adopt more advanced spectrum utilization techniques such as OFDM.

As a high-frequency outdoor radio technology, LMDS performance and range will vary depending on weather conditions. It has a range of less than 5 km and supports gigabit speeds, although services are usually offered at a much lower rate. The physics of the 30-GHz signal make it about a millimeter in length; this spectrum is sometimes referred to as the *millimeter-wave spectrum.* One effect of having such a small wavelength is that rain can effectively block the signal. In areas where rain is a factor, a lower frequency is required. A higher frequency allows faster data rates, but it also limits range, requiring more equipment to cover the same area as a lower frequency technology. LMDS bandwidth in a specific area is shared among all the users like cable. To ensure end-user performance, networks must be built with excess capacity to handle sporadic peak loads and unexpected growth in the subscriber base. In addition, there are no standards governing LMDS implementations, leading to a number of incompatible proprietary solutions. Higher network deployment costs make 1G LMDS networks more suitable for high-margin business applications rather than residential use [6, pp. 56–59].

802.16: Protocol for WMANs

An 802.16 wireless service provides a communications path between a subscriber site and a core network (the network to which 802.16 is providing access). Examples of a core network are the public telephone network and the Internet. IEEE 802.16 standards are concerned with the air interface between a subscriber's transceiver station and a base transceiver station.

Protocols defined specifically for wireless transmission address issues related to the transmission of blocks of data over a network. The standards are organized into a three-layer architecture. The lowest layer, the physical layer, specifies the frequency band, the modulation scheme, error-correction techniques, synchronization between transmitter and receiver, data rate, and the TDM structure [7].

IEEE 802.16 addresses "first-mile" applications of wireless technology to link commercial and residential buildings to high-rate core networks and thereby provide access to those networks. The 802.16 group's work has primarily aimed at a point-to-multipoint topology with a cellular deployment of base stations, each tied to core networks and in contact with fixed wireless subscriber stations.

Working Group 802.16 is now completing a draft of the IEEE-802.16 Standard Air Interface for Fixed Broadband Wireless Access Systems. The document includes a flexible MAClayer. The accompanying PHY layer is designed for 10 to 66 GHz, informally known as the LMDS spectrum. The standard is not yet final, but the draft is stable and has passed the working group's letter ballot, pending resolution of comments proposed to improve it [8].

For transmission from subscribers to a base station, the standard uses the *Demand Assignment Multiple Access–Time-Division Multiple Access* (DAMA-

TDMA) technique. DAMA is a capacity assignment technique that adapts as needed to respond to demand changes among multiple stations. TDMA is the technique of dividing time on a channel into a sequence of frames, each consisting of a number of slots, and allocating one or more slots per frame to form a logical channel.

With DAMA-TDMA, the assignment of slots to channels varies dynamically. 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. Both are TDM schemes.

Above the physical layer are the functions associated with providing service to subscribers. These functions include transmitting data in frames and controlling access to the shared wireless medium, and are grouped into the MAC layer. The MAC protocol defines how and when a base station or subscriber station may initiate transmission on the channel. Because some of the layers above the MAC layer, such as ATM, require quality of service, the MAC protocol must be able to allocate radio channel capacity to satisfy service demands.

In the downstream direction (base station to subscriber stations), there is only one transmitter, and the MAC protocol is relatively simple. In the upstream direction, multiple subscriber stations compete for access, resulting in a more complex MAC protocol. In both directions, a TDMA technique is used, in which the data stream is divided into a number of time slots.

The sequence of time slots across multiple TDMA frames that is dedicated to one subscriber forms a logical channel, and MAC frames are transmitted over that logical channel. IEEE 802.16.1 is intended to support individual channel data rates of from 2 to 155 Mbps.

Above the MAC layer is a convergence layer that provides functions specific to the service being provided. For IEEE 802.16.1, bearer services include digital audio/video multicast, digital telephony, ATM, Internet access, wireless trunks in telephone networks, and frame relay [8].

Consecutive Point Network

In a WMAN, reliability of the network can be ensured by implementing *consecutive point network* (CPN) technology (Figure 6.4). Like a SONET fiber ring, the data flow of the network around the wireless ring would reverse flow in the event of a disruption in the network. This ensures that only a limited part of the network is down due to a disruption.

Extending Range Via an Ad Hoc Peer-to-Peer Network

Ad hoc peer-to-peer technologies extend the maximum range of Wi-Fi networks from distances typically measured in hundreds of feet to several miles

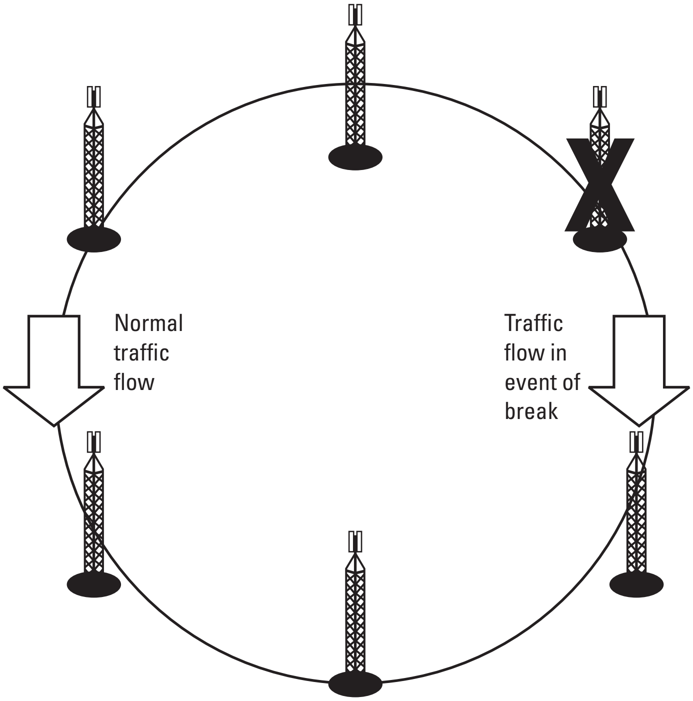


Figure 6.4 Consecutive point networks. Note that like a SONET ring, the data flow reverses itself in case of a break in the network.

(Figure 6.5). The product adds multihopping peer-to-peer capabilities to off-the-shelf 802.11 cards.

Software is utilized to turn wireless LAN cards into router-repeaters. The result is a system that enables users who are out of range of an access point to hop through one or more other nearby users until they connect to the access point. The software also automatically routes transmissions from congested access points to uncongested ones. Overall network performance is enhanced in addition to the dramatic increases in effective range. In addition, users within range of each other form a network albeit one without a connection to a larger network or the Internet.

Peer-to-peer mode is one part of the 802.11 standard. Most WLANs are operating in the infrastructure mode, in which multiple users independently connect to access points. This method severely limits the useful range of the network, forcing network administrators to add multiple access points to create an extended coverage area. The software uses the peer-to-peer capabilities included in every 802.11 card to achieve increased network coverage by making all card users a potential part of the transmission network [9].

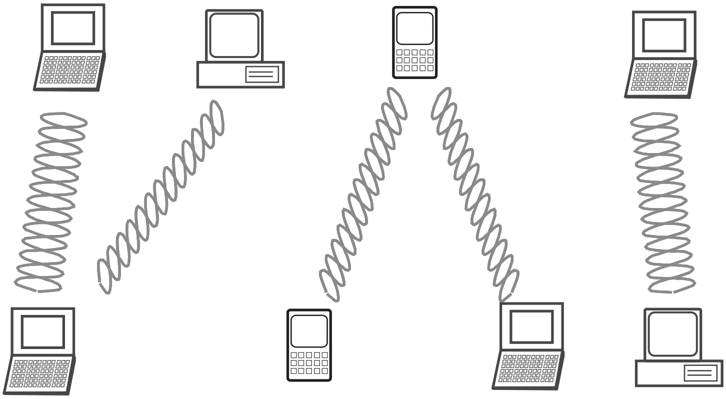


Figure 6.5 Ad hoc peer-to-peer network.

Network Features and Products

Traditional wireless solutions typically attempt to create a mobile broadband network by overlaying some IP equipment onto a circuit-switched, voice-centric system. An ad hoc peer-to-peer network offers an end-to-end IP-based, packet-switched, mesh architecture that mirrors the wired Internet's architecture and its resulting advantages. In peer-to-peer technology, the users are the network in that they add mobile routers and repeaters (or picocells) to the network infrastructure.

Because users carry much of the network with them, network capacity and coverage is dynamically shifted to accommodate changing user patterns. As people congregate and create pockets of high demand, they also create additional routes for each other, thus enabling access to network capacity through neighboring access points via multihopping. Users will automatically hop away from congested routes and access points to less congested routes and network access points. This permits the network to dynamically and automatically balance capacity and increase network utilization.

Advantages of Ad Hoc Peer-to-Peer Networks

Ad hoc peer-to-peer networks offer a number of exciting advantages for new market entrants or municipally owned and operated networks. First, the permanent, fixed components such as access points and wireless routers are small and unobtrusive relative to the cell towers found in *third generation* (3G) architectures. This presents the advantage of much less expensive deployment both in terms of physical plant and legal issues (leasing roof rights, for example). The time needed to deploy service in a given market is also greatly reduced.

Second, when enough subscriber devices are present in a given area, the reach of the network is instantly and inexpensively increased. By virtue of using a subscriber device as a router or repeater, the service provider is spared the expense of access points and wireless routers. Furthermore, a network is established among subscriber devices where there is no IAP or wireless router to connect the subscriber devices to the Internet or other networks.

In an ad hoc peer-to-peer mobile architecture, all nodes in the network, including subscriber devices, act as routers and repeaters for other subscribers in the network. This enables users to hop between any number of devices in the network to achieve the desired connection. As a result, the network allows user devices to act as wireless routers and repeaters for other users. This increases network robustness, while reducing infrastructure deployment costs. Ad hoc peer-to-peer networks make it easy for two people to directly share files, e-mail, music, video, or voice calls. Network infrastructure is not needed. Therefore, users can form high-speed voice and data networks anywhere, anytime. Instead of wireless operators subsidizing the cost of user devices (handsets, for example), users actually subsidize and help deploy the network for the operator.

Components of an Ad Hoc Peer-to-Peer Network

The network is comprised of the following elements: subscriber devices (including PDAs, laptops, mobile phones, automobiles, and so on), wireless routers, and APs. Subscriber devices can be either mobile or fixed, while the remaining elements are fixed. Wireless routers and APs can be mounted on utility poles, billboards, buildings, or any other convenient structure. It is important to note that the transceiver and modem technology within a subscriber device is identical to the transceiver technology in the fixed infrastructure. This keeps subscriber and infrastructure costs exceptionally low.

When enough subscriber devices are present in a given area, the reach of network is instantly and inexpensively increased. By virtue of using a subscriber device as a router or repeater, the service provider is spared expense of APs and wireless routers. Furthermore, a network is established among subscriber devices where there is no AP or wireless router to connect the subscriber devices to the Internet or other networks [10].

Conclusion: Range Is Not an Issue

The common misperception about 802.11b was that its maximum range was limited to 100m. With proper engineering it can reach 20 miles point to point. The rollout of the 802.16 MAN protocol allows the extension of 802.11b and associated wireless protocols over a wide geographic area. By stepping down from a MAN to lower bandwidth networks, wireless networks can reach out to residential markets and other low-density markets. Ad hoc peer-to-peer networks, by virtue of not requiring expensive infrastructure, are perhaps the most cost effective means of extending a wireless network. This has the potential to extend the network even further. Here, the subscribers are the network (Figure 6.6).

Wi-Fi systems act like small routers, with each node relaying to its nearest neighbors. Messages hop, in a peer-to-peer manner, across a broad interconnected nexus. This produces a broadband telecommunications system, built by separate, independent, interconnecting service providers with each other for their common good.

Two things make this peer-to-peer structure so interesting. First, its emergence and growth are viral. *Viral telecommunications* is a truly new, bottom-up phenomenon. In the face of a downward-falling telecom market and tight capital spending, this has further appeal. Second, its performance increases with the number of nodes. Metcalf's law states that the value of a network increases exponentially with the addition of every new node. In this topology, more nodes equals better service. By empowering the subscribers to "be" the network, the cost to the service providers is drastically reduced. The network could also be communally owned [11].

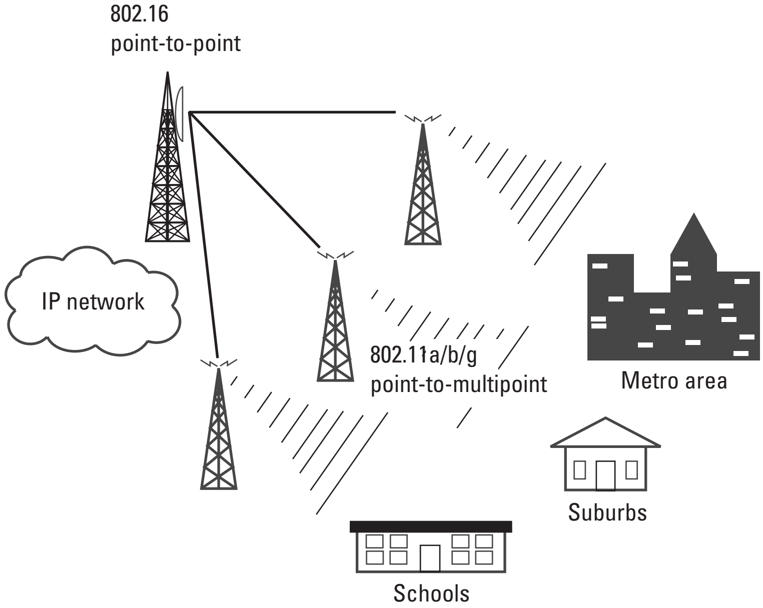


Figure 6.6 Extending the range of wireless data transmission via its architecture.

References

[1] Ohrtman, F., *Wi-Fi Handbook: Building 802.11b Wireless Networks,* New York: McGraw-Hill, 2003.

[2] Gast, M., *802.11b Wireless Networks: The Definitive Guide,* Sebastopol, CA: O'Reilly & Associates, 2002.

[3] Flickenger, R., "A Wireless Long Shot," *O'Reilly Network*, May 3, 2001, http://www.oreillynet.com/pub/a/wireless/2001/05/03/longshot.html.

[4] Brewin, B., "San Diego Wireless Net Installs 72-Mile Link," *Computer World,* November 12, 2002, http://www.computerworld.com/mobiletopics/mobile/story/0,10801,75830,00.html.

[5] Flickenger, R., *Building Wireless Community Networks,* Sebastopol, CA: O'Reilly & Associates, 2001.

[6] LaRocca, J., and R. LaRocca, *802.11 Demystified,* New York: McGraw-Hill, 2002.

[7] Stallings, W., "IEEE 802.16 for Broadband Wireless," *Network World,* September 3, 2001, http://www.nwfusion.com/news/tech/2001/0903tech.html.

[8] Marks, R., "EETimes.com: Broadband Access: IEEE Takes on Broadband Wireless," *EE Times,* January 4, 2002, http://www.business2.com/webguide/0,1660,69772,00.html.

[9] Peretz, M., "802.11 Coverage for Miles and Miles," *Wi-Fi Planet*, February 7, 2002, http://www.wi-fiplanet.com/news/article.php/970641.

[10] MeshNetworks, "Corporate and Technology Overview," white paper, http://www.meshnetworks.com/pdf/wp_corpoverview.pdf.

[11] Negroponte, N., "Being Wireless," *Wired Magazine,* Vol. 10, No. 10, October 2002, p. 119.

7

Security and Vo802.11

Early analog cellular phone systems were vulnerable to eavesdropping. As a result the adoption of that technology was not as fast as it might have been had good security been implemented by service providers. Vo802.11 must overcome any user anxieties regarding security on the network. Fears of eavesdropping and fraud can dampen consumer enthusiasm for the service. This chapter describes security measures for 802.11 networks and provides an assessment of the difficulty of "hacking" or otherwise compromising the security of 802.11 networks.

Unlike wired systems, which can be physically secured, wireless networks are not confined to inside buildings, but can be picked up as far as 1,000 feet outside of the premises with a laptop and a gain antenna. This makes WLANs inherently vulnerable to interception. Knowing this, the 802.11 committee added a first line of defense called *Wireless Equivalency Protocol.* WEP is an encryption protocol that is designed to provide the same level of security that wired cables provide. The standard provides both 40- and 128-bit (really only 104-bit) encryption at the link layer using the RC4 algorithm, which the U.S. government allows to be exported.

Electronics retailer Best Buy Co. ran into trouble in mid-2002 when customers who had purchased WLAN cards from Best Buy installed the cards in their laptops before they left the parking lot. The customers noticed unencrypted WLAN traffic that contained customer information and possibly credit card numbers. The Best Buy case provides an example of why enterprises should at a minimum encrypt their WLAN traffic with WEP. By year-end 2002, it had been estimated that 30% of enterprises would have suffered serious security exposures from deploying WLANs without implementing the proper security [1]. The 802.11i task force is currently working on extensions that will help secure the WEP. According to the Wi-Fi Alliance, formally the Wireless Ethernet Compatibility Alliance, smaller organizations should at minimum turn on WEP, password protect shared drives and resources, change the network name from the default *Service Set ID* (SSID), use MAC address filtering, use session keys, and use a VPN system. They also suggest that larger organizations consider additional security methods.

We now turn to a discussion of basic 802.11 security and the known problems. When IEEE 802.11b was first defined, its security depended on two basic security mechanisms: (1) SSID and (2) WEP. Some manufacturers added MAC address filtering to their products.

SSID

SSID is a string used to define a common roaming domain among multiple access points. Different SSIDs on APs can enable overlapping wireless networks. The SSID was thought to be a basic password without which the client could not connect to the network. However, this is easily overridden because APs broadcast the SSIDs multiple times per second and any 802.11 analysis tool such as Airmagnet, Netstumbler, or Wildpackets Airopeek can be used to read it. And, because users themselves often configure clients, this “password” is often widely known. Should you change your SSID? Absolutely. Although the SSID does not add any layer of security, it should be changed from the default value so that other people do not accidentally use your network.

WEP

The IEEE 802.11b standard also defines the WEP authentication and encryption method to mitigate security concerns. Generally, authentication methods are utilized to protect against unauthorized access to the network, whereas encryption is used to defeat eavesdroppers who may try to decrypt captured transmissions. The 802.11 standard uses WEP for both encryption and authentication.

Four options are available when using WEP:

1. Do not use WEP.
2. Use WEP for encryption only.
3. Use WEP for authentication only.
4. Use WEP for authentication and encryption only.

WEP encryption is based on RC4, which uses a 40-bit key in conjunction with a 24-bit random initialization vector to encrypt wireless data transmissions. (This is why you may see some 802.11b systems labeled as having 64-bit encryption. They are no different than those labeled as having 40-bit encryption keys.) If enabled, the same WEP key must be used on all clients and access points for communication. Most vendors today also offer 128-bit WEP (which uses a 104-bit key), a stronger encryption method that increases difficulty for eavesdroppers to decipher over-the-air transmissions. While not part of the IEEE 802.11b standard, this mode has been implemented on many different vendors' products, some of which are not interoperable.

To prevent unauthorized access, WEP also defined an authentication protocol. Two forms of authentication are defined by 802.11b: opens system and shared key. Open system authentication allows any 802.11b client to associate with the access point and skip the authentication process. There is neither any authentication of clients nor encryption of data. It can be used for public-access WLANs such as in coffee shops, airports, hotels, conference centers, and other similar venues where the public is invited to use the network.

Using shared key authentication, the AP sends a "challenge phrase" to the client radio that is requesting authentication. The client radio encrypts the challenge phrase using the shared key and returns it to the AP. If the AP successfully decrypts it back to the original challenge text, this proves that the client has the correct private key. The client is then allowed to make a network connection.

To the casual observer, it would seem that the shared key authentication process is more secure than the open system authentication process. But since both the challenge phrase (which was sent in cleartext) and the challenge are available, a hacker can derive the WEP key. Thus neither open system authentication nor shared key authentication are secure.

Because the 802.11 standard relies on external key management services to distribute the secret keys to each station, and does not specify key distribution services, most 802.11 client access cards and APs rely on manual key distribution. What this means is that the keys remain static unless changed by the network administrator. Obvious problems result from the static nature of the keys and the manual process of key management because changing the keys on each station in a large network can be extremely time consuming. If a station is lost due to theft or accident, the keys will need to be changed on all stations.

WEP provides at most four shared static encryption keys. This means that the four encryption keys are the same for all clients and APs every time a client accesses the network. With enough time and physical proximity and tools downloaded from the Web, hackers can determine the encryption key being used and decrypt data.

MAC Address Filtering

Besides the two basic security mechanisms that 802.11 provides, many products implement MAC address filtering. The MAC address filter contains the MAC addresses of the wireless NICs that may associate with any given AP. Some vendors provide tools to automate the entry and update processes. A MAC filter does not provide very strong security because it is easy to discover known good MAC addresses with a sniffer. Then, using Linux drivers available on the Internet for most 802.11 client access cards, one can configure the sniffed MAC address into the card and gain access to the network. The other two steps mentioned by the Wi-Fi Alliance, use of session keys and a VPN system, are good, workable solutions for securing Wi-Fi.

Security Risks

Security can be defined as keeping anyone from doing things you do not want them to do with, on, or from your data, computers, or peripheral devices. At risk are stored information, the accuracy and value of information, access to internal and external services, and the organization's privacy. Security risks can come from hackers, criminal intruders, corporate raiders, insiders, contractors, and disgruntled employees. Hackers are typically young hobbyists. "Script Kiddiez" copy well-known attacks from the Internet and run them. More sophisticated hackers understand the underlying protocols and their weaknesses. Criminal intruders may be after access to credit card numbers and checking accounts. Corporate raiders may be after financial information, business plans, and intellectual property.

WLAN Security Model

There are four major classes of attack on a system by intruders: interception, fabrication, modification, and interruption [2]. A fifth class of attacks—Repudiation—is an attack against the accountability of information. It is an attack from within the system by either the source entity or the destination entity. Each of these classes of attack can addressed with a security mechanism (Table 7.1). Together, the security mechanisms form a cryptosystem.

Under normal circumstances, information is sent from the source to the destination (Figure 7.1). When an attack occurs it can come in the forms listed in Table 7.1 and discussed in the following subsections.

Table 7.1
Major Classes of Security Attacks

Attack	On	Solved by
Interception	Confidentiality and privacy	Encryption/decryption
Fabrication	Authenticity	Authentication
Modification	Integrity	
Replay		
Reaction		
Interruption	Availability	
Repudiation	Nonrepudiation	

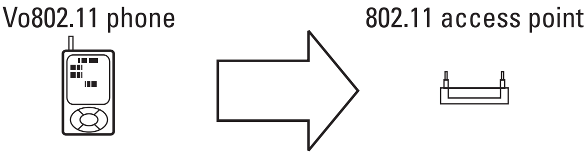


Figure 7.1 Normal flow.

Interception

Interception is a passive attack on confidentiality in which an intruding entity is able to read the information that is sent from the source entity to the destination entity (Figure 7.2). Sniffing is an example of an interception attack.

The intruder attempts to learn or make use of information from the system but does not affect system resources. The identity of the source entity can be intercepted and later used in a masquerade attack, or the intruder may be interested in releasing message contents such as authentication information, passwords, credit card numbers, intellectual property, or other sensitive information. The intruder may also be interested in performing traffic analysis on the system to derive or infer information from the traffic characteristics.

Examples of Interception

Eavesdropping and Sniffing

Eavesdropping is the passive acquisition of information from a network. Just as you can listen to other people's conversations, information can be overheard on the network. This method of gathering information about the network is getting easier with the release of several products. Airopeek, Airsnort, Netstumbler, and WEPCrack are all programs that enable you to acquire information such as

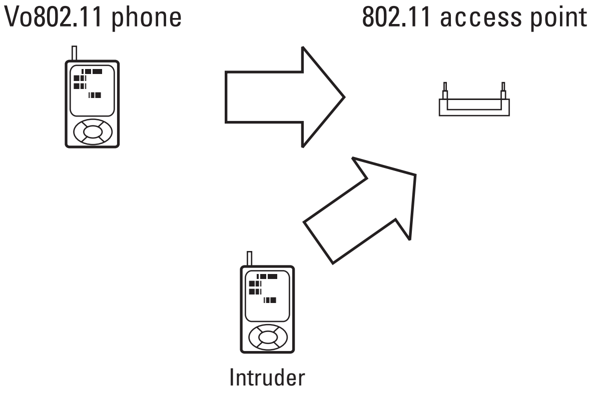


Figure 7.2 Interception in a network.

the SSID, the MAC address of the AP, and information about whether WEP is enabled [3, pp. 156–159].

The nature of an RF-based network leaves it open to packet interception by any radio within range of a transmitter. Interception can occur far outside the users' "working" range by using high-gain antennas (many of which are standard offerings from some vendors). With readily available tools, the eavesdropper is not limited to just collecting packets for later analysis, but can actually see interactive sessions like Web pages viewed by a valid wireless user. An eavesdropper can also catch weak authentication exchanges, like some Web site logins. The eavesdropper could later duplicate the logon and gain access.

The 802.11 standards committee approved WEP, a proprietary encryption design by RSA, before adequate cryptographic analysis was performed. The 802.11i task force is working specifically to correct the flaws in WEP.

WEP is a simple algorithm that uses the RC4 stream cipher to expand a short key and an *initialization vector* (IV) into an infinite pseudorandom number key stream. The sender XORs the plaintext, which is appended with a *cyclic redundancy check* (CRC), with this key stream to produce the ciphertext (Figure 7.3). The receiver has a copy of this key and uses it to generate an identical key stream. The ciphertext is XORed with the key stream and the original plaintext is recovered.

WEP operates at the link layer where packet loss is common. This is why the IV is sent in the clear. If two messages use the same IV and the same key is used with a known plaintext, the other plaintext can be recovered. IEEE 802.11 did not specify how to pick an IV. Most implementations initialize the IV with 0 and afterwards increment it by 1 for each packet sent. This means that if the unit is reset, the IV starts at 0 again.

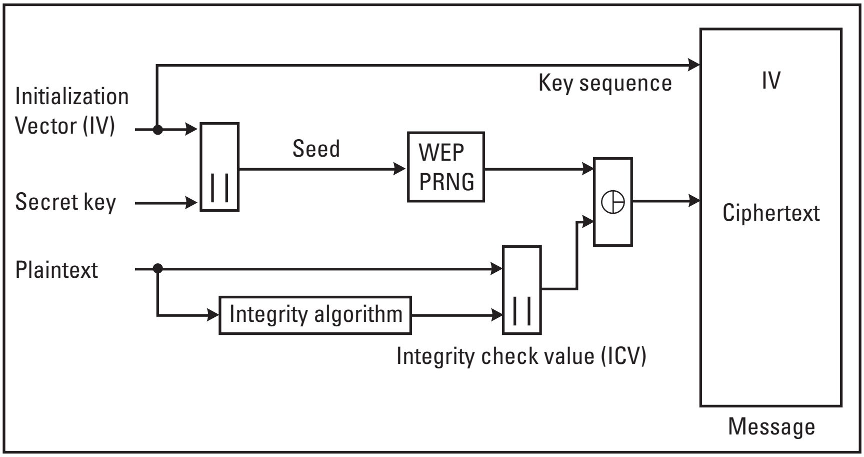


Figure 7.3 Creation of ciphertext in WEP. (*From:* [4]. © 2000 Intel Corporation, Inc. Reprinted with permission.)

There are only 24 IV choices. If the IVs were randomly chosen it only takes 12,430 frames to be 99% sure that an IV was reused. This is due to the birthday principle. For example, in a room of 23 or more people the probability of 2 people having the same birthday is 50%.

Because WEP sends the IV in the clear along with the encrypted message, it is possible to use dictionary building and statistical methods to crack the WEP key. Both the 64- and 128-bit implementations have the same flaw. The 802.11 standard leaves WEP implementation to the WLAN manufacturers, so the implementations may not be exactly the same. This adds to further weaknesses in the system.

WEP was designed for home use and small businesses. WEP has one static key for the entire system. If a laptop, PDA, or other 802.11 device gets stolen or misplaced from the enterprise, one cannot disable a single user's key, but the entire enterprise needs to be rekeyed.

Another problem is that WEP does not have a key distribution system. In a small business, it is sufficient to enter the keys into the access point and the handful of laptops. However, in a larger organization, manually entering keys is not a scalable operation. If an enterprise needs to be rekeyed, a trusted person must enter the key into the client card of every 802.11 device—manually.

Some vendors use Hex keys, others use ASCII keys, yet others use a key generation phrase, or a combination of two or three of these formats. Some client card vendors have four keys with the ability to choose one out of four. Some cards do not provide encryption at all, while others only 40 bit, and yet others allow both 40-bit and 104-bit encryption.

Fabrication

Fabrication is an active attack on authentication where the intruder pretends to be the source entity (Figure 7.4). Spoofed packets and fake e-mails are examples of a fabrication attack.

WEP has two authentication mechanisms. With the default authentication algorithm called *open system authentication,* the client only announces the intent to associate with the access point and the access point looks at the MIB and looks to see if AuthenticationType = OS. If so, access is allowed. Open system authentication, by its very nature, does not perform authentication and provides no security whatsoever (Figure 7.5).

WEP also has an optional authentication algorithm called *shared key authentication* in which the client can ask to be authenticated using shared key authentication. The AP in turn generates a random 128-bit challenge and sends it to the client (Figure 7.6). The client replies to the challenge, encrypted with the shared secret key, which is configured into both the client and AP. The AP decrypts the challenge, using a CRC to verify its integrity. If the decrypted frame matches the original challenge, the station is considered authentic. Optionally, the challenge/response handshake is repeated in the opposite direction for mutual authentication.

An attacker who captures these frames possesses all of the parts required to derive the RC4 keystream—plaintext, ciphertext, and IV—and respond to a future challenge The attacker can now pretend he is a valid client on the WLAN.

Because the key is shared with all users, there is no mechanism for authenticating individual users and hardware. If the key is leaked or cracked, anyone knowing the key can use the system. WEP also has no mechanism for the users or hardware to authenticate the access point. Without two-way authentication,

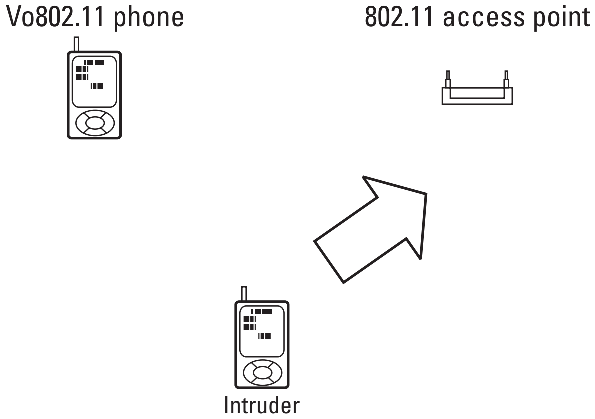


Figure 7.4 Fabrication in a network.

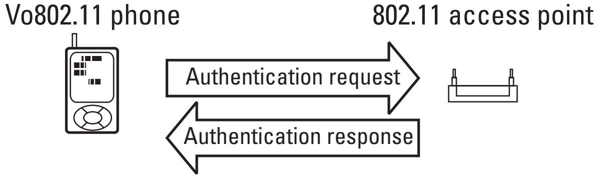


Figure 7.5 Open system authentication in an 802.11 network.

it is possible for an attacker to simulate the wireless network and get users to connect to it and to reveal additional information useful to the attacker.

MAC address filtering is sometimes used to control access to resources. However, MAC address filtering is not adequate for authentication of users. It is relatively simple to sniff valid MAC addresses out of the air and change the MAC address of a client card to masquerade as a legitimate user. Once access is gained to the network, all computers on the network are accessible because WEP and 802.11 do not provide access control mechanisms to limit which resources can be accessed. In a home, SOHO, or small business environment, this may not be an issue. However, in an enterprise environment, it may be important to control access to resources based on access policies.

Examples of Fabrication

Man-in-the-Middle Attacks

To execute a *man-in-the-middle attack,* two hosts must be convinced that the computer in the middle is the other host. The classic version of this attack

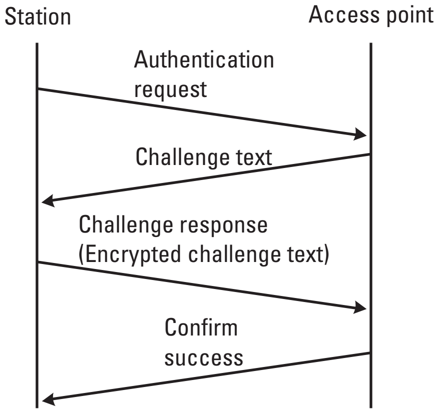


Figure 7.6 Shared key authentication in an 802.11 network. (*From:* [4]. © 2000 Intel Corporation. Reprinted with permission.)

occurs when an attacker intercepts packets from the network, modifies them, and reinserts them into the network.

Spoofing

Spoofing is pretending to be someone or something that you are not, such as using another person's user ID and password. DNS spoofing is accomplished by sending a DNS response to a DNS server on the network. IP address spoofing depends on the fact that most routers only look at the destination EP address, not the sending address. Validating the sending IP address can prevent this type of spoofing [5, pp. 72–74].

Insertion Attacks

Configuring a device to gain access to a network or inserting unauthorized devices into a network in order to gain access is called an *insertion attack.* By installing wireless network cards and being in the vicinity of a target network, a device can be configured to gain access. Unauthorized APs can be installed in an attempt to get users to connect to a hacker's AP rather than to the intended network AP. If these APs are installed behind the corporate firewall, the risk of attack is much greater. This can sometimes be done by well-meaning, but misinformed employees [3, p. 157].

Brute-Force Password Attacks

Also known as *password cracking* or *OT dictionary attacks,* a *brute-force password attack* uses a dictionary and repeated attempts to test passwords to attempt to gain access to the network. This type of attack is possible even if password authentication is implemented [3, p. 157].

Invasion and Resource Stealing

Once an attacker has gained the knowledge of how a WLAN controls admittance, he or she may be able to either gain admittance to the network on his own or steal a valid station's access. Stealing a station's access is simple if the attacker can mimic the valid station's MAC address and use its assigned IP address. The attacker waits until the valid system stops using the network and then takes over its position in the network. This would allow an attacker direct access to all devices within a network, or to use the network to gain access to the wider Internet, all the while appearing to be a valid user of the attacked network [5].

Modification

Modification is an active attack on integrity in which an intruding entity changes the information that is sent from the source entity to the destination entity (Figure 7.7). Insertion of a Trojan horse program or virus is an example of a modification attack.

WEP is wide open to a modification attack without detection because the ICV is a linear function that only uses addition and multiplication; that is,

$$\text{crc}(x \text{ XOR } y) = \text{crc}(x) \text{ XOR } \text{crc}(y)$$

With the CRC-32 integrity check, it is possible to change one or more bits in the original plaintext and one can predict which bits in the checksum need to be changed for the message to remain valid. This means it is possible to take messages from the source entity, modify them, and reinsert them in the data stream without detection. Basic 802.11 security does not guarantee message integrity. WEP or its replacement cipher needs to have a secure integrity check.

Examples of Modification Attacks

Loss of Equipment

The *loss of equipment* is an issue that has recently received quite a bit of attention due to events within the FBI. The loss of a laptop or other piece of equipment poses the issue of what data were contained within the device. It is possible for an unscrupulous person to dial into the wired network using lost or stolen equipment and stored passwords and masquerade as an authorized user. This scenario is possible with current wired networks and is not dependent on having access to a WLAN. The loss of a device equipped with wireless access certainly carries the same risks.

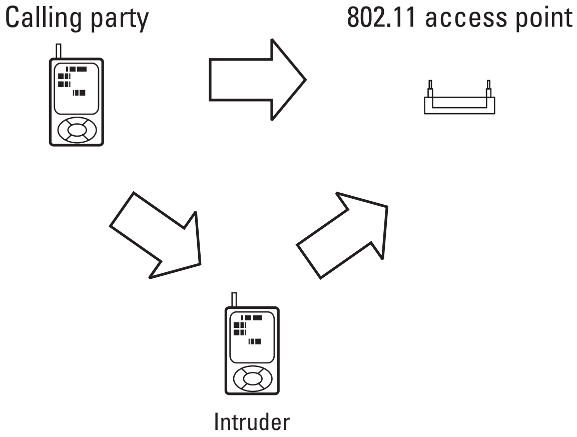


Figure 7.7 Modification attack in an 802.11 network.

Virus Infection

Virus infection is another issue that affects both wired and wireless networks. To date, there have been no reported viruses that infect cell phones; however, there have been viruses that are capable of sending text messages to cell phones. Two of these are VBS/Timo-A and the LoveBug. There have been reports of viruses that infect Palm OS units as well as viruses carried on diskette, CD-ROM, and e-mail. These viruses can infect laptops whether or not they are wireless equipped and can be introduced into and spread via either the larger wired or wireless network [3, p. 153].

Replay

Replay is an active attack on integrity in which an intruding party resends information that is sent from the source entity to the destination entity (Figure 7.8).

Basic 802.11 security has no protection against replay. It does not contain sequence numbers or time stamps. Because IVs and keys can be reused, it is possible to replay stored messages with the same IV without detection to insert bogus messages into the system. Individual packets must be authenticated, not just encrypted. Packets must have sequence numbers or time stamps.

Examples of Replay Attacks

Traffic Redirection

An attacking STA can poison the ARP tables in switches on the wired network through the AP causing packets for a wired station to be routed to the attacking STA. The attacker can either passively capture these packets before forwarding

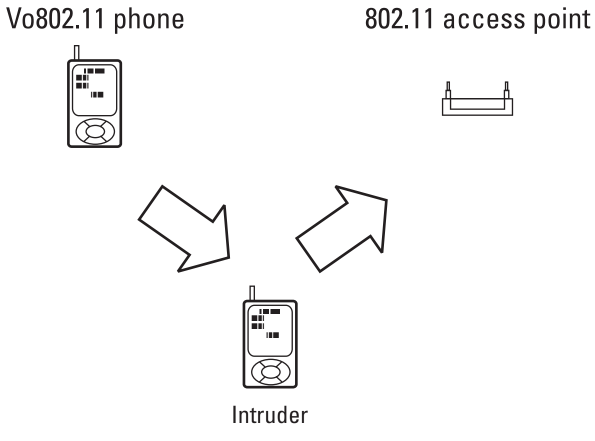


Figure 7.8 Replay attack on a network.

them to the attacked wired system or can attempt a man-in-the-middle attack. In such an attack, all the susceptible systems could be on the wired network.

Reaction

Reaction is an active attack where packets are sent by the intruder to the destination (Figure 7.9). The reaction is monitored by the intruder. Additional information can be learned from this new side channel.

Interruption

An active attack on availability in which an intruding entity blocks information sent from the originating entity to the destination entity is referred to as an *interruption* (Figure 7.10). Examples are denial-of-service attacks and network flooding.

The intruder may try to exhaust all network bandwidth using ARP flooding, ping broadcasts, TCP SYN flooding, queue flooding, smurfs, synk4, and other flood utilities. Also, the intruder may use some physical mechanism such as RF interference to successfully interrupt a network. Related to this is a degradation of service attack where service is not completely blocked, but the quality of service is reduced. With basic 802.11 security, little can be done to keep a serious intruder from mounting a denial of service attack.

Denial of Service Attacks

Denial of service (DoS) attacks do not allow a hacker to gain access to the network; rather, they basically make computer systems inaccessible by overloading

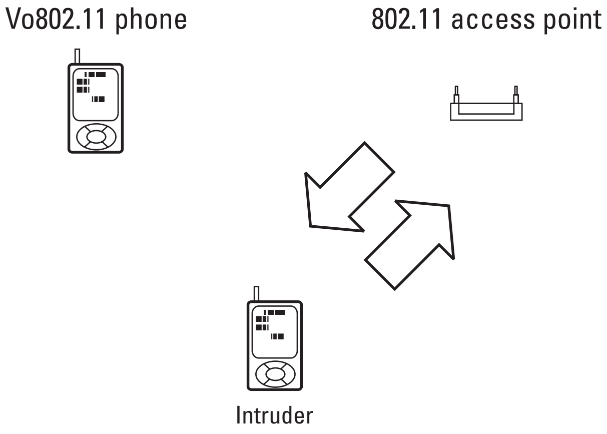


Figure 7.9 Example of a reaction attack.

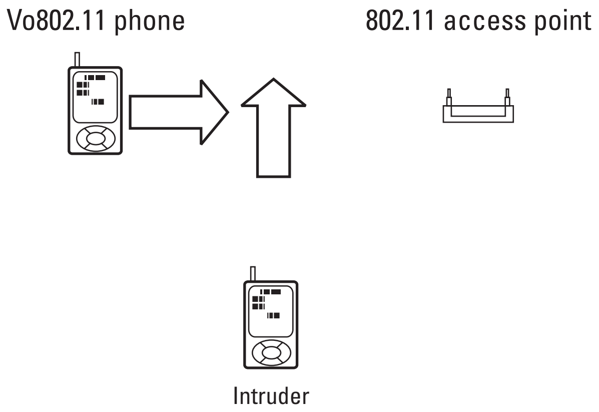


Figure 7.10 Example of interruption.

servers or networks with useless traffic so legitimate users can no longer access those resources. The intention is to prevent the network from providing services to anyone. Usually this is accomplished by overloading a resource to cause a failure. The overload causes the host to become unavailable, much like those annoying messages of "all circuits are busy." There are many variations on these types of attacks depending on the type of resource blocked (disk space, bandwidth, internal memory, and buffers), and some are more easily prevented than others. In the simplest case, turning off the service when it is not needed prevents this type of attack. In other cases, they cannot be easily blocked without limiting the use of a necessary resource. In a wireless network, because the airwaves are shared by other devices such as cordless telephones, microwave ovens, and baby monitors, an attacker with the proper equipment can flood the airwaves with noise and disrupt service to the network [3, pp. 152–158].

Examples of DoS Attacks

Rogue Networks and Station Redirection

An 802.11 wireless network is very susceptible to a rogue AP attack. A rogue AP is one owned by an attacker that accepts STA connections and then at a minimum intercepts traffic if not also performing man-in-the-middle attacks before allowing traffic to flow to the proper network. The goal of a rogue is to get valid traffic off the WLAN onto a wired network for attacking (or to conduct the attack directly within the rogue AP), and then reinsert the traffic into the proper network. Such rogue APs could readily be deployed in public areas as well as shared office space areas.

Repudiation

A *repudiation* attack is an active attack on nonrepudiation by either the source or the destination in which either the source entity denies sending a message, or the destination entity denies receiving a message (Figure 7.11).

Basic 802.11 security does not have nonrepudiation. Without nonrepudiation, the source entity can deny ever having sent a message and the destination entity can deny ever having received the message.

Network Architecture

Typical Network Architecture with WLAN Added

One should defend the LAN network from users on the wireless access points. Figure 7.12 shows a typical corporate infrastructure today. The Internet connects to a router on the WAN side. On the LAN side of the router, one may optionally connect a *demilitarized zone* (DMZ) server that is accessible from the Internet for file transfer as an example. A firewall separates the Internet from the corporate network on the LAN side. Often, this firewall function is included in the router. The LAN side serves computers and, more recently, APs and wireless laptops. However, the new AP accidentally creates a way to get in behind the firewall through the air link. One can see how this leaves the network open to vulnerabilities. Once the user has achieved wireless access, the user also has access to the LAN inside the company.

Typical Network Architecture with WLAN and Wireless Firewall Added

One can change the network architecture as shown in Figure 7.13 by adding a wireless authentication firewall that regulates access to the LAN by allowing users to pass only after they have been authenticated. An optional wireless DMZ server or capture portal may exist on the WLAN side of the network. The wireless authentication firewall separates the WLAN from the LAN, thus protecting the enterprise's network from access through the wireless equipment. In an 802.1x/*Extensible Authentication Protocol* (EAP) arrangement, the access point

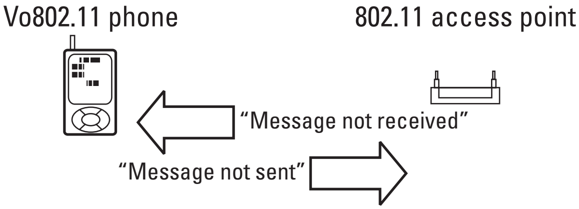


Figure 7.11 Example of repudiation.

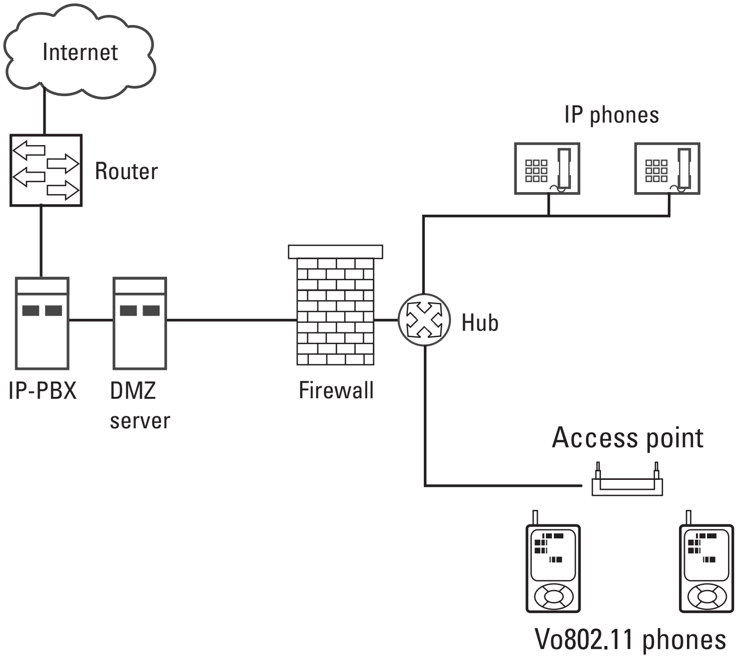


Figure 7.12 Wireless LAN architecture.

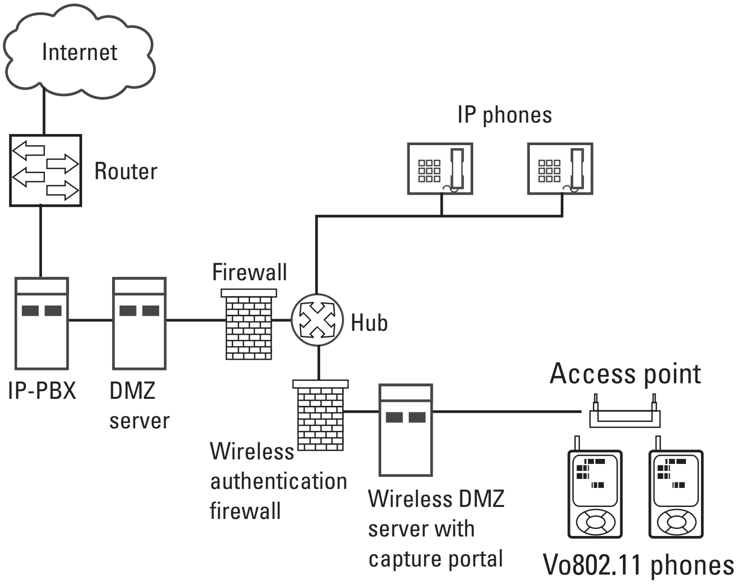


Figure 7.13 A wireless authentication firewall protects the LAN.

will contain the firewall and an additional *remote access dial-in user service* (RADIUS) server will need to be located on the LAN. In a VPN arrangement, the LAN hosts a VPN server, which forms the termination point of the VPN tunnel. Both of the firewalls will need a hole to carry VPN traffic from the WAN and WLAN side to the LAN.

Mobility and Security

If mobility is used, the solution must be secure during handoff. Handoffs open the network up to a redirection attack. If not properly secured, the intruder can take over the communication with the destination entity after the handoff.

Security Policy: A Range of Options

To build a security system, one needs to know what is being protected. These could be devices such as servers, routers, and modem banks and information such as e-mail, intellectual property, trade secrets, customer lists, business plans, and medical records. Sometimes, the information has to be protected by law. One also needs an idea of who this material is being protected from: hackers, customers, insiders (employees and contractors), competitors. From this one can do a simple risk analysis to determine what is at risk—data or the network—and the level of countermeasures required to solve the problem.

In risk management one can either ignore, accept, defend, or pass on a problem. Unfortunately, there is no canned security policy that you can obtain or use. Each business has its own unique requirements and practices that dictate how implementations are made. Table 7.2 shows varying levels of security, the configuration, what is secured by the configuration, and in what applications such a configuration might be used.

No Security

No security is like leaving your door wide open. Anyone can come in and use the network access. Basically it means a company is ignoring potential security problems.

Public Access

Public access is like handing out keys to everyone you know and trust. NoCat Auth, Sputnik, and Wayport all restrict access to public users by authenticating them. The authentication process protects the network by preventing fabrication of access credentials. In some cases, billing is exchanged for granting access to the system. In NoCat Auth's case, access can be removed for people who abuse the system. Many of these solutions do not provide a secure tunnel for

Table 7.2
A Range of Security Options for Wireless Networks

		Configuration	Entity Secured	Applications
0	No security	Network out of the box; no configuration; no WEP	Nothing	?
1	Public access	User authentication; user must supply VPN through Internet back to enterprise	Network access	Hot spots, libraries, coffee shops, hotels, airports, and so on, with portability
2	Limited security	40- or 128-bit WEP; MAC access control list; no broadcast	Some network access; some data privacy	Home and SOHO with portability
3	Basic security	Wi-Fi protected access; later 802.11i	Network access and data privacy	Home, SOHO, and small enterprise with portability
4	Advanced security	802.1x/EAP-x, RADIUS	Network access and data privacy	Enterprise with portability
5	End-to-end security	VPNs such as PPTP, PPTPv2, L2TP, Kerberos, and IPsec	Network access and data privacy	Special applications, business travelers, telecommuting, and so on; enterprise with outside users

their users. The data sent over the air are in the clear. Users must provide their own protection against breeches of confidentiality such as using a VPN to tunnel back to their enterprise network.

Basic Access Control

Currently, basic access is like hiding the key under the mat. It is hidden out there for clever people to find, but the network access and data accessed may not be worth the trouble to access. At minimum, one should turn on WEP, password protect shared drives and resources, change the network name from the default (SSID), use MAC address filtering, and turn off broadcasts if possible. The IEEE offers two solutions: WEP enhancements in the short term and, in the long term, WEP replacements.

802.11 Security Measures Beyond WEP

Wi-Fi Protected Access

In November 2002, the Wi-Fi Alliance announced the *Wi-Fi Protected Access* (WPA) security standard [5, pp. 81–86]. It replaced the comparably weak WEP standard previously offered for Wi-Fi equipment.

WPA uses the *Temporal Key Integrity Protocol* (TKIP), a more hardened encryption scheme than that used in WEP. TKIP uses *key hashing* (KeyMix) and a nonlinear *message integrity check* (MIC). TKIP also uses a *rapid-rekeying* (ReKey) protocol that changes the encryption key about every 10,000 packets. However, TKIP does not eliminate fundamental flaws in Wi-Fi security. If one can hack TKIP, one not only breaks confidentiality, but also access control and authentication.

WPA will work in two different ways, depending on the type of network. In homes and small offices lacking authentication servers, the technology will work in a so-called "preshared" key mode. Users simply enter the network key to gain access.

In the managed mode, it will work with authentication servers and will require the support of 802.1x and EAP. The 802.1x/EAP combination enables a client network adapter to negotiate via an access point with a back-end authentication server using securely encrypted transactions to exchange session keys.

Every device on a wireless network must be upgraded to WPA in order for it to work. If a network is sewn together from several manufacturers' devices and the WPA upgrade for one of the manufacturers is not available yet, the user will have to wait to be able to deploy WPA. A mixed network can run with WPA and WEP, both installed. However, security in the networks will default to WEP, which offers less protection.

WPA contains many parts of 802.11i. However, some of the key elements are not included such as support for a new encryption algorithm called the *Advanced Encryption Standard* (AES), which will replace the RC4-based encryption algorithm when 802.11i becomes available. Migrating to AES encryption will require hardware changes because AES is computationally more complex than RC4. Also secure fast handoff preauthentication, secure deassociation and deauthentication, and security for peer-to-peer communications (ad hoc mode) will follow when 802.11i is released. When 802.11i is a deployed standard, products will be labeled "Wi-Fi WPA2-certified" [6].

WPA, when it becomes available, will be a good step toward securing a home or SOHO WLAN. However, for larger enterprises, 802.1/EAP and VPN are still viable means of securing the WLAN. The 802.1x/EAP combination offers more EAP choices and VPN offers workers access to the enterprise's network from any location.

802.1x and EAP Advanced Security

The 802.1x standard provides an authentication framework for wireless LANs, allowing a user to be authenticated by a central authority. The actual algorithm that is used to determine whether a user is authentic is left open and multiple algorithms are possible. Examples are certificate-based solutions such as

EAP-TLS, password-based solutions such as EAP-OTP and EAP-MD5, smart card-based solutions such as EAP-SIM, and hybrids such as EAP-TTLS, which use both certificates and passwords. Some companies offer their own proprietary EAP solution, such as LEAP from Cisco.

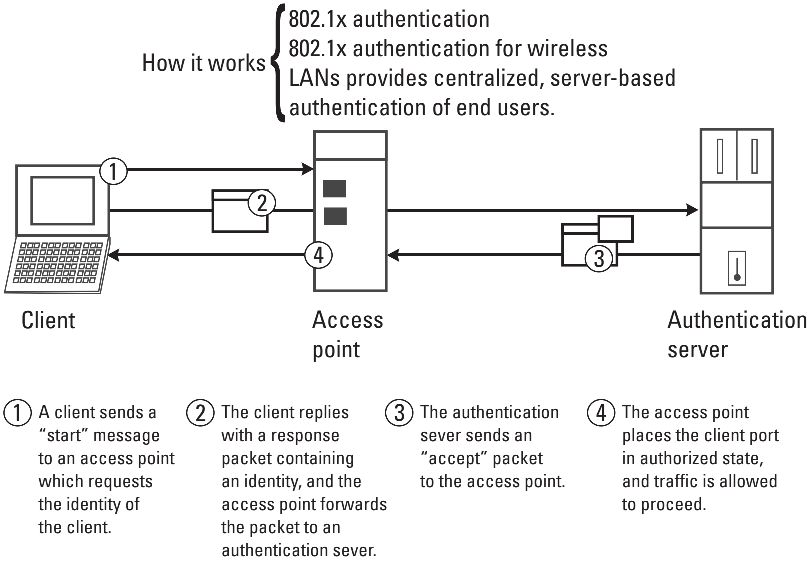
802.1x uses an existing protocol, EAP (RFC 2284), which works on Ethernet, token ring, or wireless LANs, for message exchange during the authentication process.

802.1x Network Port Authentication

The 802.1x authentication feature (Figure 7.14) for wireless LANs has three main components: the supplicant (usually the client software), the authenticator (usually the access point), and the authentication server (usually a remote authentication dial-in user service server, although RADIUS is not specifically required by 802.1x) [7]. The authenticator connects to the LAN network.

The normal flow for the 802.11x authentication process is as follows: The supplicant (in the client) tries to connect to the access point by sending a start message. The access point detects the supplicant and enables the supplicant's port in an unauthorized state, so only 802.1x/EAP messages are forwarded. All other traffic is blocked.

The supplicant then sends an EAP-start message. The access point replies with an EAP-request identity message to obtain the supplicant's identity. The



802.1x authentication
802.1x authentication for wireless LANs provides centralized, server-based authentication of end users.

1. A client sends a "start" message to an access point which requests the identity of the client.
2. The client replies with a response packet containing an identity, and the access point forwards the packet to an authentication sever.
3. The authentication sever sends an "accept" packet to the access point.
4. The access point places the client port in authorized state, and traffic is allowed to proceed.

Figure 7.14 The 802.1x authentication feature. (*From:* [8]. © 2003 Network World, Inc. Reprinted with permission.)

client's EAP-response packet containing the supplicant's identity is forwarded to the authentication server.

The authentication server authenticates the supplicant and either responds by accepting or rejecting the supplicant. As the response is passed back through the authenticator, a passed authentication response will enable the port, whereas it remains blocked for a failed authentication response.

EAP

To address the shortcomings of WEP for authentication, the industry is working towards solutions based on the 802.1x specification, which is itself based on the IETF'S Extensible Authentication Protocol. EAP was designed with flexibility in mind, and it has been used as the basis for a number of network authentication

The 802.1x specification is based on EAP. EAP is formally specified in RFC 2284 and was initially developed for use with *Point-to-Point Protocol* (PPP). When PPP was first introduced, two protocols were available to authenticate users, each of which required the use of a PPP number. Authentication is not a "one size fits all" problem, and it was an advanced area of research at the time. Rather than burn up PPP numbers for authentication protocols that might become obsolete, the IETF standardized EAP [9].

IEEE 802.1x is not a single authentication method; rather it utilizes EAP as its authentication framework. This means that 802.1x-enabled switches and access points can support a wide variety of authentication methods, including certificate-based authentication, smart cards, token cards, one-time passwords, and so on. However, the 802.1x specification itself does not specify or mandate any authentication methods. Since switches and access points act as a "pass through" for EAP, new authentication methods can be added without the need to upgrade the switch or access point, by adding software on the host and back-end authentication server.

Because IEEE 802.1x does not involve encapsulation [unlike *Point-to-Point Protocol over Ethernet* (PPPOE) or VPN] it adds no per-packet overhead and can be implemented on existing switches and access points with no performance impact. This means that IEEE 802.1x can scale from speeds of 11 Mbps (802.11) to 10+ Gbps, and can be enabled on existing switches with a firmware upgrade, without the need to buy new hardware. On hosts, because IEEE 802.1x can be implemented in the NIC driver, support can be enabled by obtaining updated drivers from the NIC vendor; there is no need to install a new operating system.

IEEE 802.1x integrates well with open standards for authentication, authorization, and accounting (including RADIUS and LDAP) and so it fits in well with an existing infrastructure for managing dial-up networks and VPNs.

RADIUS servers (including Windows 2000 IAS) that support EAP can be used to manage IEEE 802.1x-based network access. These specifications describe how IEEE 802.1x works and how it can be managed via RADIUS and SNMP. Through RADIUS, IEEE 802.1x permits management of authorization on a per-user basis. Per-user services include filtering (layer 2 or layer 3), tunneling, dynamic VLANs, rate limits, and so on [5, pp. 91–92].

802.1x/EAP Authenticators

Many commercial access points such as Cisco, Lucent/Orinoco/Agere, and Enterasys feature 802.1x authentication support. Support for homebrew authenticators is provided with the Open1X project.

RADIUS

RADIUS [10] is currently the de facto standard for remote authentication. It is a widely deployed protocol for network access *authentication, authorization, and accounting* (AAA) in both new and legacy systems. Although it has a few security and transport issues associated with it, it is very likely that RADIUS will continue to be widely used for many years to come. Eventually, RADIUS may be replaced by a new protocol called DIAMETER. RADIUS is simple, efficient, and easy to implement, making it possible for RADIUS to fit into even the most inexpensive embedded devices.

The security issues revolve around shallowness of the protocol and poor implementation of the specification. The protocol lacks confidentiality, authentication of client messages, and protection against a replay attack (integrity). The shared secret scheme does not allow for enough entropy in the keys and it seems to be open to off-line dictionary attacks. RADIUS needs to be secured with an external protocol like *Internet Protocol Security* (IPsec).

The issues with transport are most relevant for accounting in situations where services are billed according to usage. RADIUS runs on UDP and has no defined retransmission or accounting record retention policy and does not support application-layer acknowledgments or error messages. Lost packets can mean revenue loss. This makes RADIUS accounting unreliable for usage-based billing services, particularly in interdomain usage (such as roaming), where substantial packet loss can occur when using the Internet.

RADIUS has two specifications that make up the protocol suite: authentication and accounting. The authentication portion can be used to determine if a user can gain access to the network. The authentication can be done locally or by proxy to another RADIUS server.

FreeRADIUS is available for a wide range of platforms, including Linux, FreeBSD, OpenBSD, OSF/Unix, and Solaris. OpenRADIUS is a RADIUS server that can be compiled to run on many variations of Unix [11].

EAP MD5

EAP-MD5 or CHAP [12] represents a kind of base-level EAP support among 802.1x devices. It is the least secure version of EAP because it uses user names and passwords for authentication, which are easily socialized. Also, it is vulnerable to dictionary attacks. In addition, EAP-MD5 does not support dynamic WEP keys, which is a critical liability.

LEAP

Cisco was one of the first vendors to market with its proprietary *Lightweight EAP* (LEAP) [13]. LEAP works only with Cisco client 802.11 cards, RADIUS servers, and Cisco access points. LEAP is vulnerable to man-in-the-middle dictionary attacks.

EAP-TLS

EAP-TLS (*transport layer security*) is an open standard [14] that is supported by many vendors. It uses *public key infrastructure* (PKI), which takes advantage of asymmetric public and private keys and thus is very secure. EAP-TLS is supported natively in Windows XP and by Windows 2000 servers. The only burden is that one must set up a PKI because every device needs an x.509 certificate. However, once EAP-TLS is set up it is virtually transparent to the user.

EAP-TTLS

EAP-TTLS and EAP-TLS are similar in that both use TLS, the successor to SSL, as the underlying strong cryptography. However, EAP-TTLS differs in that only the RADIUS servers, not the users, are required to have certificates. The user is authenticated to the network using ordinary password-based credentials, whose use is made secure against active and passive attack by enclosing it in the TLS security wrapper.

EAP-SIM

EAP-SIM (*subscriber identification module*) is an EAP method designed by Nokia that allows hardware authentication to a SIM chip. A SIM is a secure processor about the size of a small postage stamp. SIMs are currently used in GSM mobile phones to authenticate the user on a mobile network. After clicking connect and optionally entering a *personal identification number* (PIN), the system authenticates with the network and then connects to the Internet. The beauty of a SIM chip is that it makes cloning of authentication secrets very difficult. It is likely that many of the GSM carriers such as T-Mobile and Sonera will implement EAP-SIM to secure their public wireless LAN offerings.

PEAP

PEAP is another Cisco-developed protocol. Whereas EAP was originally created for use with PPP, it has since been adopted for use with IEEE 802.1x network port authentication. Since its deployment, a number of weaknesses in EAP have become apparent. These include lack of protection of the user identity or the EAP negotiation, no standardized mechanism for key exchange, no built-in support for fragmentation and reassembly, and lack of support for fast reconnect.

By wrapping the EAP protocol within TLS, *Protected EAP* (PEAP) [15] addresses these deficiencies. Any EAP method running within PEAP is provided with built-in support for key exchange, session resumption, and fragmentation and reassembly. PEAP provides the ability to seamlessly roam between access points. In the near future, PEAP will support the same EAP types that EAP supports. One of the disadvantages of 802.1x/EAP is the lack of supplicants.

VPNs

A virtual private network enables a specific group of users to access private network data and resources securely over the Internet or other networks. VPNs are characterized by the concurrent use of tunneling, encryption, authentication, and access control over a public network.

VPNs create "virtual" point-to-point connections using a technique called *tunneling.* As the name suggests, tunneling acts like a "pipe" that bores through a network cloud to connect two points. Typically started by a remote user, the tunneling process encapsulates data and encrypts it into standard TCP/IP packets, which can then securely travel across the Internet to a VPN server on the other side where they are decrypted and de-encapsulated onto the private LAN network.

The two basic VPN types are as follows:

1. *Remote access VPNs:* Securely connect remote users, such as mobile users and telecommuters, to the enterprise. This type of VPN can be used by 802.11 users to initiate a session back to their corporate LAN, for example, salespeople equipped with laptops and telecommuters that would like to connect intermittently from various diverse locations such as hotels, airports, convention centers, and coffee shops. The key concerns are encryption and authentication; performance and bandwidth can be sacrificed due to the fact the connection is broadband.
2. *LAN-to-LAN VPNs:* Securely connect remote and branch offices to the enterprise (intranet VPNs). Securely connect third parties, such as customers, suppliers, and business partners, to the enterprise (extranet VPNs). Intranet or extranet VPNs can be used to secure wireless point-to-point links. For example, a wireless backhaul may connect a hot spot back to a central location. This type of VPN needs to be encrypted and authenticated as well as meet strict performance and

bandwidth requirements since this kind of connection carries network traffic.

How VPN Works with 802.11

To support 802.11 wireless LANs, a VPN client software application is deployed on all machines that will use the wireless LAN, and a VPN gateway is introduced into the network between the access point and the wireless LAN segment. An encrypted VPN tunnel is built from the laptop through the wireless gateway and terminated at the VPN gateway in order to gain access to the wired LAN through the wireless access point. All traffic passing through the access point must go through the VPN gateway before entering the LAN. The cleartext data on the other side of the secure tunnel can then continue on to their destination inside the local network. The VPN tunnel provides authentication, data confidentiality, and data integrity. Thus, other encryption mechanisms such as WEP are no longer needed.

A VPN solution also allows mobile workers to access their network from a remote location where security may or may not be provided. The secure tunnel extends from the client's computer, through the Internet, through the firewall/VPN gateway to the VPN server. From the VPN server, the data continue to their destination inside the corporate network.

Vulnerabilities of VPN

While VPNs are touted as a secure solution for wireless LANs, VPNs using one-way authentication are still vulnerable to exploitation from, for instance, man-in-the-middle attacks. Deployment of wireless LANs in large organizations can create a nightmare of distributing and maintaining client software to all clients. Almost all VPN solutions shipping today are proprietary (not IETF standard) in some form or another and are generally not interoperable. Because of this fact, not all devices may have client software available for any one VPN supplier. Also, it is often the case that once a VPN is installed, a different VPN will not operate on the same machine. Thus VPNs are impractical for securing a public access WLAN.

It is also important to note that many of the proprietary security extensions may have security flaws due to the lack of cryptographic rigor applied to them. Despite these vulnerabilities, encryption, authentication, and integrity remain essential elements of wireless LAN security [16].

VPN Standards

Many protocols have been written to use with VPNs. These protocols attempt to close some of the security holes inherent in VPNs. These protocols continue to compete with each other for acceptance in the industry and are not compatible with each other.

IPsec

IPsec VPNs have nearly become accepted as the de facto standard for securing IP data transmission over shared public data networks since VPN software has been developed for a wide variety of clients. It addresses authentication, data confidentiality, integrity, and key management, in addition to tunneling.

Basically, IPsec encapsulates a packet by wrapping another packet around it. It then encrypts the entire packet. This encrypted stream of traffic forms a secure tunnel across an otherwise unsecured network. Interoperability issues still exist, however, between different vendors' implementations.

Point-to-Point Tunneling Protocol

The *Point-to-Point Tunneling Protocol* (PPTP) is a protocol specification developed by several companies. Nearly all flavors of Windows include built-in support for the protocol. PPTP was the dominant VPN before IPsec was deployed. PPTP tunnels data, encrypts user data, and authenticates users.

PPTP is a tunneling protocol that provides remote users with encrypted, multiprotocol access to a corporate network over the Internet. PPTP uses *generic routing encapsulation* (GRE). PPTP wraps IP packets in GRE packets before sending them down the tunnel. Network layer protocols, such as IPX and NetBEUI, are encapsulated by the PPTP protocol for transport over the Internet.

The initial releases of PPTP for Windows by Microsoft contained security features that some experts claimed were too weak for serious use, but Microsoft continues to improve its PPTP support.

Layer Two Tunneling Protocol

The *Layer Two Tunneling Protocol* (L2TP) was developed by Cisco. L2TP supports non-TCP/IP clients and protocols (such as frame relay, ATM, and SONET) but fails to define any encryption standard. Although L2TP is compatible with most network protocols it is not widely deployed; however, it is common in certain telco and *Internet service provider* (ISP) networks.

L2TP over IPsec

L2TP over IPsec offers tunneling, user authentication, mutual computer authentication, encryption, data authentication, and data integrity. L2TP's offers multiprotocol support.

SSL

The *Secure Sockets Layer* (SSL), working only with TCP/IP protocols, is the primary protocol for secure connections from Web browsers to Web servers, usually for secure credit card connections or for sensitive data. SSL requires a valid site certificate issued from an authorized certificate authority. SSL provides tunneling, data encryption, mutual authentication, integrity, and nonrepudiation.

UPN-Related Security Protocols

SOCKS Network Security Protocol

SOCKS is a VPN protocol that operates on layer 5, whereas most others operate at layer 2 or 3. SOCKS version 5 is a circuit-level proxy protocol that was originally designed to facilitate authenticated firewall traversal. Functioning at a higher level means that SOCKS only operates with certain applications. SOCKS v5 supports a broad range of authentication, encryption, tunneling, and key management schemes. It is generally considered to be a market failure.

IP Addresses NAT/PAT

Many VPNs require fully routable/public IP addresses and no port blocking on any ports except incoming port 80 and port 139. (VPNs do not use these ports.) The reason for this is that the VPNs cannot tolerate *network address translation* (NAT) or *port address translation* (PAT). However, NAT/PAT is necessary to provide network security, to prevent subscribers from running host servers on their LAN, and to preserve valuable IP address space. A growing number of VPN clients support emerging standards for UDP encapsulation to push IPsec through NAT/PAT. PPTP often passes through NAT/PAT without trouble, but L2TP over IPsec also requires encapsulation.

Kerberos

Kerberos provides a third method of securing the 802.11 over the air link. It is used primarily by Symbol Technologies, Inc., with their Spectrum24 WLANs. Kerberos provides robust security and uninterrupted network connectivity for voice and data devices and addresses the security needs and concerns of network managers.

Kerberos provides both user authentication and encryption key management and can guard networks from attacks on data in transmission, including interruption, interception, modification, and fabrication. Kerberos was voted as the “mandatory-to-implement” security service for 802.11e authentication and encryption key management. Kerberos provides confidentiality, authentication, integrity, access control, and availability. Kerberos also works very well during handoffs between access points, resulting in uninterrupted application connectivity. Reauthentication to the network is very quick.

How Kerberos Works with 802.11

Kerberos is based on the key distribution model developed by Needham and Schroeder [16]. Network authentication using Kerberos involves four processes: authentication exchange, ticket-granting service exchange, user/server exchange, and secure communications between user and server [17].

Authentication Exchange

The user sends a request to the authentication server for a ticket to the *ticket granting server* (TGS). The *authentication server* (AS) looks up the user in its database and finds the client's secret key, then generates a session key (SK1) for use between the client and the TGS. The AS encrypts the session key using the user's secret key to form a message. The AS also uses the TGS's secret key (known only to the authentication server and the TGS) to encrypt the session key and the user's name to form a *ticket granting ticket* (TGT). The TGT and the message are sent back to the user.

Ticket Granting Service Exchange

The user decrypts the message and recovers the session key. The user creates an authenticator by encrypting the user's name, IP address, and a time stamp with the session key. The user sends this authenticator, along with the TGT, to the TGS, requesting access to the target server. The TGS decrypts the TGT to recover SK1 and then uses the SK1 inside the TGT to decrypt the authenticator. It verifies information in the authenticator, the ticket, the user's network address, and the time stamp. If everything matches, it lets the request proceed.

Then the TGS creates a new session key (SK2) for the user and target server to use, encrypts it using SK1, and sends it to the user. The TGS also sends a ticket containing the user's name, network address, a time stamp, and an expiration time for the ticket—all encrypted with the target server's secret key—and the name of the server.

User/Server Exchange

The user decrypts the message and gets the SK2. Finally ready to approach the target server, the user creates a new authenticator encrypted with SK2. The user sends the session ticket (already encrypted with the target server's secret key) and the encrypted authenticator. Because the authenticator contains plaintext encrypted with SK2, it proves that the user knows the key. The encrypted time stamp prevents an eavesdropper from recording both the ticket and authenticator and replaying them later. The target server decrypts and checks the ticket, authenticator, user address, and time stamp. For applications that require two-way authentication, the target server returns a message consisting of the time stamp plus 1, encrypted with SK2. This proves to the user that the server actually knew its own secret key and thus could decrypt the ticket and the authenticator.

Secure Communications

The target server knows that the user is who he claims to be, and the two now share an encryption key for secure communications. Because only the user and target server share this key, they can assume that a recent message encrypted in that key originated with the other party.

Disadvantages of Kerberos

The Kerberos authentication system has an inherent disadvantage: If an attacker logs on to the same computer at the same time as an authorized user, the cached keys located on that computer are accessible to the attacker.

The Kerberos system relies on the synchronization of the clocks located on the different machines. If an intruder can mislead a host in terms of the correct time, the authentication ticket to the network can be used repeatedly as a result of the nonexpiring time stamp.

Kerebos must trust that all three machines [time/authenticator servers (KDC), the client, and the network server] are void of an intruder. If a ticket is forwarded, the system must trust all of the other systems that the ticket has traveled through before reaching the current server. However, the server in which the ticket arrives cannot tell where it has come from—it can only tell that it has been on other servers by a flag, which has been set to 1.

Passwords can be guessed by plugging a password "guess" into the public encryption key algorithm. The longer a ticket is granted, the more likely it is to be stolen and used by an unauthorized user. In a wireless system using MAC address registration as an authentication method, if the NIC is stolen, the card has the inherent authentication of the user that is tied to that NIC and will be granted access to the network [18].

Standards

Kerberos version 5 is standardized under RFC 1521 [19].

Conclusion

At the time of this writing, there has been much coverage in the press regarding potential security holes in 802.11 security measures. In many instances in those press stories, the 802.11 network managers failed to enable even the most basic security measures built into 802.11. This is the equivalent of leaving one's door unlocked and has little to do with the security of 802.11. Any security planning should start with an equation that figures in what is to be secured (bank records, military intelligence, jokes from Aunt Nancy, and so on) and what is perceived to be the threat (foreign intelligence services, cyber bank robbers, the casual hacker, an eavesdropping neighbor) as measured by the resources (financial) available to defend against those perceived threats to network security.

The 802.11 specification has a number of measures, including WEP, built into it to protect a network from external threats. Should the network manager not feel that WEP is adequate to protect the network based on the above equation, a number of other measures can be added to the network to heighten the level of security in the network. This chapter has explored these added security measures in detail. It should be stated that no network is absolutely secure. With the addition of external security measures, 802.11 networks can be as secure as most wired networks.

By securing the 802.11 network, the odds of a Vo802.11 conversation being listened in on or fraud perpetrated via some form of network intrusion becomes rather remote. Security remains a top concern for 802.11 service providers and vendors alike. Expect new and exciting products and services to emerge in this space.

References

[1] Reynolds, M., "What's Up with WEP: Strategy, Trends and Tactics," Gartner Group, August 2001.

[2] Stallings, W., *Network and Internetwork Security: Principles and Practice,* Upper Saddle River, NJ: Prentice Hall, 1995.

[3] LaRocca, J., and R. LaRocca, *802.11 Demystified,* New York: McGraw-Hill, 2002.

[4] Weatherspoon, S., "Overview of IEEE 802.11b Security," *Intel Technology Journal,* Q2, 2000.

[5] Ohrtman, F., *Wi-Fi Handbook: Building 802.11b Wireless Networks,* New York: McGraw-Hill, 2003.

[6] "IEEE 802.1x-2001 Standard for Local and Metropolitan Area Networks: Port-Based Network Access Control," IEEE, 2001, http://standards.ieee.org/getieee802/download/802.1x-2001.pdf.

[7] Gast, M., *802.11 Wireless Networks: The Definitive Guide,* Sebastopol, CA: O'Reilly and Associates, 2002, p. 100.

[8] "802.1X," Network World Fusion, 2003, http://www.nwfusion.com/links/Encyclopedia/0-9/474.html.

[9] Cisco Systems, "Cisco Aironet Response to University of Maryland's Paper, 'An Initial Security Analysis of the IEEE 802.1x Standard," San José, CA, 2002, http://www.cisco.com/en/US/products/hw/wireless/ps430/prod_bulletin09186a00800a9e74.html.

[10] "Radius Protocol and Best Practices," http://www.microsoft.com/windows2000/docs/RADIUS_Sec.doc.

[11] "PPP Challenge Handshake Authentication Protocol (CHAP)," http://www.ietf.org/rfc/rfc1994.txt.

[12] Application Note, "Authentication with 802.1x and EAP Across Congested WAN Links," Cisco Systems, San José, CA, August 2002, http://www.cisco.com/warp/public/cc/pd/witc/ao350ap/prodlit/authp_an.htm.

[13] "PPP EAP TLS Authentication Protocol," http://www.ietf.org/rfc/rfc2716.txt.

[14] Josefsson, S., et al., "Protected Extensible Authentication Protocol (PEAP)," Internet Engineering Task Force, 2002, http://www.ietf.org/internet-drafts/draft-josefsson-pppext-eap-tls-eap-07.txt.

[15] AirDefense, "5 Practical Steps to Secure Your Wireless LAN," white paper, p. 3, http://www.airdefense.com.

[16] Needham, R. M., and M. D. Schroeder, "Using Encryption for Authentication in Large Networks of Computers," *Communications of the ACM,* Vol. 21, No. 12, December 1978, pp. 993–999.

[17] Kay, R., "Kerberos," *Computer World,* July 3, 2000, http://www.computerworld.com/news/2000/story/0,11280,46517,00.html.

[18] Bellovin, S. M., and M. Merritt, "Limitations of the Kerberos Authentication System," *Computer Communication Rev.*, Vol. 20, No. 5, 1990, pp. 119–132.

[19] Kohl, J., and C. Neuman, "The Kerebos Network Authentication Service (V5)," *Internet Engineering Task Force,* 1993, http://www.ietf.org/rfc/rfc1510.txt.

8

Objections Due to Interference and QoS on Vo802.11 Wireless Networks

If Vo802.11 networks are to be a viable bypass of the PSTN, they must deliver a subscriber experience comparable to or better than that of the PSTN. This is especially important with regard to voice services. Incumbent telcos take great pride in delivering good voice quality on their legacy networks with relatively reliable service. The concern many have when it comes to replacing the copper wires or fiber cables of the PSTN with the air waves of 802.11 is that the air waves, given that they are not as controllable or predictable as copper wire or fiber cables, will deliver an inferior QoS or may be susceptible to interference from other emitters in the electromagnetic spectrum.

Voice is a challenging medium to deliver in packet-switched networks. QoS in wired VoIP networks remains a topic of almost endless discussion. Network managers strive to shave milliseconds off the delivery time of voice packets on state-of-the-art IP networks. Given the emphatic focus on QoS in wired packet networks, one of the foremost concerns regarding Vo802.11 is that it cannot deliver the QoS necessary for intelligible voice quality. As with any other telecommunications engineering puzzle, it is only a matter of good engineering to deliver voice quality as good or better than that of the PSTN over Vo802.11 networks. This chapter covers how QoS can be optimized on an 802.11 network. Good voice quality is the by-product of optimal QoS engineering on an 802.11 network.

What will entice consumers, both business and residential, to give up tried and true PSTN service for Vo802.11 service? The primary attraction may be greater bandwidth (11 Mbps versus 56 Kbps) that delivers data services including streaming video, video on demand, videoconferencing, music file sharing, and local and long-distance telephone service for a monthly service fee that is marginally more than that of a combined monthly phone, cable TV, and Internet access bill. With proper engineering it is possible to deliver Vo802.11 with a voice quality that is equal to or even better than the PSTN. Voice quality superior to that of cell phones is certain.

Perhaps the main objection to Vo802.11 is the misperception that the signal will severely suffer from interference from sources external to the network. These sources are reputed to be garage door openers, microwave ovens, cordless phones, and so on. A further concern is that the subscriber must have a direct line of sight from the service provider's transmitter. These concerns segue into concerns about the QoS of 802.11 networks. The transmission of a packetized medium (IP) over wires has its own series of challenges in addition to mere interference. This chapter addresses concerns about interference, line-of-sight, and QoS improvements made possible by 802.11e. There are no problems in telecommunications—only solutions.

As illustrated in Figure 8.1, QoS has to be measured across the total network; that is, it must encompass both the wired and wireless portions of the network. It does no good to have a very high quality service level agreement for IP services via a wired network if one's wireless connection to the AP via a wireless connection suffers from interference or severe latency and vice versa. Latency is measured from endpoint to endpoint across a network.

Interference

What most people think of when referring to QoS in a wireless network actually has to do with interference from other transmission sources. An immediate concern is a profusion of wireless appliances in day-to-day use such as garage door openers, microwave ovens, and cordless phones. The truth is, many of these household appliances do not operate on the same frequency as 802.11 or the power of their emissions is too low or too distant to interfere with 802.11 traffic.

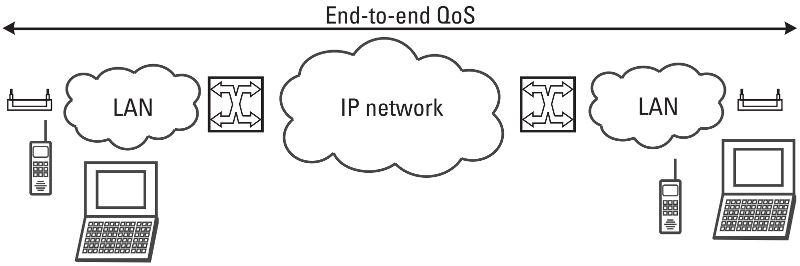


Figure 8.1 QoS is measured from endpoint to endpoint, encompassing both the wired and wireless portions of the network.

A wide variety of other devices (bar-code scanners, industrial lighting, industrial heaters, and home microwave ovens) also use the same frequencies. Because these LANs (and other devices in the ISM band) operate at fairly low power levels, the actual risk of interference is relatively slight, but it does exist. As the popularity of such LANs has increased, situations have developed in which such interference has, indeed, become an issue [1].

External Sources of Interference

Interference can be categorized as having two sources: external and internal. External sources are not related to the 802.11 network itself and are often categorized as some cordless phones, baby monitors, and so on. Internal sources originate in the 802.11 network.

Debunking External Interference Myths

Garage door openers are purported to provide interference to 802.11 LANs (Table 8.1). This is a myth. Garage door openers operate in the 286- to 390-MHz band, so they do not interfere with 802.11. The 900-MHz cordless phones operate in the 802- to 829-MHz ISM band and also do not interfere with 802.11 at all. However, 2.4-GHz cordless phones do operate on the same band as 802.11 and can cause interference. So how does one deal with interference from other applications of the 2.4- and 5.8-GHz bands since FCC Part 15 users are granted use on a noninterference basis?

The FCC licenses 802.11 wireless access points to operate under Class B, §15.247 of the FCC regulations in the 2.4-GHz ISM band. The regulations

Table 8.1
Potential External Sources of Interference to 802.11 Networks

Source of Interference	Discounting Factor or Solution
Garage door opener	Wrong frequency
Microwave oven	Commercial microwaves may have the power to generate enough interference to interfere with a WLAN; residential microwaves do not have the power to generate enough interference to be a factor beyond subscriber's premises.
Cordless phone	Considered to be a nonissue in the industry. Too little power to interfere beyond the immediate residence or office. If subscriber's cordless phone is interfering with subscriber's service, then subscriber should replace 2.4-GHz phone with a 900-MHz cordless phone. Also, why would a residence with cell phone and VoIP-capable 802.11 service still use a PSTN-connected cordless phone?

Source: [2].

state that any device licensed to operate under Part 15 may not interfere with or otherwise disrupt the operation of licensed devices coexisting in the same spectrum. In other words, unlicensed Part 15 devices are the lowest priority, after the federal government, FCC licensed services, and Part 18 devices (ISM transmit-only devices) such as telemetry, radiolocation, and RF heating and lighting, and Part 97 (amateur radio). Also, other unlicensed Part 15 devices under the wrong conditions will interfere such as 2.4-GHz cordless phones, Bluetooth applications, microwave ovens, and 2.4-GHz baby monitors [2, pp. 108–109].

Engineering WLANs to Minimize External Interference

Five parameters should be brought under the control of network planners to minimize external sources of interference:

1. Which channel/band is used;
2. Distance to the interference (further is better)/distance to intended signal (closer is better);
3. Power levels of interference (lower is better);
4. Antenna beam widths;
5. Which protocol is used.

Changing Channels

Sometimes the easiest thing to do is to change the channel to an unused or less congested channel. The specifications for both 802.11a and 802.11 stipulate multiple channels or frequencies. If interference is being encountered on one frequency, then it is merely a matter of switching frequencies to a channel that is not being interfered with. The 802.11b specification provides 11 overlapping channels for North America (Table 8.2), each channel being 22 MHz in width, and each channel centered at 5-MHz intervals (beginning at 2.412 GHz and ending at 2.462 GHz). This means that there are only three channels that do *not* overlap (channels 1, 6, 11).

The 802.11a specification provides 12 channels, each channel being 20 MHz wide, and each centered at 20-MHz intervals (beginning at 5.180 GHz and ending at 5.320 GHz for the upper and middle U-NII bands, and beginning at 5.745 GHz and ending at 5.805 GHz for the upper U-NII band). It is important to note that none of these channels overlaps [3].

Hope for the amelioration of this problem lies in the deployment of 802.11a and 802.11g standards. The 5-GHz UNII band is far less congested, and Wi-Fi has a greater amount of spectrum in which to operate. More channels are permitted, and the standards bodies are working on protocols that allow multiple access points to negotiate among themselves automatically for the proper

Table 8.2
The 11 Overlapping 802.11b Channels

Channel	Frequency (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

Source: [2].

frequency allocation. The 802.11g standard uses OFDM on 2.4 GHz, which is less susceptible to interference and provides more channels. However, the operational range of both 802.11g and 802.11a may be an issue in larger environments [4].

Once a source of interference has been identified, a common practice among WISPs is to negotiate among broadcasters (the WISPs) which WISPs will transmit on what frequency. If such an arrangement cannot be achieved, there are multiple channels to switch to in order to avoid interference.

Dealing with Distance

As described earlier in Chapter 6 on the link budget, the delivery of an intelligible signal is a function of both power of the signal and distance between transmitter and receiver. A signal on the same frequency as the 802.11 WLAN, for example, will not interfere if the source is too distant. That is, the interfering signal becomes too weak to present interference. In addition, if the distance between the AP and the subscriber device is greater than optimal, the signal becomes weak over the distance and becomes susceptible to interference because the interfering signal can be greater than the desired signal.

Where 802.11 is used as a last mile solution, providing access to a residence or small business, the potential sources of interference must be considered. If sources of interference (cordless phones or microwave ovens) can be eliminated within the residence or small enterprise, then the second possible source of interference would come from neighboring residences. The potential for those sources of interference are limited by the distance to the subscriber's network and the power level of that interference. Household appliances such as microwave ovens and cordless phones generate too little power to offer interference beyond the building in which they are located, unless the device is defective. For example, the door seal may need to be replaced. In this case, the defective microwave oven is a hazard in itself [2].

Engineering with Power

Power levels of the primary and interfering signals must also be taken into account. If the power level of the interfering signal gets close to the power level of the intended 802.11 or other WLAN signal, then interference will occur. The simplest solution is to increase the power level of the WLAN signal in order to overcome the interfering signal. The limitation here is that the service provider must not interfere with licensed spectrum operators on similar (unlikely) spectrum. The other solution is that the power level of the interfering signal must be reduced. However, it is important to understand that increasing power can cause interference for other users of the band and that there are legal power output limits set by FCC regulations [5].

Antenna Beamwidths

Another way to eliminate interference is to use antennas to shape where the transmitter's signal goes and where the receiver will listen. A narrow beamwidth antenna can increase the effective power toward the receiver and also increase the signal strength of the received signal.

Another engineering approach to over come QoS issues is to use smart antenna technology. Steering the antenna can do the following for the wireless network: improve the SNR with beam forming, reduce interference due to channel reuse, and mitigate intersymbol interference in multipath environments. Much of this technology falls under the heading of MIMO (*multiple-in, multiple out*).

San Francisco–based Vivato is now marketing their Wi-Fi, the first of its kind. Wi-Fi switches deliver the power of network switching with phased-array radio antennas. These Wi-Fi switches use phased-array radio antennas to create highly directed, narrow beams of Wi-Fi transmissions. The Wi-Fi beams are created on a packet-by-packet basis. Vivato calls this technology *PacketSteering*.

Unlike current wireless LAN broadcasting, Vivato's switched beam is focused in a controlled pattern and pointed precisely at the desired client device. These narrow beams of Wi-Fi enable simultaneous Wi-Fi transmissions to many devices in different directions, thus enabling parallel operations to many users—the essence of Wi-Fi switching. These narrow beams also reduce cochannel interference, because they are powered only when needed [6].

Protocol

By using 802.11g instead of 802.11b, one can gain the advantage of OFDM, which is less susceptible to interference and multipath.

Other Options for Controlling External Interference

Other things that can be done to control interference include the following: (1) Control the environment so that the interference is limited by controlling the devices used. For example, many airports are requiring all RF applications to be run through their frequency coordinators. (2) Choose the "correct" band for each application. A wireless ISP may consider using 802.11a so that their WWAN does not interfere with a possible WLAN solution.

Internal Sources of Interference

Thus far we have focused on sources of interference that are external to the wireless network. As mentioned in the introduction, there are a number of challenges that arise from *within* a wireless network due to the nature of wireless transmissions. These sources of interference include multipath and channel noise. Both can be engineered out of the network.

Multipath and Fade Margin

Multipath interference occurs when waves emitted by the transmitter travel along a different path and interfere destructively with waves traveling on a direct line-of-sight path (Figure 8.2). This is sometimes referred to as *signal fading.*

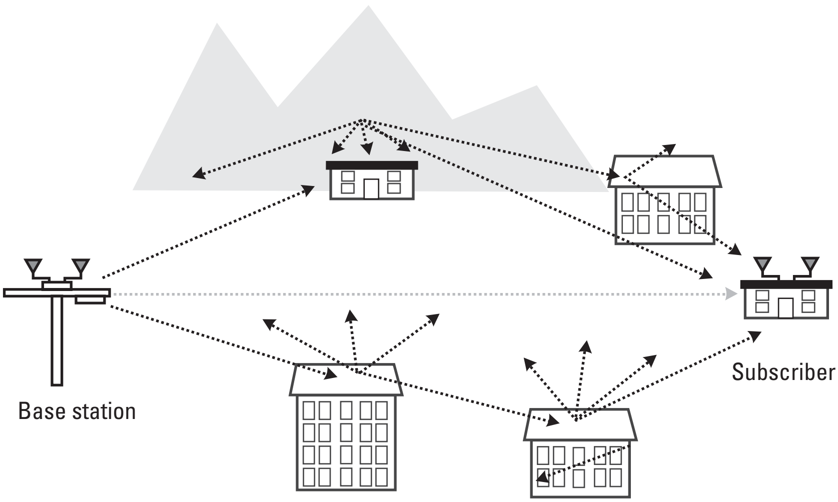


Figure 8.2 Multipath interference. (*From:* [7]. © 1999 Cisco Systems, Inc. Reprinted with permission.)

This phenomenon occurs because waves traveling along different paths may be completely out of phase when they reach the antenna, thereby canceling each other. Because signal cancellation is almost never complete, one method of overcoming this problem is to transmit more power. In an indoor environment, multipath is almost always present and tends to be dynamic (constantly varying). Severe fading due to multipath can result in a signal reduction of more than 30 dB. It is therefore essential to provide an adequate link margin to overcome this loss when designing a wireless system. Failure to do so will adversely affect reliability. The amount of extra RF power radiated to overcome this phenomenon is referred to as *fade margin* or *system operating margin.* The exact amount of fade margin required depends on the desired reliability of the link, but a good rule of thumb for 802.11 protocols is 15 to 20 dB for a 95% confidence interval [8].

One method of mitigating the effects of multipath is by implementing antenna diversity. Because the cancellation of radio waves is geometry dependent, use of two (or more) antennas separated by at least half of a wavelength can drastically mitigate this problem. On acquisition of a signal, the receiver checks each antenna and simply selects the antenna with the best signal quality. This reduces, but does not eliminate, the required link margin that would otherwise be needed for a system that does not employ diversity. The downside is this approach requires more antennas and a more complicated receiver design. Another method of dealing with the multipath problem is via the use of an adaptive channel equalizer. Adaptive equalization can be used with or without antenna diversity.

After the signal is received and digitized, it is fed through a series of adaptive delay stages, which are summed via feedback loops. This technique is particularly effective in slowly changing environments such as transmission over telephone lines, but is more difficult to implement in rapidly changing environments like factory floors, offices, and homes where transmitters and receivers are moving in relation to each other. The main drawback is the impact on system cost and complexity. Adaptive equalizers can be expensive to implement for broadband data links.

Spread spectrum systems are fairly robust in the presence of multipath. DSSS systems will reject reflected signals, which are significantly delayed relative to the direct path or strongest signal. This is the same property that allows multiple users to share the same bandwidth in CDMA systems. However, 802.11's DSSS does not have enough processing gain and orthogonal spreading codes to do this. FHSSs also exhibit some degree of immunity to multipath. Because an FHSS transmitter is continuously changing frequencies, it will always hop to some frequencies that experience little or no multipath loss. In a severe fading environment, throughput of an FHSS system will be reduced, but it is unlikely that the link will be lost completely. OFDM systems such as

802.11a and 802.11g transmit on multiple subcarriers on different frequencies at the same time. Multipath is limited in much the same way that it is limited in an FHSS system. Also, OFDM specifies a slower symbol rate to reduce the chance that a signal will encroach on the following signal, thus minimizing multipath interference [2].

Channel Noise

When evaluating a wireless link, the three most important questions to be answered are these:

1. How much RF power is available?
2. How much bandwidth is available?
3. What is the required reliability as defined by the *bit error rate* (BER)?

In general, RF power and bandwidth effectively place an upper bound on the capacity of a communications link. The upper limit in terms of data rate is given by Shannon's channel capacity theorem

$$C = B \times \log 2\left(1 + S/N\right)$$

where

C = channel capacity (bps);
B = channel bandwidth (Hz);
S = signal strength (W);
N = noise power (W).

Note that this equation means that for an ideal system, the BER will approach zero if the data transmission rate is below the channel capacity. In the "real world," the degree to which a practical system can approach this limit is dependent on modulation technique and receiver noise.

For all communications systems, channel noise is intimately tied to bandwidth. All objects that have heat emit RF energy in the form of random (Gaussian) noise. The amount of radiation emitted can be calculated by

$$N\left[\text{watts}\right] = kT\ B$$

where

k = Boltzmann's constant (1.38×10^{-23} J/K);
T = system temperature (K), usually assumed to be 290K;
B = channel bandwidth (Hz), predetection.

This is the lowest possible noise level for a system with a given physical temperature. For most applications, temperature is typically assumed to be room temperature (290K). The preceding two equations demonstrate that RF power and bandwidth can be traded off to achieve a given performance level (as defined by the BER) [2]. This implies that using a lower data rate that occupies a lower channel bandwidth will provide better range.

If You Want Interference, Call the Black Ravens

One of the coauthor's first "real" jobs was that of intelligence officer for the Tactical Electronic Warfare Squadron 135 (abbreviated VAQ-135 and nicknamed the "World Famous Black Ravens") of the U.S. Navy. This squadron flew the EA-6B tactical jamming aircraft. The airplane is equipped with an ALQ-99 jamming system and is used to jam enemy radar and radio communications in a tactical role. It has been rumored for many years that the squadron's four aircraft, strategically positioned, could shut down most of the electromagnetic spectrum of the United States (TV, radio, and so on).

In a strategic role during the Cold War, the U.S. Air Force developed the B-52G, which was a bomber equipped with an extensive suite of electronic jamming equipment designed to defeat the Soviet air defenses. This would require overwhelming air defense overlapping radar networks that operated at a variety of frequencies. It would also deliver overwhelming interference on air defense radio communications making the airwaves unusable for the Soviets. By shutting down Soviet air defense radars and negating their ability to communicate by radio, the B-52G would clear a path for itself and other strategic bombers to targets for destruction by nuclear attack. A trivia question on student exams at the U.S. Navy's Electronic Warfare School in the 1980s was "What is the electromagnetic coverage of the B-52G jamming system?" The correct answer was "DC (direct current) to daylight."

Given the potential for interference at tactical and strategic military applications relative to the electromagnetic spectrum, it is trivial to negate the benefits of wireless networks based on perceived interference from garage door openers, home microwave ovens, or cordless phones. With proper network engineering, interference from household appliances can be negated [2].

Line of Sight, Near Line of Sight, and Nonline of Sight

Another objection raised to 802.11 networks is the perception that the subscriber must be located on a direct line of sight from the transmitter. This, some detractors fear, would necessitate an unsightly forest of roof antennas in residential settings not unlike the TV antennas of the 1950s and 1960s. Given the threat to the aesthetics of some residential neighborhoods, some zoning commissions might attempt to outlaw antennas for wireless networks. Others fret that many households would not be able to benefit from wireless broadband services because their homes would not be located on a direct line of sight from the service provider's transmitters. While these are not insignificant considerations, the following sections address engineering for delivery of service to the maximum number of subscribers in reach of the service provider's transmitters.

Fresnel Zone and Line-of-Sight Considerations

Line of sight in microwave includes an area around the path called the *Fresnel zone*. The Fresnel zone is an elliptical area immediately surrounding the visual path. It varies depending on the length of the signal path and the frequency of the signal. The Fresnel zone can be calculated, and it must be taken into account when designing a wireless link. Any object within the Fresnel zone will attenuate the transmission path between two points. The maximum radius of the Fresnel zone can be calculated by the following formula:

$$R = 43.3 \times \text{sqrt}(d/4f)$$

where d is the distance in miles and f is the frequency in gigahertz. Thus, for a 5-mile link at 2.4 GHz, the radius is 31.25 feet.

Line of sight (LOS) refers to the situation in which there is a direct, unobstructed path from the transmitter to the receiver. This usually means in any wireless network that the transmission will suffer less degradation if there is an object(s) obstructing that path. LOS is the best possible configuration for transmission on 802.11 networks.

Nonline of sight (NLOS) refers to a situation in which the radio link is blocked. However, with proper engineering, it is possible to receive 802.11 services without having a direct LOS to the service provider's transmitters. This term usually applies where the service provider has deployed its transceivers in a cell network where there is a backbone that services individual cells. If a subscriber is in a location that is NLOS, they will not be served by the WISP. To reach that prospective customer, the service provider would have to deploy a new, costly base station.

One alternative to a new base station would be "any point-to-multipoint" technology (Figure 8.3) or an ad hoc peer-to-peer network. In the case of an "any point-to-multipoint" network topology, any node already in the network can be used as a relay point to reach the central site. If location 5 is within LOS of location 2, for example, node 2 will start functioning as a "repeater" by simply installing a wide focus antenna connected to its port B. At the new subscriber site (location 5) a transceiver is installed with a directional antenna pointing at location 2 [9].

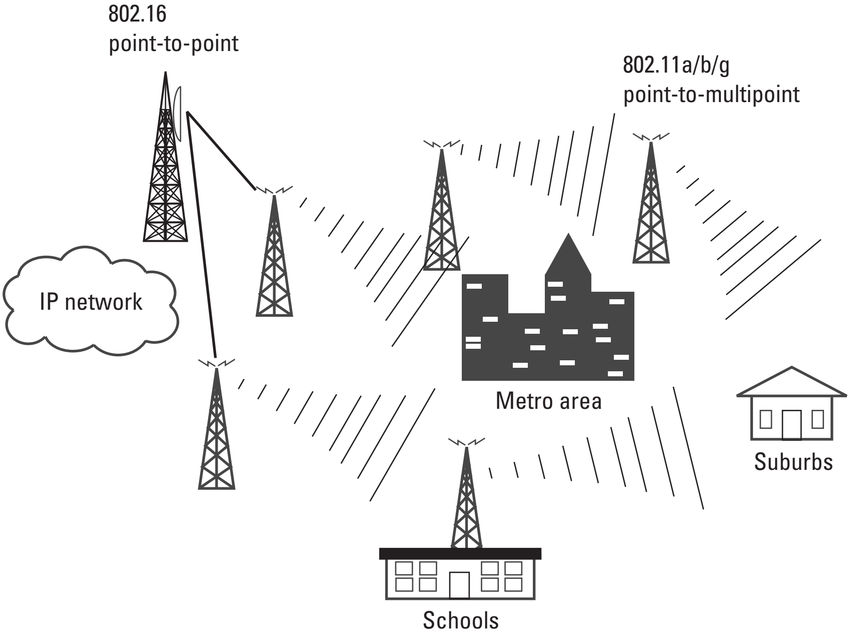


Figure 8.3 Using "anypoint-to-multipoint" technology to reach an NLOS subscriber.

Ad hoc peer-to-peer systems, which are also known as *mesh networks,* can also provide a cost-effective means of providing service to NLOS locations. Single long radio links are replaced with several shorter ones that are less susceptible to noise and multipath. In an ad hoc peer-to-peer network, something as simple as a subscriber device (handheld PDA, cell phone, laptop, and so on) can be used as a repeater to reach an access point or base station. The downside is that current client adapters must use new client software to control the routing function of the subscriber device and change it from infrastructure and ad hoc as needed. The cost of access points and base station technology as well as wireless routers is becoming less expensive as time goes by. Ergo, the ability of a service provider to reach subscribers increases with time. Ditto for potential subscribers if they wish to provide the equipment to receive wireless broadband. Not being LOS to a base station or access point should not prevent them from receiving the benefits of wireless broadband [2, pp. 120–121].

Importance of QoS on 802.11 Networks

When the suggestion is made that 802.11 networks and associated protocols could potentially replace the PSTN as we know it, one of the first considerations is to provide an alternative to the primary service for which the PSTN was built: voice. Voice over a data network requires a great deal of attention to detail in engineering such a network. The primary objection to carrying voice over the Internet Protocol, the primary means of transmitting voice over a packet network, is that the QoS of an IP network is inadequate to deliver intelligible voice to the subscriber. Limitations of an IP network to deliver adequate QoS for voice and video include latency, jitter, and packet loss. By delivering adequate QoS for voice service, 802.11 presents an alternative to the PSTN's voice services. By delivering good QoS for video delivery, the 802.11 network provides an alternative to a cable or satellite TV service.

Need for QoS in Wireless Networks

To deliver voice quality that compares to the PSTN, a network operator must minimize latency, jitter, and packet loss on a Vo802.11 network. An additional network requirement must be supported if the user experience in wireless broadband is to be similar to the user experience in wired broadband (e.g., T1 access). The previous paragraphs detailed how wired IP networks can be engineered to limit latency and other factors that detract from QoS. The IEEE has been grappling with the issue of QoS on wireless networks and has recently approved 802.11e, which is backward compatible with other variants of 802.11, which means that improvements in QoS contained in 802.11e can be applied to 802.11 or 802.11a. This section outlines the mechanisms required to ensure QoS is contained in both 802.11 and 802.11e.

Challenges to Wireless QoS

Many previous attempts at WLAN QoS (and non-QoS channel access schemes), show that strategies that work well in a wired environment do not translate to WLANs. Things that break assumptions include the following: The packet error rate can be in the range of 10% to 20%; bit rates vary according to channel conditions; and bandwidth managers are subject to the "rubber pipe problem," in which the managers do not know how much bandwidth they have to manage, because a neighboring, unrelated bandwidth manager can take some of it at any time. In addition, if a wireless network is to bypass or substitute for the PSTN, it must be able to prioritize voice and video packets over data packets [10].

Latency in Wireless Networks

As discussed earlier in this chapter, the chief threat to an IP network is latency, or delay of the delivery of packets via the network. *Latency* is defined as the time it takes for the network to respond to a user command. If latency is high, causing noticeable delays in downloading Web pages, then the experience feels nothing at all like broadband, no matter how high the data rates are. Low latency (less than 50 ms) is a requirement that must be met if the mass-market adoption of wireless services and devices is to be successful.

The latency experienced by the wireless user has a number of contributing sources, including air link processing, propagation, network processing and transport, a far-end server (if applicable), the application being used, and the user device (Table 8.3). The sum of these latencies must be minimized to ensure a positive end-user experience. Because of the many contributing sources in wired networks, there is little room for latency contributed by the wireless system.

The processing delay leads to another very unique disadvantage of systems that are not built on an all-IP basis. Many networks cannot transmit native IP packets and require "IP assistance" through either a protocol change (transcoding and encapsulation) or through the addition of equipment in the network to "simulate" IP performance. Those measures introduce complexity and packet delays, further impacting the latency of a given system and driving up costs.

Throughput and latency are two essentials for network performance. Taken together, these elements define the "speed" of a network. Whereas throughput is the quantity of data that can pass from source to destination in a specific time, round-trip latency is the time it takes for a single data transaction to occur (i.e., the time between requesting data and receiving it). Latency can also be thought of as the time it takes from data send-off on one end to data retrieval on the other end (from one user to the other).

Latency is crucial to the broadband experience because the Internet is based on TCP. TCP requires the recipient of a packet to acknowledge its receipt. If the sender does not receive a receipt in a certain amount of time

Table 8.3

Types of Delay Encountered on an 802.11 Network

Delay	Definition
Air link processing	The time necessary to convert user data to air link packets (code, modulate, and frame user data) and transmit it
Propagation	The time necessary for a signal to travel the distance between the base station and the subscriber device and vice versa
Network transmission	The time necessary to send the packet across the backhaul and backbone networks, including routing and protocol processing delays and transmission time
Far-end processing	The time required for processing by the far-end servers and other devices

Source: [11].

(milliseconds), then TCP assumes that the connection is congested and slows down the rate at which it sends packets. TCP is very effective in dealing with congestion on the wired networks.

A system's ability to efficiently handle a large user population depends significantly on its ability to service many small TCP/IP messages per unit time and, hence, to multiplex many active data users within a given cell. Hence, high latency translates directly into lower system capacity for serving data users, which equates to higher cost. The ideal mobile data network supports both high peak data rates (3 Mbps) and low packet latency (2 ms), and a unique approach is needed to do this over the wireless medium [11].

QoS in 802.11

The consensus in the industry is that 802.11 by itself does not offer adequate QoS. The IEEE has forwarded a new protocol designed to improve QoS in the original 802.11 MAC to enhance support for QoS-sensitive applications such as VoIP, videoconferencing, and streaming video. The original 802.11 MAC included two modes of operation, DCF and PCF. The 802.11e draft specification introduces two new modes of operation, *enhanced DCF* (EDCF) and *hybrid coordination function* (HCF). As with the original 802.11 MAC, the 802.11e enhancements are designed to work with all possible 802.11 physical layers (original 802.11, 802.11, 802.11a, and 802.11g). The following sections describe QoS efforts in 802.11 and the mechanisms in 802.11e that are designed to improve QoS in wireless networks [12].

Legacy 802.11 MAC

To dissect the progression to 802.11e as a QoS mechanism, it is first necessary to examine the legacy 802.11 MAC. The legacy 802.11 MAC includes support for two access mechanisms, the DCF and the PCF (Figure 8.4). In practice,

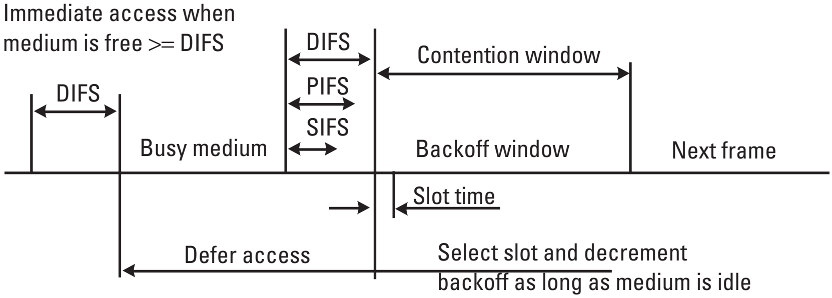


Figure 8.4 Basic access method in DCF and PCF. (*From:* [12]. © 2002 Stanford University. Reprinted with permission.)

almost all, if not all, commercial implementations use DCF exclusively. Refer to Chapter 2 of this book for review of the access mechanisms.

DCF

The basic 802.11 MAC protocol is the Distributed Coordination Function. DCF is based on a CSMA/CA mechanism. CSMA/CA is very similar to Ethernet's CSMA/CD, however, due to the implementation of the wireless transceiver, collision detection is not possible.

In CSMA/CA, stations listen to the medium to determine when it is free. Once a station detects that the medium is free it begins to decrement its back-off counter (a sort of "preemptive" back-off). Each station maintains a *contention window* (CW) that is used to determine the number of slot times a station has to wait before transmission. The back-off counter only begins to decrement after the medium has been free for a DIFS period. If the back-off counter expires and the medium is still free, the station begins to transmit. It is possible that two nodes begin to transmit at the same time in which case a collision occurs. Collisions (or other transmission problems) are detected by the lack of an acknowledgment from the receiver. After the detection of a collision, the station randomly picks a new back-off period from its CW (the CW grows in a binary exponential fashion similar to Ethernet) and then attempts to gain control of the medium again. Due to collisions and the binary back-off mechanism, there are no transmit guarantees with DCF. A pictorial representation of the DCF access mechanism is included in Figure 8.4.

Collision Avoidance Mechanisms

To avoid collisions, the DCF uses mechanisms for sensing whether the medium is in use before transmitting. If the medium is in use, the station will wait according to a predetermined algorithm before attempting to transmit. The DCF supports complementary physical and virtual carrier sense mechanisms.

Because each medium has different characteristics, physical sensing of the medium is called *clear channel assessment* (CCA). For example, a direct sequence radio PHY can be directed to report the medium to be in use in any of three separate conditions. The first condition reports an in-use condition if any energy above a defined threshold is detected on the medium. The second condition reports an in-use condition if any DSSS signal is detected. The last condition reports an in-use condition if a DSSS signal above a defined threshold is detected on the medium. Physical sensing is very efficient, but it is susceptible to the hidden-node problem (cannot sense that which is out of range).

In virtual carrier sensing, no actual physical sensing of the medium occurs. Information about the use of the medium is exchanged through the use of control frames. As opposed to physical carrier sensing, virtual sensing greatly reduces the probability of collisions between hidden nodes on a network. It also reduces the overall throughput. This is due to the additional control frames that must be exchanged. Because this overhead is fixed, the smaller the data frames being sent, the higher the percentage of overhead that is added. In networks with a large amount of small packets or low collision rates, it is best to use only physical sensing. For this reason, the DCF virtual carrier sensing mechanism is optional. The virtual carrier sense control messages are called *request to send* (RTS) and *clear to send* (CTS) frames. A frame size threshold (RTS threshold) can be set that enables a virtual carrier sense procedure only for packets greater than a specified size. The RTS/CTS procedure is not used for broadcast or multicast frames (single frames with multiple destinations) because this could generate multiple conflicting CTS responses. The virtual carrier sense mechanism also helps to avoid collisions when two overlapping BSSs utilize the same radio channel for transmission.

When node A wants to send data, it sends an RTS frame to the AP with addressing and timing information. It sends the address of the node that will receive the impending data frame [*receiver address* (RA)], its own address [*transmitter address* (TA)], and how long it wants to transmit (duration). The calculation of the duration has several elements. An AP receiving an RTS frame replies with a CTS frame. The CTS can be heard by all nodes within the AP's range. In forming the CTS frame, the AP copies the TA from the RTS into the RA of CTS frame, and the AP copies the TA from the RTS into the RA of CTS frame. It also copies the duration field into the CTS after adjusting it for the actual transmission of the CTS. The receipt of a CTS causes the receiver to store the duration field as its *network allocation vector* (NAV). The NAV is a timer that indicates the amount of time that remains before the medium can be used. This value counts down on a regular basis, and when it reaches zero, it indicates that the medium is free. It is updated every time an RTS or CTS with a larger value is received. By combining the physical sensing of the medium with the RTS/CTS procedure, it is possible for a hidden node that is unable to receive from the originating node to avoid collisions with an impending data transmission.

In addition to the RTS and CTS control frames, the DCF CSMA/CA procedure requires an *acknowledgment* (ACK) frame to be sent upon successful receipt of certain types of frames. There is no *negative acknowledgment* (NACK), only a timer that indicates how long to wait for an ACK before the transmission is assumed to be in error. The DCF also provides several frame interval timers based on PHY-specific values. These interval timers represent the time that a station must sense that the medium is idle before starting a transmission. There are two PHY-specific intervals that serve as the basis for the other frame interval timers: the slot time and SIFS. The slot time for a DSSS PHY (20 μs) is defined as the sum of the receive-transmit turnaround time and the energy-detect time including any propagation delay. The slot time for the IEEE 802.11 frequency-hopping PHY is 50 μs. The SIFS is the shortest of the frame interval spaces and is used to allow the completion of an in-progress transmission. The SIFS for the DSSS PHY is 10 μs. The SIFS for the FHSS PHY is 28 μs.

The slot time and the SIFS are used as components in three other frame intervals. These are the DIFS, the *extended interframe space* (EIFS), and the *PCF interframe space* (PIFS). The DIFS is used by the DCFR to enable the transmission of data and management MPDUs. The EIFS is used to enable the processing of frames reported to be erroneous by the PHY layer. The PIFS enables a station to have priority access to the medium when operating in the PCF contention-free mode.

One other timer is used in the DCF virtual CSMA/CA capability: the back-off interval. If a station that wants to transmit detects that a transmission is in progress, it will wait before retrying the transmission. The time it will wait is determined by the back-off algorithm, which is an exponential progression between minimum and maximum values. The starting value for this progression is calculated by multiplying a random number between the minimum and maximum back-off values with the slot time of the PHY. The back-off time is subsequently calculated as sequentially ascending integer powers of 2, minus 1. For example, if the random value is 3 and the slot time is 10 μs, the station would wait 7 or (2 cubed − 1) × 10 μs (70 μs). The retries would then continue using 15 (2^4 −1) and then 31 (2^5 − 1) up to the maximum value between retries. Because a random number is used, two stations entering a transmission entry sequence will usually not arrive at the same back-off interval. This prevents two stations from repeatedly colliding because their retry sequences become synchronized. The station also has a retry counter that can limit the number of retries. Figure 8.5 illustrates the virtual carrier sense protocol.

The DCF carrier sense protocol is a robust method of overcoming the challenges of radio data transmission between network peers. Centralized traffic management such as that provided by an access point is discussed next.

Data Fragmentation

The longer a transmission lasts, the greater the probability that it will be corrupted by interference. To allow the transmission of shorter frames and reduce the likelihood of interference, the IEEE 802.11 MAC provides a method of breaking transmissions into smaller units. This is called *fragmentation.* A value called the *fragmentation threshold* specifies that frames over a specified size should be divided into multiple transmissions. The frame header contains a sequence control field that shows the order of the fragments. Fragments constituting a frame are transmitted immediately after one another without any contention for the medium. Each fragment has its own CRC, and an individual ACK is transmitted for each fragment. The fragment transmissions are separated

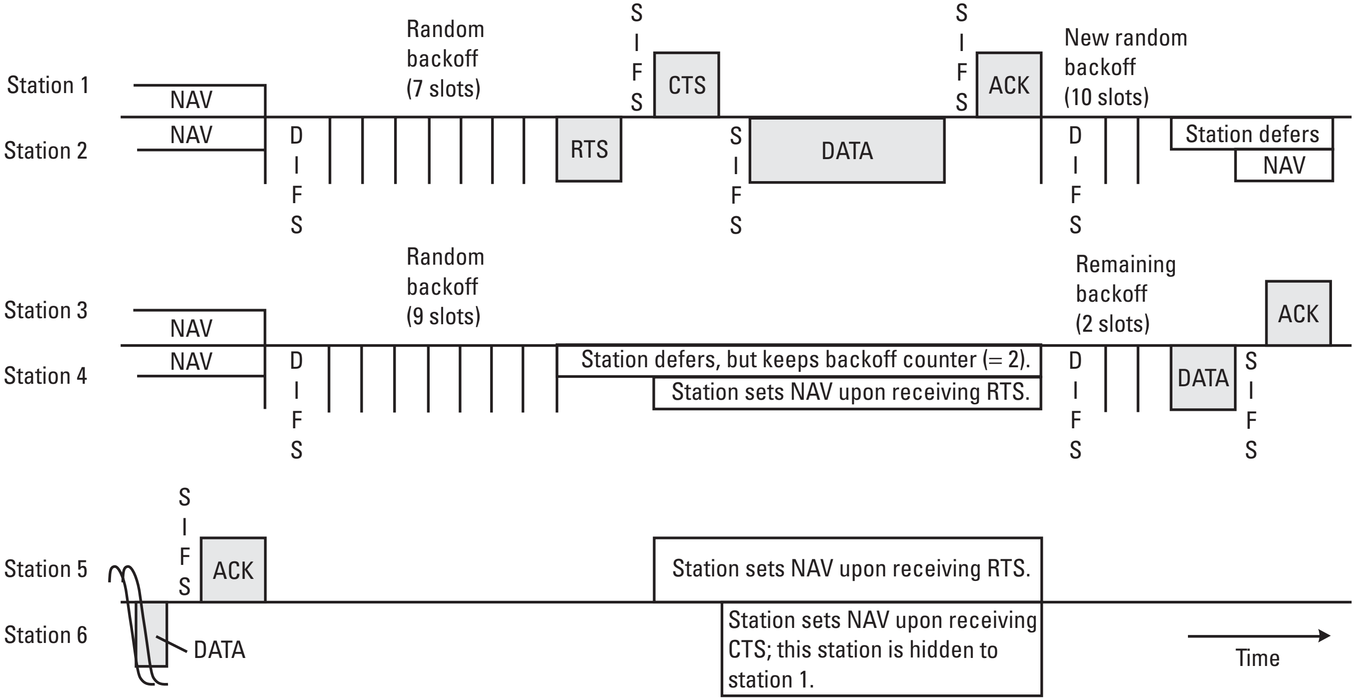


Figure 8.5 Virtual carrier sense protocol. (*From:* [13]. © 2002 Mustafa Ergen. Reprinted with permission.)

by the appropriate frame interval space. The transmission of a sequence of fragments is called a *frame burst.* If an error occurs on a fragment, subsequent fragments are not transmitted until the previous frame is acknowledged. The retransmission and back-off rules apply to fragmented frame transmissions. Duration information in the fragments and ACK frames sets the NAV. Broadcast and multicast frames are not fragmented even if their size exceeds the fragmentation threshold.

Through the use of CSMA/CA and the definition of rules for peer-to-peer as well as centrally managed data transfer, the MAC layer provides reliable structured access to the PHY layer. Physical, diosynchrosies are masked from the upper layers, enabling the LLC functions and the whole suite of TCP/IP protocols [14, pp. 141–142].

PCF

In an attempt to support limited QoS, 802.11 also defined the Point Coordination Function. With PCF, the period after each beacon transmission is divided into two sections, the contention-free period and the contention period, which together constitute a superframe. The point coordinator (generally assumed to be colocated at the AP) is guaranteed access to the medium in the beginning of the contention-free period by beginning transmission before the expiration of the DIFS. During the contention-free period, the point coordinator lets stations have priority access to the medium by polling the stations in a round-robin fashion. The contention-free period is then followed by the contention period, during which access to the medium is governed by DCF.

In PCF the point coordinator has no knowledge of the offered load at each station. The point coordinator simply round-robin polls all stations that have indicated the desire to transmit during the contention-free period. Any station can request to be added to the poll sequence by a special frame exchange sequence during the contention period [14, p. 140].

EDCF defines eight traffic classes. Various parameters governing back-off can be individually set per traffic class. Medium access is similar to DCF with the addition of an *arbitration interframe space* (AIFS). A station cannot begin decrementing the back-off timer until after AIFS. Within a node, each traffic class has a dedicated queue. Traffic class queues contend for access to the virtual channel. Frames that gain access to the virtual channel then contend for medium.

HCF is analogous to PCF but allows a hybrid coordinator to maintain state for nodes and allocate contention-free transmit opportunities intelligently. The hybrid coordinator uses the offered load per traffic class at each station for scheduling.

802.11e MAC Enhancements

To support QoS, many priority schemes are currently being discussed. IEEE 802.11 Task Group E currently defines enhancements to the above-discussed 802.11 MAC, which are called 802.11e. These enhancements introduce two new MAC modes: EDCF and HCF. Both of these QoS-enhanced MAC protocols support up to eight priority levels of traffic, which map directly to the RSVP protocol and other protocol priority levels.

EDCF

The major enhancement provided by EDCF versus DCF is the introduction of eight distinct traffic classes. Aside from this, EDCF, as the name suggests, works in a fashion similar to the DCF MAC, except that some of the elements of the MAC are parameterized on a per-class basis. Figure 8.6 illustrates the functioning of EDCF. Here, each *traffic class* (TC) starts a back-off after detecting the channel being idle for an AIFS. The AIFS is at least as large as the DIFS, and can be chosen individually for each TC. This is the first per-class MAC parameter added in EDCF.

Second, the minimum value of the CW for each traffic class, denoted by CWMin, can be selected on a per-TC basis. In DCF a global constant CWMin is used to initialize all CW values.

Third, when a collision is detected and the CW has to be increased, the value of CW is increased by a *persistence factor* (PF), which is also determined on a per-TC basis. A value of 1 for the PF gives a CW that stays constant even in the case of collisions, whereas a value of 2 (which is the default) gives binary exponential back-off identical to DCF. The equation to calculate the CW in case of a collision is given by

$$newCW[TC] >= ((oldCW[TC] + 1) \times PF[TC]) - 1$$

The CWMax value sets the maximum possible value for the CW on a per-TC basis; however, CWMax is typically intended to remain the same for all traffic classes (at the default valued used in DCF).

Within a station, the eight TCs have independent transmission queues. These behave as virtual stations with the above-mentioned parameters determining their ability to transmit. If the back-off counters of two or more parallel TCs in a single station reach zero at the same time, a scheduler inside the station treats the event as a virtual collision. The *transmit opportunity* (TXOP) is given to the TC with the highest priority of the “colliding” TCs, and the others back off as if a collision on the medium occurred.

The QoS parameters, which are provided on a per-TC basis, can be adapted over time. The base station does this by announcing them periodically

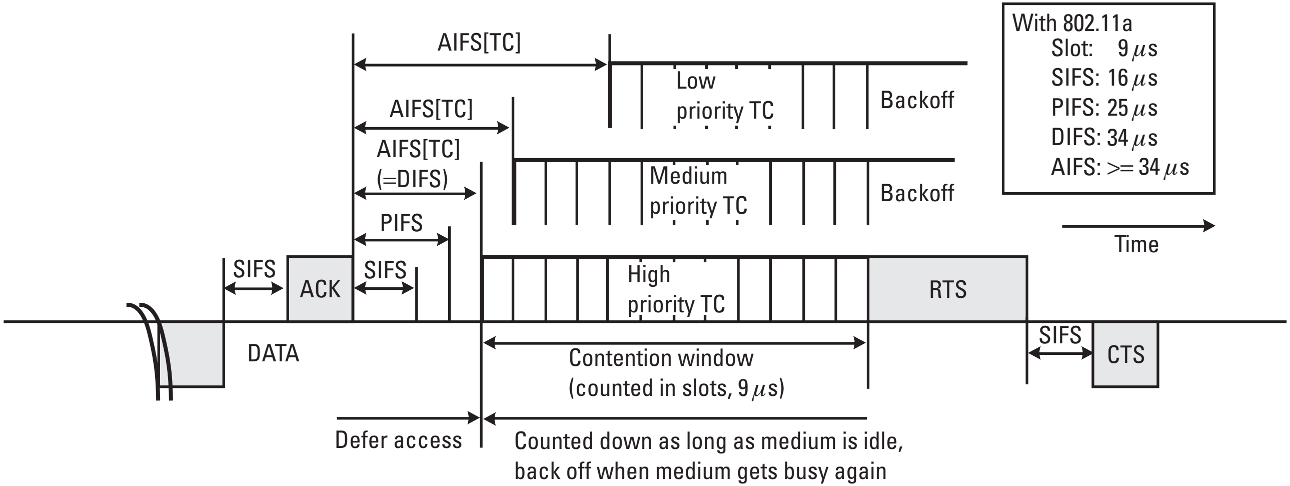


Figure 8.6 IEEE 802.11e. (*From:* [13]. © 2002 Mustafa Ergen. Reprinted with permission.)

via the beacon frames, which are transmitted at the beginning of every superframe.

HCF

HCF is an extension of the polling idea in PCF. Just like in PCF, under HCF, the superframe is divided into the *contention-free period* (CFP) that starts with every beacon, and the *contention period* (CP). During the CP, access is governed by EDCF, although the *hybrid coordinator* (HC, generally colocated at the AP) can initiate HCF access at any time. (Due to its higher priority it can begin transmitting before the expiration of the DIFS.)

During the CFP, the HC issues a QoS CF-Poll to a particular station to give it a TXOP. The HC specifies the starting time and maximum duration as part of the CF-Poll frame. During the CFP, no stations attempt to gain access to the medium, so when a CF-Poll is received, they assume a TXOP and transmit any data they have. The CFP ends after the time announced by the beacon frame or by a CF-End Frame.

If a station is given a CF-Poll, it is expected to start responding with data within an SIFS period. If it does not, the HC can take over the medium after a PIFS period, and allocate another CF-Poll to another station. This allows very efficient use of the medium during the CFP.

To determine which station to give the TXOP to, the HC uses per-station/per-TC queue length data that it collects and maintains to reflect the current snapshot of the infrastructure BSS. The QoS control field that has been added to the MAC frame definition allows stations implementing 802.11e to send queue lengths per TC to the HC.

Scheduling

The MAC defines protocols and mechanisms to perform HCF and EDCF. However, there are a couple of opportunities to perform scheduling decisions that are not determined by pure random number selection as in DCF.

HC Scheduling

The HC has available over time a snapshot view of the per-TC per-station queue length information, including that of the AP itself. With this, it has to decide who to allocate TXOPs to during the CFP. This involves considering, at minimum, the following:

- Priority of the TC;
- Required QoS for the TC (low jitter, high bandwidth, low latency, and so on);
- Queue lengths per TC;

- Queue lengths per station;
- Duration of TXOP available and to be allocated;
- Past QoS seen by the TC.

The practice is to implement a simple scheme of calculating a weighted average queue length per station (weights based on TC queues within a station) and to allocate the maximum available TXOP within the CFP to the station with the largest average. However, various schemes are possible to meet different goals.

Endpoint Scheduling Within TXOP

When a wireless station gets a TXOP by polling from the HC, the HC does not specify a particular TC for the TXOP. This leaves the decision of the TC to service in the TXOP up to the wireless station. This decision can depend on the same factors as for the HC scheduler, except the multiple-station cell-wise aggregation that the HC scheduler uses is not applicable.

By decentralizing this decision, the protocol allows a scalable mechanism for maintaining TC history and servicing them as per QoS seen in the past without collecting this information and detail at the AP, making it unwieldy. We have currently implemented a simple scheme that always sends data for the highest priority TC pending during the TXOP.

EDCF and HCF: QoS in 802.11 Networks

EDCF provides significant improvements for high-priority QoS traffic; however, these improvements are typically provided at the cost of worse performance for lower priority traffic. It also appears that the EDCF parameters can require significant tuning to achieve performance goals. Despite these problems, we find EDCF attractive because of its simplicity as compared to HCF.

HCF, just like its predecessor PCF, provides for much more efficient use of the medium when the medium is heavily loaded. Unlike PCF, HCF does a good job of channel utilization even when the channel is operating well below capacity. Due to reduced overhead, HCF can provide better QoS support for high-priority streams while allocating reasonable bandwidth to lower priority streams.

Both coordination functions are backward compatible with DCF and PCF. This fact, along with our results, leads us to believe that EDCF and HCF will soon see ubiquitous adaptation into mainstream wireless LAN technology [12].

Conclusion

The 802.11e specification is based on more than a decade of experience in design of WLAN protocols and was built from the ground up for real-world wireless conditions. Also, 802.11e is backward compatible with 802.11; that is, non-802.11e terminals can receive QoS-enabled application streams.

This chapter described measures aimed at improving QoS in 802.11 networks with the goal of reducing latency, jitter, and packet loss, which detract from good voice quality. These wireless networks are potentially capable of delivering QoS and voice quality comparable to the PSTN. Note that the RBOCs were losing phone lines to cell phone service providers at an alarming rate (for the RBOCs) during 2002. In fact, the RBOCs have recorded, percentage-wise, their first decline in lines in use since the Great Depression. Cell phone service is admittedly inferior in quality to that of the PSTN, yet given the trade-off in mobility, consumers are accepting a cell phone delivering inferior voice quality over a land line from the PSTN.

The motivating factor for land-line customers to drop their service from the RBOC is the convenience in mobility offered by the cell phone as well as certain price advantages (free long distance in off-peak hours). The point here is that, ultimately, the QoS of the PSTN is not an absolute requirement for consumers. The PSTN is doomed if it must compete with 802.11 in that 802.11 using 802.11e potentially delivers at least comparable QoS in both voice and data services while offering data rates up to 11 Mbps (compared with most DSL plans at 256 Kbps). Given that consumers will trade QoS for convenience and price as witnessed by the loss of lines to cell phone service providers, it is not hard to imagine they would trade off the PSTN for the convenience of greater bandwidth and the wider range of services (video on demand, videoconferencing, and so on) available with that greater bandwidth.

References

[1] Horak, R., “Wireless LANs (WLANs): Focus on 802.11,” http://www.commweb.com/article/COM20020827S0003.

[2] Ohrtman, F., *Wi-Fi Handbook: Building 802.11b Wireless Networks,* New York: McGraw-Hill, 2003.

[3] Linksys, “A Comparison of 802.11a and 802.11 Wireless LAN Standards,” white paper, http://www.linksys.com/products/images/wp_802.asp.

[4] Fine, C., *Watch Out for Wi-Fi*, Goldman Sachs report, September 26, 2002, p. 35.

[5] FCC Regulations Parts 15.247 and 15.407, http://www.fcc.gov.

[6] Vivato, “Vivato Switches Are Changing the Physics of Wireless,” white paper, http://www.vivato.net/prod_tech_technology.html.

[7] Reid, N., “Breakthroughs in Fixed Wireless,” Cisco Systems, 1999.

[8] Zyren, J., and A. Petrick, “Tutorial on Basic Link Budget Analysis,” Intersil white paper, June 1998, http://www.intersil.com.

[9] Bandeira, N., and L. Poulsen, “Broadband Wireless Network Overcomes Line-of-Sight (LOS) Constraints and Lowers Deployment Cost,” Wi-LAN white paper, 2001, p. 5, http://www.wi-lan.com.

[10] Intel, “IEEE 802.11b High Rate Wireless Local Area Networks,” 2000, http://www.intel.com/network/connectivity/resources/doc_library/documents/pdf/wireless_lan.pdf.

[11] Flarion, “Low Latency—The Forgotten Piece of the Mobile Broadband Puzzle,” white paper, http://www.flarion.com.

[12] Priyank, G., et al., “Achieving Higher Throughput and QoS in 802.11 Wireless LANs,” Stanford University white paper, 2002, p. 1, http://nondot.org/~radoshi/cs444n/802_11-Final.html.

[13] Ergen, M., “IEEE 802.11 Overview,” University of California at Berkeley, presentation, May 20, 2002, http://www.eecs.berkeley.edu/~ergen/docs/IEEE-802.11overview.ppt.

[14] LaRocca, J., and R. LaRocca, *802.11 Demystified,* New York: McGraw-Hill, 2002.

9

Engineering Vo802.11 Networks for Maximum QoS

The previous chapter dealt with engineering an 802.11 network that would achieve the best possible QoS for packet delivery over the air waves. This chapter explains measures particular to voice that will deliver the best possible voice quality on a Vo802.11 network.

QoS on Vo802.11 Networks

Despite the fact that telephone companies are losing thousands of lines per month in the United States to cell phone service providers, many perceive that voice over a cell phone connection would deliver inferior voice quality and, as a result, is not a viable alternative to the copper wires of the PSTN. As explored in the previous chapter, a number of new measures (primarily 802.11e) improve the QoS on 802.11.

But what about voice? As wired service providers and network administrators have found, voice is the hardest service to provision on an IP network. New developments in the Vo802.11 industry point to some exciting developments that overcome the chief objection to Vo802.11. Before we discuss these developments, we must first determine what metrics to use in comparing Vo802.11 to the voice quality of the PSTN.

Measuring Voice Quality in Vo802.11

How does one measure the difference in voice quality between a Vo802.11 network and the PSTN? As the VoIP industry matured, new means of measuring voice quality came on the market. Currently, two tests are available that provide a metric for voice quality. The first is a holdover from the circuit-switched voice industry known as the *mean opinion score* (MOS). The other has emerged with the rise in popularity of VoIP and is known as *perceptual speech quality measurement* (PSQM).

MOS

Can voice quality as a function of QoS be measured scientifically? The telephone industry employs a subjective rating system known as the mean opinion score to measure the quality of its telephone connections. The measurement techniques are defined in ITU-T P.800 and are based on the opinions of many testing volunteers who listen to a sample of voice traffic and rate the quality of that transmission. The volunteers listen to a variety of voice samples and are asked to consider factors such as loss, circuit noise, side tone, talker echo, distortion, delay, and other transmission problems. The volunteers then rate the voice samples from 1 to 5 with 5 being "excellent" and 1 being "bad." The voice samples are then awarded a mean opinion score or "MOS." A MOS of 4 is considered "toll quality," that is, equal to the PSTN.

Note here that the voice quality of VoIP applications can be engineered to be as good or better than the PSTN. Recent research performed by the Institute for Telecommunications Sciences in Boulder, Colorado, compared the voice quality of traffic routed through VoIP gateways with the PSTN. Researchers were fed a variety of voice samples and were asked to determine if the sample originated with the PSTN or from the VoIP gateway traffic. The result of the test was that the voice quality of the VoIP gateway routed traffic was "indistinguishable from the PSTN" [1]. Note that the IP network used in this test was a closed network and not the public Internet or other long-distance IP network. This report indicates that quality media gateways can deliver voice quality on the same level as the PSTN. The challenge then shifts to ensuring the IP network can deliver similar QoS to ensure good voice quality. This chapter explains how measures can be taken to engineer voice-specific solutions into a wireless network to ensure voice quality equal to that of the PSTN.

PSQM

Another means of testing voice quality in Vo802.11 networks is known as perceptual speech quality measurement. It is based on ITU-T Recommendation P.861, which specifies a model to map actual audio signals to their representations inside the head of a human. Voice quality consists of a mix of objective and subjective parts and varies widely among the different coding schemes and the types of network topologies used for transport. In PSQM, measurements of processed (compressed, encoded, and so on) signals derived from a speech sample are collected and an objective analysis is performed comparing the original and the processed version of the speech sample (Figure 9.1). From that, an opinion is rendered as to the quality of the signal processing functions that processed the original signal. Unlike MOS scores, PSQM scores result in an absolute number, not a relative comparison between the two signals [2]. The value in this is that vendors can state the PSQM score for a given platform (as assigned by an impartial testing agency). Service providers can then make at least part of their buying decision based on the PSQM score of the Vo802.11 platform.

Detractors to Voice Quality in Vo802.11 Networks

What specifically detracts from good voice quality in an 802.11 environment? Latency, jitter, packet loss, and echo detract from good voice quality in an 802.11 network. With proper engineering, the impact of these factors on voice quality can be minimized and voice quality equal to or better than that of the PSTN can be achieved on 802.11 networks.

Countering Latency on Vo802.11 Networks

Voice as a wireless IP application presents unique challenges for 802.11 networks. Primary among these is acceptable audio quality resulting from minimized network latency (also known as delay) in a mixed voice and data environment. Ethernet, wired or wireless, was not designed for real-time streaming media or guaranteed packet delivery. Congestion on the wireless network, without traffic differentiation, can quickly render voice unusable. QoS measures must be taken to ensure that voice packet delays stay under 100 ms.

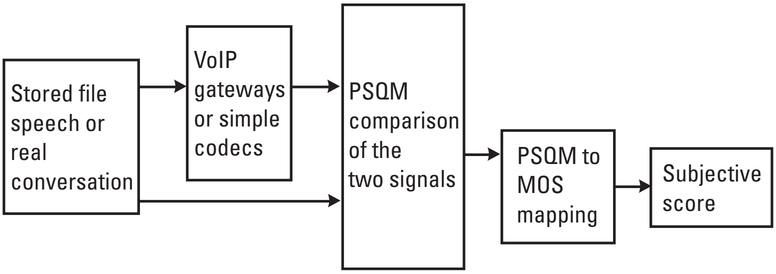

Stored file speech or real conversation
VoIP gateways or simple codecs
PSQM comparison of the two signals
PSQM to MOS mapping
Subjective score

Figure 9.1 Process of PSQM. (*From:* [2]. © 2000 McGraw-Hill, Inc. Reprinted with permission.)

Voice signal processing at the sending and receiving ends, which includes the time required to encode or decode the voice signal from the analog or digital form into the voice-coding scheme selected for the call and vice versa, adds to the delay. Compressing the voice signal will also increase the delay. The greater the compression the greater the delay. Where bandwidth costs are not a concern, a service provider can utilize G.711, which is uncompressed voice (64 Kbps), which imposes a minimum of delay due to the lack of compression.

On the transmitting side, packetization delay is another factor that must be accounted for in the calculations. The packetization delay is the time it takes to fill a packet with data. The larger the packet size the more time is required. Using shorter packet sizes can shorten this delay but will increase the overhead because more packets have to be sent, all containing similar information in the header. Balancing voice quality, packetization delay, and bandwidth utilization efficiency is very important to the service provider [2, pp. 230–231].

How much delay is too much? Of all the factors that degrade Vo802.11, latency (or delay) is the greatest. Recent testing by Mier Labs offers a metric as to how much latency is acceptable or comparable to "toll quality" (i.e., that voice quality offered by the PSTN). Latency of less than 100 ms does not affect "toll-quality" voice. However, latency of greater than 120 ms is discernible to most callers, and at 150 ms the voice quality is noticeably impaired, resulting in less than a toll-quality communication. The challenge for Vo802.11 service providers and their vendors is to get the latency of any conversation on their network to not exceed 100 ms [3]. Humans are intolerant of speech delays of more than about 200 ms. As mentioned earlier, ITU-T G.114 specifies that delay is not to exceed 150 ms one way or 300 ms round-trip. The dilemma is that while elastic applications (e-mail for example) can tolerate a fair amount of delay, they usually try to consume every bit of network capacity they can. In contrast, voice applications need only small amounts of the network, but that amount has to be available immediately [3, 4].

The delay experienced in a call occurs on the transmitting side, in the network, and on the receiving side. Most of the delay on the transmitting side is due to codec delay (packetization and look-ahead) and processing delay. In the network, most of the delay stems from transmission time (serialization and propagation) and router queuing time. Finally, the jitter buffer depth, processing, and, in some implementations, polling intervals add to the delay on the receiving side.

The delay introduced by the speech coder can be divided into algorithmic and processing delay. The algorithmic delay occurs due to framing for block processing, since the encoder produces a set of bits representing a block of speech samples. Furthermore, many coders using block processing also have a look-ahead function that requires a buffering of future speech samples before a block is encoded. This adds to the algorithmic delay. Processing delay is the amount of time it takes to encode and decode a block of speech samples.

Dropped Packets

In Vo802.11 networks, a percentage of the packets can be lost or delayed, especially during periods of congestion. Also, some packets are discarded due to errors that occurred during transmission. Lost, delayed, and damaged packets result in substantial deterioration of voice quality. In conventional error correction techniques used in other protocols, incoming blocks of data containing errors are discarded, and the receiving computer requests the retransmission of the packet. Thus, the message that is finally delivered to the user is exactly the same as the message that originated. Because Vo802.11 systems are time sensitive and cannot wait for retransmission, more sophisticated error detection and correction systems are used to create sound to fill in the gaps. This process stores a portion of the incoming speaker's voice, then, using a complex algorithm to approximate the contents of the missing packets, new sound information is created to enhance the communication. Thus, the sound heard by the receiver is not exactly the sound transmitted, but rather portions of it have been created by the system to enhance the delivered sound [5].

Most of the packet losses occur in the routers, either due to high router load or high link load. In both situations, packets in the queues might be dropped. Another source of packet loss is errors in the transmission links, resulting in CRC errors for the packet. Configuration errors and collisions might also result in packet losses. In nonreal-time applications, packet losses are solved at the protocol layer by retransmission (TCP). For telephony this is not a viable solution since retransmitted packets would arrive too late and be of no use.

Perhaps the chief challenge to Vo802.11 is that, relative to wired networks, packets are dropped at an excessive rate (upwards of 30%). This can lead to distortion of the voice to the extent that the conversation is unintelligible. In VoIP gateways designed for wired networks, one solution is to use a jitter buffer with a "bit bucket." The solution in the wired VoIP industry had been to simply eliminate ("drop") voice packets that arrive late and out of order. This is acceptable if the percentage of late and out-of-order packets is fairly small (say, less than 10%). When the packet loss grows due to the many vagaries of wireless transmissions, the voice quality falls off precipitously.

Jitter

Jitter occurs because packets have varying transmission times. It is caused by different queuing times in the routers and possibly by different routing paths. The jitter results in unequal time spacing between the arriving packets and requires a jitter buffer to ensure smooth, continuous playback of the voice stream.

The chief correction for jitter is to include an adaptive jitter buffer. The jitter buffer described in the solution above is a fixed jitter buffer. An improvement above that is an adaptive jitter buffer that can dynamically adjust to accommodate for the high levels of delay that can be encountered in wireless networks.

Factors Affecting QoS in Vo802.11 Networks

The four most important network parameters for effective transport of Vo802.11 traffic are bandwidth, delay, jitter, echo, and packet loss (Table 9.1). Voice and video quality are highly subjective things to measure. This presents a challenge for network designers who must first focus on these issues in order to deliver the best QoS possible. This section explores the solutions available to service providers that will deliver the best QoS possible.

It is necessary to scrutinize the network for any element that might induce delay, jitter, packet loss, or echo. This includes the hardware elements such as routers and media gateways and also the routing protocols that prioritize voice packets over all other types of traffic on the IP network.

Improving QoS in IP Routers and Gateways

End-to-end delay is the time required for a signal generated at the caller's mouth to reach the listener's ear. Delay is the impairment that receives the most attention in the media gateway industry. It can be corrected via functions contained in the IP network routers, the VOIP gateway, and in engineering in the IP network. The shorter the end-to-end delay, the better the perceived quality and overall user experience.

Sources of Delay: IP Routers

Packet delay is primarily determined by the buffering, queuing, and switching or routing delay of the IP routers. Packet capture delay is the time required to

Table 9.1

Factors Affecting Vo802.11 Voice Quality

Factor	Description
Delay	Latency between transmitting IP packet to receiving packet at destination
Jitter	Variation in arrival times between continuous packets transmitted from point A to point B; caused by packet routing changes, congestion, and processing delays
Bandwidth	Greater bandwidth delivers better voice quality
Packet loss	Percentage of packets never received at the destination

Source: [6].

receive the entire packet before processing and forwarding it through the router. This delay is determined by the packet length, link layer operating parameters, and transmission speed. Using short packets over high-speed networks can easily shorten the delay. Vo802.11 networks use packetization rates to balance connection bandwidth efficiency and packet delay.

Measures for Delivering Optimal QoS on Vo802.11 Networks

QoS requires the cooperation of all logical layers in the IP network—from application to physical media—and of all network elements, from end to end. Clearly, optimizing QoS performance for all traffic types on a Vo802.11 network presents a daunting challenge. To partially address this challenge, several IETF groups have been working on standardized approaches for IP-based QoS technologies. The IETF's approaches fall into the following categories:

- Prioritization using the *Resource Reservation Protocol* (RSVP) and differentiated services (DiffServ);
- Label switching using *multiprotocol label switching* (MPLS);
- Bandwidth management using the subnet bandwidth manager.

To greatly simplify the objection that VoIP voice quality is not equal to that of the PSTN, the network has been engineered to diminish delay and jitter by instituting RSVP, DiffServ, and/or MPLS on the network.

RSVP

A key focus in this industry is to design Vo802.11 networks that will prioritize voice packets over data packets. One of the earlier initiatives, *Integrated Services* (int-serv), developed by the IETF, is characterized by the reservation of network resources prior to the transmission of any packets. The RSVP, defined in RFC 2205, is the signaling protocol that is used to reserve bandwidth on a specific transmission path. RSVP is designed to operate with the OSPF and BGP routing protocols. The int-serv model is comprised of RSVP; an admission control routine, which determines network resource availability; a classifier, which puts packets in specific queues; and a packet scheduler, which schedules packets to meet QoS requirements. The latest development is *Resource Reservation Protocol–Traffic Engineering* (RSVP-TE), a control/signaling protocol that can be used to establish a traffic-engineered path through the router network for high-priority traffic. This traffic-engineered path can operate independently of other traffic classes.

RSVP currently offers two levels of service. The first level is *guaranteed,* which comes as close as possible to circuit emulation. The second level is *controlled load,* which is equivalent to the service that would be provided in a best effort network under no-load conditions. Table 9.2 lists the mechanisms available in conventional packet-forwarding systems that can handle isochronous traffic.

RSVP works where a sender first issues a PATH message to the far end via a number of routers. The PATH message contains a *traffic specification* (Tspec) that provides details about the data packet size. Each RSVP-enabled router along the way establishes a path state that includes the previous source address of the PATH message. The receiver of the PATH message responds with a *reservation request* (RESV) that includes a *flow specification* (flowspec). The flowspec includes a Tspec and information about the type of reservation service requested, such as controlled-load service or guaranteed service.

The RESV message travels back to the sender along the same route that the PATH message took (in reverse). At each router, the requested resources are allocated, assuming that they are available and that the receiver has the authority to make the request. Finally, the RESV message reaches the sender with a confirmation that resources have been reserved [7, pp. 362–363].

Delay is a function of two components. The first is a fixed delay due to the processing within the individual nodes and is only a function of the path taken. The second component of delay is the queuing delay within the various nodes. Queuing is an IP-based QoS mechanism that is available in conventional packet-forwarding systems and can differentiate and appropriately handle isochronous traffic to deliver optimal QoS on Vo802.11 networks. Numerous

Table 9.2

Reservation, Allocation, and Policing Mechanisms Available in Conventional Packet-Forwarding Systems That Can Differentiate and Appropriately Handle Isochronous Traffic

Reservation, Allocation, and Policing	
RSVP	Provides reservation setup and control to enable the resource reservation that integrated services prescribes. Hoses and routers use RSVP to deliver QoS requests to routers along data stream paths and to maintain the router and host state to provide the requested service—usually bandwidth and latency.
TRP	Offers another way to prioritize voice traffic. Voice packets usually rely on the user datagram protocol with RTP headers. RTP treats a range of UDP ports with strict priority.
Committed access rate	CAR, a traffic-policing mechanism, allocates bandwidth commitments and limitations to traffic sources and destinations while specifying policies for handling traffic that exceeds the bandwidth allocation. Either the network’s ingress or application flows can apply CAR thresholds.

Source: [8].

mechanisms are in place to make queuing as efficient as possible, as described in Table 9.3.

Controlled load service (see RFC 2211) is a close approximation of the QoS that an application would receive if the data were being transmitted over a network that was lightly loaded. A high percentage of packets will be delivered successfully and the delay experienced by a high percentage of the packets will not exceed the minimum delay experienced by any successfully delivered packet.

DiffServ

A follow-on IETF initiative is Differentiated Services (diff-serv; see RFC 2474). DiffServ sorts packets that require different network services into different classes. Packets are classified at the network ingress node according to *service level agreements* (SLAs). DiffServ is a set of technologies proposed by the IETF to

Table 9.3
Queuing Mechanisms for Handling Isochronous Traffic

Queuing	**Description**
First-in, first-out (FIFO)	Also known as the best effort service class, FIFO simply forward packets in the order of their arrival.
Priority queuing (PQ)	PQ allows prioritization on some defined criteria, called policies. Four queues—high, medium, normal, and low—are filled with arriving packets according to the policies defined. DSCP packet marking can be used to prioritize such traffic.
Custom queuing (CQ)	CQ allows specific amount of a queue to be allocated to each class while leaving the rest of the queue to be filled in round-robin fashion. It essentially facilitates prioritization multiple classes in queuing.
Weighted fair queuing (WFQ)	WFQ schedules interactive traffic to the front of the queue to reduce response time, then fairly shares the remaining bandwidth among high-bandwidth flows.
Class-based weighted fair queuing (CBWFQ)	CBWFQ combines custom queuing and weighted fair queuing. This strategy gives higher weight to higher priority traffic, defined in classes using WFQ processing.
Low-latency queuing (LLQ)	LLQ brings strict priority queuing to CBWFQ. It gives delay-sensitive data (voice) preferential treatment over other traffic. This mechanism forwards delay-sensitive packets ahead of packets in other queues.

Source: [6].

allow Internet and other IP-based network service providers to offer differentiated levels of service to individual customers and their information streams. On the basis of a *DiffServ code point* (DSCP) marker in the header of each IP packet, the network routers would apply differentiated grades of service to various packet streams, forwarding them according to different *per-hop behaviors* (PHBs). The preferential *grade of service* (GoS), which can only be attempted and not guaranteed, includes a lower level of packet latency because those preferred packets advance to the head of a packet queue should the network suffer congestion [8].

DiffServ improves QoS on Vo802.11 networks by making use of the IP version 4 *type of service* (ToS) field and the equivalent IP version 6 traffic class field. The portion of the ToS/traffic class field that DiffServ uses is known as the *DS field.* The field is used in specific ways to mark a given stream as requiring a particular type of forwarding. The type of forwarding to be applied is known as per-hop behavior, of which DiffServ defines two types: *expedited forwarding* (EF) and *assured forwarding* (AF).

PHB is the treatment that a DiffServ router applies to a packet with a given DSCP value. A router deals with a multiple flows from many sources to many destinations. Many of the flows can have packets marked with a DSCP value that indicates a certain PHB. The set of flows from one node to the next that shares the same DSCP codepoint is known as an *aggregate.* From a DiffServ perspective, a router operates on packets that belong to specific aggregates. When a router is configured to support a given PHB, then the configuration is established in accordance with aggregates rather than to specific flows from a specific source to a specific destination.

EF (RFC 2598) is a service in which a given traffic stream is assigned a minimum departure rate from a given node, that is, one that is greater than the arrival rate at the same node. The arrival rate must not exceed a prearranged maximum. This process ensures that queuing delays are removed. Because queuing delays are the chief cause of end-to-end delay and are the main cause of jitter, this process ensures that delay and jitter are minimized. The objective is to provide low loss, low delay, and low latency such that the service is similar to a virtual leased line. EF can provide a service that is equivalent to a virtual leased line.

The EF PHB can be implemented in a network node in a number of ways. Such a mechanism could enable unlimited preemption of other traffic such that EF traffic always receives access first to outgoing bandwidth. This could, however, lead to unacceptably low performance for non-EF traffic through a token bucket limiter.

AF (RFC 2597) is a service in which packets from a given source are forwarded with a high probability assuming the traffic from the source does not exceed a prearranged maximum. If it does exceed that maximum, the source of the traffic runs the risk that the data will be lumped in with normal best effort IP traffic and will be subject to the same delay and loss possibilities. In a DiffServ network, certain resources will be allocated to certain behavior aggregates, which means that a smaller share is allocated to standard best effort traffic. Receiving best effort service in a DiffServ network could be worse than receiving best effort service in a non-DiffServ network. A given subscriber to a DiffServ network might want the latitude to occasionally exceed the requirements of a given traffic profile without being too harshly penalized. The AF PHB offers this possibility.

The AF PHB allows a provider to offer different levels of forwarding assurances for packets received from a customer. The AF PHB enables packets to be marked with different AF classes and within each class to be marked with different drop-precedence values. Within a router, resources are allocated according to the different AF classes. If the resources allocated to a given class become congested, then packets must be dropped. The packets to be dropped are those that have higher drop-precedence values. The objective is to provide a service that ensures that high-priority packets are forwarded with a greater degree of reliability than packets of a lower priority.

In a DiffServ network, the AF implementation must detect and respond to long-term congestion by dropping packets and respond to short-term congestion, which thus derives a smoothed long-term congestion level. When the smoothed congestion level is below a particular threshold, then no packets should be dropped. If the smoothed congestion level is between a first and second threshold level, then packets with the highest drop precedence level should be dropped. As the congestion level rises, more of the high drop-precedence packets should be dropped until a second congestion threshold is reached. At that point, all of the high drop-precedence packets are dropped. If the congestion continues to rise, then packets of the medium drop-precedence level will also start to be dropped.

The implementation must treat all packets within a given class and precedence level equally. If 50% of packets in a given class and precedence value are to be dropped, then that 50% should be spread evenly across all packets for that class and precedence. Different AF classes are treated independently and are given independent resources. When packets are dropped, they are dropped for a given class and drop-precedence level. Packets of one class and precedence level might possibly experience a 50% drop rate, whereas the packets of a different class with the same precedence level are not dropped at all. Regardless of the number of packets that need to be dropped, a DiffServ node must not reorder AF packets within a given AF class, regardless of their precedence level [7, p. 384].

MPLS-Enabled IP Networks

Multiprotocol label switching has emerged as the preferred technology for providing the best QoS for Vo802.11, traffic engineering, and VPN capabilities on the Internet. MPLS contains forwarding information for IP packets that is separate from the content of the IP header such that a single forwarding paradigm (label swapping) operates in conjunction with multiple routing paradigms. The basic operation of MPLS is to establish *label switched paths* (LSPs) through the network into which certain types of traffic are directed. MPLS provides the flexibility of being able to form *forwarding equivalence classes* (FECs) and the ability to create a forwarding hierarchy via label stacking. All of these techniques facilitate the operation of QoS, traffic engineering, and VPNs. MPLS is similar to DiffServ in that it marks traffic at the entrance to the network. The function of the marking is to determine the next router in the path from source to destination.

MPLS involves the attachment of a short label to a packet in front of the IP header. This procedure is effectively similar to inserting a new layer between the IP layer and the underlying link layer of the OSI model. The label contains all of the information that a router needs to forward a packet. The value of a label can be used to look up the next hop in the path and forward to the next router. The difference between this routing and standard IP routing is that the match is exact. This enables faster routing decisions in routers [7, p. 364].

An MPLS-enabled network, on the other hand, is able to provide low latency and guaranteed traffic paths for voice. Using MPLS, voice traffic can be allocated to an FEC that provides the differentiated service appropriate for this traffic type. Significant work has been done recently to extend MPLS as the common control plane for optical networks [9].

MPLS is not primarily a QoS solution. MPLS is a new switching architecture. Standard IP switching requires every router to analyze the IP header and to make a determination of the next hop, based on the content of that header. The primary driver in determining the next hop is the destination address in the IP header. A comparison of the destination address with entries in a routing table and the longest match between the destination address and the addresses in the routing table determines the next hop. The approach with MPLS is to attach a label to the packet. The content of the table is specified according to an FEC, which is determined at the point of ingress to the network. The packet and label are passed to the next node, where the label is examined and the FEC is determined. This label is then used as a simple look-up in a table that specifies the next hop and a new label to use. The new label is attached and the packet is forwarded.

The major difference between label switching and standard routing based on IP is that the FEC is determined at the point of ingress to the network where information might be available that cannot be indicated in the IP header. The FEC can be chosen based on a combination of destination address, QoS requirements, the ingress router, or a variety of other criteria. The FEC can indicate such information and routing decisions in the network and automatically take that information into account. A given FEC can force a packet to take a particular route through the network without having to cram a list of specific routers into the IP header. This is important for ensuring QoS where the bandwidth that is available on a given path has a direct impact on the perceived quality [7, p. 399].

To date, MPLS is considered one of the best means of engineering a Vo802.11 network to deliver the best possible voice quality. As this technology becomes more widely deployed in IP networks, voice quality on Vo802.11 networks will be of a quality equal to or better than the PSTN.

Bit Rate on Vo802.11 Networks

The *bit rate* (or *compression rate*) is the number of bits per second delivered by the speech encoder; therefore, it determines the bandwidth load on the network. It is important to note that the packet headers (IP, UDP, RTP) also add to the bandwidth. Speech quality generally increases with the bit rate. Very simply put, the greater the bandwidth, the greater the speech quality.

Voice Codecs Designed for Vo802.11 Networks

Many of the detractors to good speech quality in Vo802.11 can be overcome by engineering a variety of fixes into the speech codecs used in both circuit- and packet-switched telephony. The following subsections describe speech coding and how it applies to speech quality, but first we look at the QoS solution:

> **QoS Solution:** Fix circuit-switched voice codecs in a packet-switched, wireless world with enhanced speech processing software.

Circuit-Switched Speech Coding in IP Telephony

The most commonly used codecs for IP telephony today are G.711, G.729, and G.723.1 (at 6.3 Kbps). All of these codecs were designed for, or based on technology designed for, circuit-switched telephony. Mobile telephony has been the major driver for development of speech coding technology in recent years. All of the coders used in mobile telephony, as well as G.729 and G.723.1, are based on the CELP paradigm. These codecs are designed for use in circuit-switched networks and do not work well for packet-switched networks, because their design focuses on handling bit errors rather than packet losses. The important points regarding G.711 as a Vo802.11 codec are that (1) the coder was designed for circuit-switched telephony and (2) it does not include any means to counter packet loss. Insertion of zeros is commonly used when packet loss occurs, leading to a disrupted voice stream (i.e., coming in "broken") and steep degradation of quality with increasing packet losses.

It is possible to introduce error concealment by extrapolating and interpolating received speech segments, which improves quality over zero stuffing. An example is the new Annex I to G.711 called G.711 PLC, which does not always work well and does not guarantee robust operation.

G.729 and G.723.1 belong to a different class of coders compared to G.711. The important points regarding G.729 and G.723.1, as well as other CELP coders, are as follows: (1) The coding paradigm used in these coders was developed for circuit-switched and mobile telephony; (2) the basic speech quality is worse than PSTN quality, that is, they have mobile telephony quality; (3) the coding process is based on interframe dependencies leading to interpacket dependencies; (4) packet loss performance is very poor because of error propagation resulting from interpacket dependencies, thus speech quality degrades rapidly with increasing packet losses; (5) the coders have built-in heuristic error concealment methods and they also suffer from interframe dependencies (for some coders, more frames than the lost one need error concealment); and (6) the coders produce an inflexible bit stream and the packet size is restricted to an integer number of frames, which reduces flexibility.

Modifying Voice Codecs to Improve QoS in Vo802.11 Networks

One of the first processes in the transmission of a telephone call is the conversion of an analog signal (the wave of the voice entering the telephone) into a digital signal. This process is called *pulse code modulation*. This is a four-step process consisting of PAM sampling, companding, quantization, and encoding. Encoding is a critical process in Vo802.11. To date voice codecs used in VoIP (packet switching) are taken directly from PSTN technologies (circuit switching). Cell phone technologies use PSTN voice codecs. New software in the Vo802.11 industry utilizes modified PSTN codecs to deliver voice quality comparable to that of the PSTN.

If circuit-switching voice codecs are the challenge to good QoS in wireless, packet-switched networks, what, then, is the fix for outdated voice codecs? There is an emerging market of enhanced speech processing software that corrects for the shortcomings of traditional voice codecs, which were designed decades ago for a circuit-switched PSTN. These recent developments in Vo802.11b software provide QoS enhancement solutions for IP telephony in the terminal with very high voice quality even with severe network degradations caused by jitter and packet loss. These Vo802.11b QoS enhancements should provide Vo802.11b speech quality comparable to that of the PSTN. Also, speech quality should degrade gradually as packet loss increases. Moderate packet loss percentages should be inaudible.

Enhanced Speech Processing Software

New speech processing algorithms provide for diversity, which means that an entire speech segment is not lost when a single packet is lost. Diversity is achieved by reorganizing the representation of the speech signal. Diversity does not add redundancy or send the same information twice. Ergo, it is bandwidth efficient and ensures that packet losses lead to a gradual and imperceptible degradation of voice quality. The trade-off is that diversity leads to increased delays. Enhanced speech processing software includes advanced signal processing to dynamically minimize delay. Therefore, the overall delay is maintained at approximately the same level as it would be without diversity. Furthermore, the basic quality (no packet loss) is equivalent to or better than PSTN (using G.711).

Enhanced speech processing software is built to enhance existing standards used in IP telephony. This software enables high speech quality on a loaded network with jitter, high packet losses, and delays. Cost savings are realized using enhanced speech processing software because there is no need to overprovision network infrastructure. The high packet loss tolerance also reduces the need for and subsequent cost of network supervision resulting in further cost savings.

Examples of Enhanced Speech Processing Products

Various enhanced speech processing products have been designed:

- *Adaptive jitter buffer.* The use of an adaptive jitter buffer optimizes sound quality by using an advanced adaptive jitter buffer control combined with an error concealment algorithm. This works with any codec such as G.711, G.729, and G.723.1. This arrangement improves the sound quality significantly without any interoperability problems. This solution quickly adapts to the dynamic network conditions of packet-switched networks. This ensures high speech quality with significant latency savings compared to conventional jitter buffering technology.
- *Enhanced G.711.* G.711 with enhancement provides superior packet loss robustness. Enhanced G.711 consists of the G.711 codec combined with an enhancement to provide packet loss robustness. During call setup, the system determines if the recipient also has Enhanced G.711, if so the call will continue using Enhanced G.711; if there is no match, the call will proceed using G.711 on both ends. The enhancement unit is similar in function to encryption methods. The packets are transcoded to prevent packet loss as opposed to privacy. Enhanced G.711 in combination with an adaptive jitter buffer provides a PSTN speech quality level at packet loss/delay rates up to 30%. This is achieved without increasing the bit rate, and without significant increases in latency.
- *Packet loss robustness.* Using a low-bit-rate codec is one method of increasing packet loss robustness: Low-bit-rate codecs use less bandwidth, providing a more efficient use of the available bandwidth. The basic speech quality of one low-bit-rate codec offers better voice quality

than G.729 and G.723.1 and operates at a rate of 13.3 Kbps. Other methods of increasing robustness to packet loss is using an error concealment algorithm.

- *Acoustic echo cancellation.* Echo is often prevalent when using a PC or IP phone. Acoustic echo cancellation is contained in enhanced speech software to counter echo in those platforms [10].

Conclusion

This chapter explained QoS considerations for Vo802.11. QoS on a Vo802.11 network can now be engineered to be superior to that of the PSTN. As with any engineering issue, overcoming shortcomings is merely a matter of good engineering. That good engineering would include such measures as RSVP, Diff-Serv, MPLS, and codecs modified for Vo802.11. With the right mix of QoS measures, Vo802.11 networks can deliver voice quality that is as good as or better than that of the PSTN.

References

[1] Craig, A., “Qualms of Quality Dog Growth of IP Telephony,” *Network News,* November 11, 1999, p. 3.

[2] Douskalis, B., *IP Telephony: The Integration of Robust VoIP Services,* New York: McGraw-Hill, 2000.

[3] Mier Communications, “Lab Report-QoS Solutions,” February 2001, p. 2, http://www.sitaranetworks.com/solutions/pdfs/mier_report.pdf.

[4] McCullough, J., and D. Walker, “Interested in VOIP? How to Proceed,” *Business Communications Review,* April 1999, pp. 16–22.

[5] *White Paper: IP Voice Services*, Report to Congress on Universal Service, CC Docket No. 96-45, March 18, 1998, http://www.von.org/docs/whitepap.pdf.

[6] Ohrtman, F., *Wi-Fi Handbook: Building 802.11b Wireless Networks,* New York: McGraw-Hill, 2003.

[7] Collins, D., *Carrier Grade Voice over IP,* 2nd ed., New York: McGraw-Hill, 2002.

[8] Agarwal, A., “Quality of Service (QoS) in the New Public Network Architecture,” *IEEE Canadian Review,* Fall 2000, p. 1.

[9] Integral Access, “The Evolution Toward Multiservice IP/MPLS Networks,” white paper, 2001, pp. 4–5, http://www.integralaccess.com.

[10] Global IP Sound, “Third-Party Evaluation of Global IP Sound Edge Device QoS Solutions for VoIP,” white paper, http://www.globalipsound.com.

10

Scalability in Wireless VoIP Networks

If Vo802.11 is to replace legacy TDM voice networks, it must be able to scale to the same level as the network it replaces. Unlike a legacy wired network, it must take into consideration not just switching capacity, but bandwidth and spectrum allocation as well. This chapter outlines scalability concerns that network planners should take into consideration when planning a Vo802.11 application (Figure 10.1).

Bandwidth Considerations for Wireless VoIP

The IEEE specification for 802.11b calls for a maximum of 11 Mbps of bandwidth. The G.711 standard for voice traffic calls for 64 Kbps. Very simply put, this would suggest that more than 170 simultaneous G.711 conversations could take place on a single access point. However, VoIP adds considerable overhead per conversation, so the actual amount of bandwidth used in an uncompressed VoIP conversation is well in excess of 64 Kbps, depending on what information is to be included in the headers of the voice packets (could exceed 10 Kbps). Also, the use of 802.11b does not guarantee 11 Mbps of bandwidth. Rather, in the case of hotspots, for example, the source of the bandwidth is a T1 (1.54-Mbps) system ordered from the local telephone company. This would suggest more than 20 simultaneous conversations would be possible based on a 64-Kbps stream for the VoIP conversation before figuring in VoIP overhead. The use of other variants of 802.11, namely, 802.11a, which has a maximum bandwidth of 54 Mbps, would suggest, before figuring in VoIP overhead, the possibility of more than 800 simultaneous conversations. Compressing the voice stream to 8 Kbps (G.729) could increase the number of simultaneous conversations per AP.

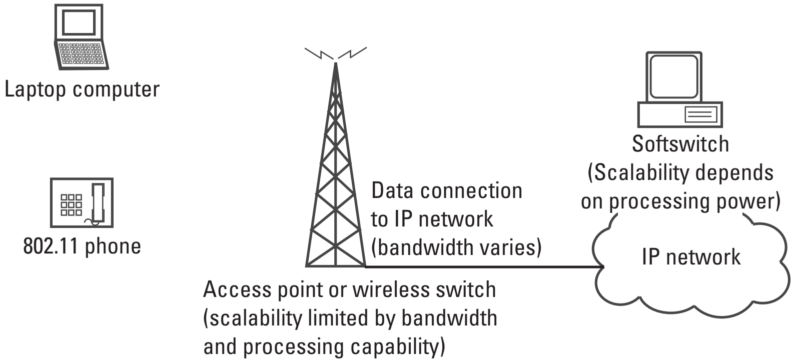


Figure 10.1 Considerations in the scalability of wireless VoIP networks.

Importance of Bandwidth to Scalability

The bandwidth of an AP, which is one determinant of the maximum number of simultaneous conversations, is not infinite through space. Rather, bandwidth diminishes with distance from the access point. Factors such as trees, buildings, and weather can degrade the penetration capabilities of 802.11 through space. This is called *path loss* and is described later. Figure 10.2 illustrates path loss or the degradation of the data link rate, that is, the bandwidth of 802.11 variants through space.

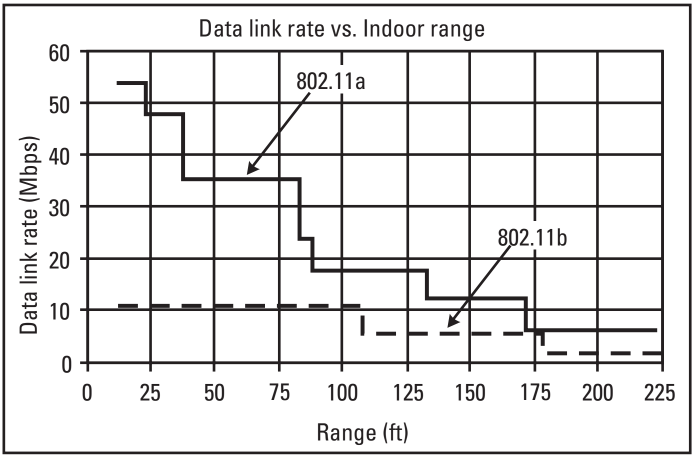


Figure 10.2 The greater the distance from the access point, the greater the degradation in bandwidth, which limits the total number of simultaneous calls per access point. (*From:* [1]. © 2003 Trapeze Networks. Reprinted with permission.)

Which 802.11 Protocols Are Best for Which Vo802.11 Applications?

Four primary standards-based protocols are currently available: 802.11, 802.11b, 802.11a, and 802.11g. At risk of repeating Chapter 2, a brief overview is provided in order to review the strengths and weaknesses for Vo802.11 for each of the 802.11 protocols. The trade-offs between these protocols include bandwidth, frequency, range, and penetration.

802.11b

The most widely used standard protocol, 802.11b, requires DSSS technology, specifying a maximum over-the-air data rate of 11 Mbps and a scheme to reduce the data rate when higher data rates cannot be sustained. This protocol supports 5.5-, 2-, and 1-Mbps over-the-air data rates in addition to 11 Mbps using DSSS and CCK. This IEEE 802.11b standard uses CCK as the modulation scheme to achieve data rates of 5 and 11 Mbps.

The IEEE 802.11b specification allows for the wireless transmission of approximately 11 Mbps of raw data at indoor distances to about 300 feet and outdoor distances of perhaps 20 miles in a point-to-point use of the 2.4-GHz band. The distance depends on impediments, materials, and LOS.

802.11a

The 5.1-GHz band is specified for indoor use only, the 5.2-GHz band is designated for indoor/outdoor use, and the 5.7-GHz band is designated for outdoor use only. RF interference is much less likely because the 5-GHz bands are less crowded. The 5-GHz bands each have four separate nonoverlapping channels. The 802.11a standard specifies OFDM using 52 subcarriers for interference and multipath avoidance, supports a maximum data rate of 54 Mbps using 64QAM, and mandates support of 6-, 12-, and 24-Mbps data rates. Equipment designed for the 5.1-GHz band has an integrated antenna and is not easily modified for higher power output and operation on the other two 5-GHz bands.

802.11g

Variant 802.11g is an extension to 802.11b and operates in the 2.4-GHz band. It increases 802.11b's data rates to 54 Mbps using the same OFDM technology that is used in 802.11a. The range at 54 Mbps is less than that of existing 802.11b APs operating at 11 Mbps. As a result, if an 802.11b cell is upgraded to 802.11g, the high data rates will not be available throughout all areas. Also, 802.11g offers higher data rates and more multipath tolerance than 802.11b. Although there is more interference on the 2.4-GHz band, 802.11g may the protocol of choice for best range and bandwidth combination and it is upwardly compatible with 802.11b equipment.

Why Frequency Bands Are Important

The 802.11 technologies can be deployed on four unlicensed frequency bands in two bands called ISM and U-NII. The 2.4-GHz ISM band has an inherently stronger signal with a longer range and can travel through walls better than the 5-GHz U-NII bands. However, the U-NII band allows more users to be on the same channel simultaneously. The 2.4-GHz ISM band has a maximum of three nonoverlapping 22-MHz channels, whereas the 5-GHz band has four nonoverlapping 20-MHz channels in each of the U-NII bands.

Path Loss Illustrated

The most difficult part of calculating a link budget is the path loss. Outdoors, the free-space loss is well understood. The path loss equation [2] for outdoors can be expressed as follows:

$$\text{Free-space path loss} = 20 \log(d\,[\text{m}]) + 20 \log (f[\text{MHz}]) + 36.6 \text{ dB}$$

At 2.4 GHz, the formula simplifies to

$$\text{Free-space path loss} = 20 \log(d\,[\text{m}]) + 40 \text{ dB}$$

This formula holds true as long as one can see along the LOS from the receiver and the transmitter and have a sufficient amount of area around that path called the *Fresnel zone.* For indoors, this formula is more complicated and depends on factors such as building materials, furniture, and occupants. At 2.4 GHz, one estimate follows this formula:

$$\text{Indoor path loss (2.4 GHz)} = 55 \text{ dB} + 0.3 \text{ dB}/d\,[\text{m}]$$

At 5.7 GHz, the formula looks like this:

$$\text{Indoor path loss (5.7 GHz)} = 63 \text{ dB} + 0.3 \text{ dB}/d\,[\text{m}]$$

Receiving Antenna Gain

The receiving antenna gain adds to the link budget just like the transmitting antenna. Adding gain to an antenna is balanced gain because it adds gain for both transmitting and receiving.

Link Margin

Fade margin is the difference, in decibels, between the magnitude of the received signal at the receiver input and the minimum level of signal determined for reliable operation. The higher the fade margin, the more reliable the link will be. The exact amount of fade margin required depends on the desired reliability of the link, but a good rule-of-thumb is 20 to 30 dB. Fade margin is often referred to as *thermal* or *system operating margin.*

Diffraction Losses

Diffraction occurs when the radio path between the transmitter and receiver is obstructed by a surface that has sharp irregularities or edge. The secondary waves resulting from the obstructing surface are present behind the obstacle. On close-to-LOS obstacles, diffraction losses can be as little as 6 dB. On NLOS obstacles, diffraction losses can be 20 to 40 dB.

Coax and Connector Losses

Cable losses are a function of cable type, thickness, and length. Generally speaking, the thicker and better built the cable, the lower the losses (and the higher the cost). Coax losses are nearly prohibitive in the 2.4- and 5.8-GHz bands. The best option is to use coax cable as sparingly as possible and locate the microwave transceiver as close to the antenna as possible in an environmental enclosure. Connector losses can be estimated at 0.5 dB per connection.

A network planner calculating the capacity of a Vo802.11 network should take into account path loss by using the formulas given earlier. Given the direct relationship between bandwidth and the maximum number of simultaneous conversations, a network planner can determine the maximum number of Vo802.11 users on a given AP.

Frequency Reuse Planning for Vo802.11 Networks

Another determinant of scalability for a Vo802.11 network is frequency reuse in a given service area. If a service provider or enterprise operator were to use the same frequency over a wide area, there would be inevitable interference, which would limit the maximum number of users of a given Vo802.11 network. Determining frequency reuse is an important factor in determining the capacity of a Vo802.11 network. Cell phone service providers have used reuse techniques for years.

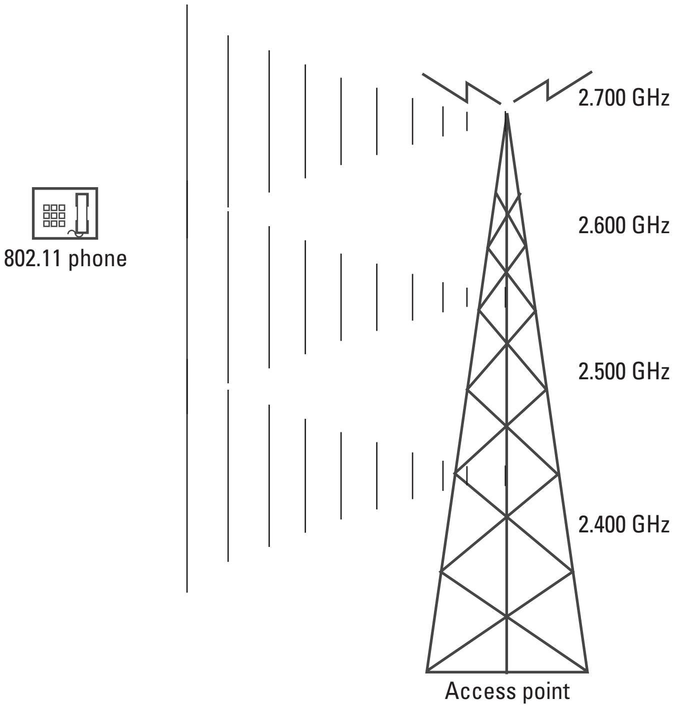


Figure 10.3 The 2.4-GHz band has three nonoverlapping channels. (*From:* [3]. © 2002 Neeli Prasad and Anand Prasad. Reprinted with permission.)

Frequency Reuse at 2.4 GHz

The 2.4-GHz band has eleven 22-MHz-wide channels defined starting at 2.412 GHz and going for every 5 MHz through 2.462 GHz. Three nonoverlapping channels are available—1, 6, and 11—as shown in Figure 10.3. These nonoverlapping channels can be used in a 3-to-1 reuse pattern as shown in Figure 10.4.

Frequency Reuse at 5 GHz

The operating channel center frequencies are defined at every integral multiple of 5 MHz above 5 GHz. The valid operating channel numbers are 36, 40, 44, 48, 52, 56, 60, 64, 149, 153, 157, and 161. The lower and middle 802.11a sub-bands accommodate eight channels in a total bandwidth of 200 MHz. The upper 802.11a band accommodates four channels in a 100-MHz bandwidth. The centers of the outermost channels are 30 MHz from the bands' edges for the lower and middle 802.11a bands and 20 MHz for the upper U-NII band (Figure 10.5).

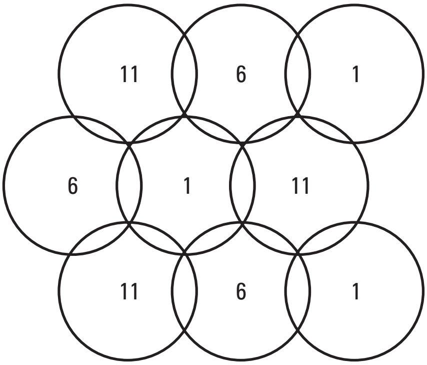


Figure 10.4 A 3-to-1 reuse pattern. (*From:* [3]. © 2002 Neeli Prasad and Anand Prasad. Reprinted with permission.)

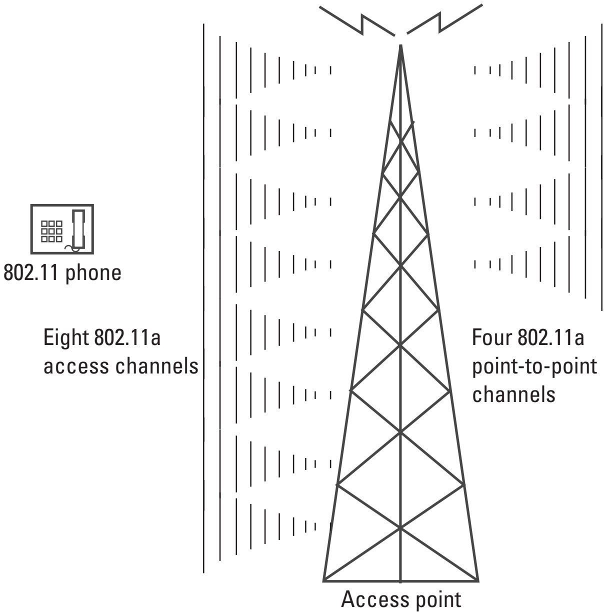

802.11 phone
Eight 802.11a access channels
Four 802.11a point-to-point channels
Access point

Figure 10.5 Valid 5-GHz operating channels.

Point-to-point links operate on the other four channels: 149, 153, 157, and 161. This allows four channels to be used in the same area. The 802.11a APs and client adapter cards operate on eight channels: 36, 40, 44, 48, 52, 56, 60, and 64. This allows two 4-to-1 reuse patterns to be used, as shown in Figure 10.6.

By using both the low-frequency and midfrequency ranges together, we can take advantage of a 7-to-1 reuse pattern with a spare (Figure 10.7). The

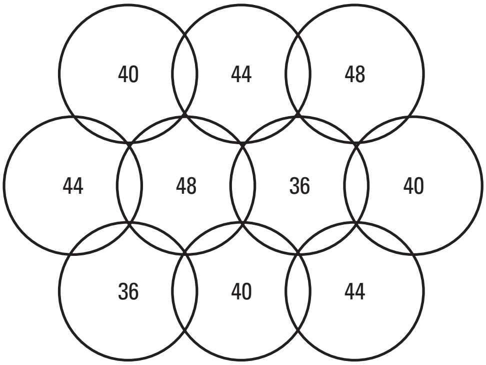


Figure 10.6 A 4-to-1 reuse pattern. (*From:* [3]. © 2002 Neeli Prasad and Anand Prasad. Reprinted with permission.)

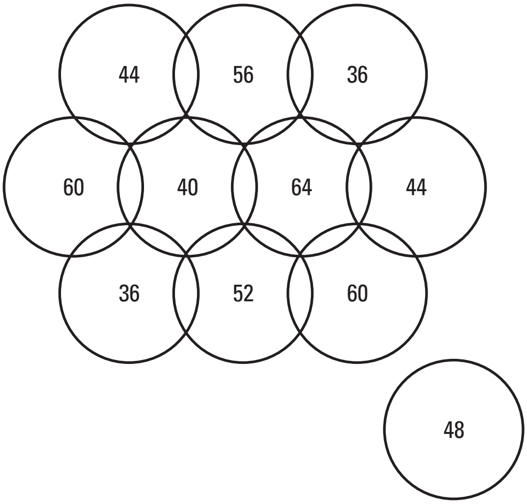


Figure 10.7 A 7-to-1 reuse pattern with a spare. (*From:* [3]. © 2002 Neeli Prasad and Anand Prasad. Reprinted with permission.)

spare can be added for a fill to extend coverage or to add capacity in areas such as conference rooms where more capacity is needed.

Frequency Allocation

For a simple project like one or two APs, simply assign the least used frequencies from the site survey. For more complex projects involving three or more APs, pick a frequency reuse pattern for the frequencies that are used for the project, starting with the most complicated part of the site survey and assigning frequencies. Avoid overlapping channels if possible. If there is an area that has to have an overlap, plan it such that it is naturally an area where the most capacity would be required.

FCC Regulations and Power of Vo802.11 Transmissions

The use of the 802.11 bands is regulated under Parts 15.247 and 15.407 of the FCC regulations [4]. Table 10.1 lists the relevant parts of Part 15.247 regarding power at the time of this writing.

Point-to-Multipoint Links

Part 15.247(b)(1) limits the maximum power at the antenna to 1W. Part 15.247(b)(3) allows antennas that have more than 6 dB as long as the power to the antenna is reduced by an equal amount in the 2.4-GHz band. This implies that the maximum *effective isotropic radiated power* (EIRP) is 4W or 36 dBm. This limit of 4W EIRP regardless of the antenna gain is illustrated in Table 10.2.

Point-to-Point Links

Point-to-point links have a single transmitting point and a single receiving point. Typically, a point-to-point link is used in a building-to-building application. Part 15.247(b)(3)(i) allows the EIRP to increase beyond the 4W limit for point-to-multipoint links in the 2.4-GHz ISM band. For every additional 3-dB gain on the antenna, the transmitter only needs to be cut back by 1 dB.

The so-called "three-for-one rule" for point-to-point links can be observed in Table 10.3. According to Part 15.247(b)(3)(ii), in the 5.8-GHz band, there is

Table 10.1
Frequency Bands and Their Associated FCC Part 15 Power Limits for Point-to-Multipoint Applications

Frequency Range (MHz)	Bandwidth (MHz)	Max Power at Antenna	Max Effective Isotropic Radiated Power	Notes
2,400–2,483.5	83.5	1W (+30 dBm)	4W (+36 dBm)	Point-to-point
—	—	1W (+30 dBm)	—	Point-to-multipoint following 3:1 rule
5,150–5,250	100	50 mW	200 mW (+23 dBm)	Indoor use; must have integral antenna
5,250–5,350	100	250 mW (+24 dBm)	1W (+30 dBm)	—
5,725–5,825	100	1W (+30 dBm)	200W (+53 dBm)	—

Source: [4].

Table 10.2
Point-to-Multipoint Operation in the 2.4-GHz ISM Band

Power at Antenna (mW)	Power at Antenna (dBm)	Max Antenna Gain (dBi)	EIRP (W)	EIRP (dBm)
1,000	30	6	4	36
500	27	9	4	36
250	24	12	4	36
125	21	16	4	36
63	18	19	4	36
31	15	21	4	36
15	12	24	4	36
8	9	27	4	36
4	6	30	4	36

Source: [4].

Table 10.3
Point-to-Point Operation in the 2.4-GHz ISM Band

Power at Antenna (mW)	Power at Antenna (dBm)	Max Antenna Gain (dBi)	EIRP (W)	EIRP (dBm)
1,000	30	6	4	36
794	29	9	6.3	38
631	28	12	10	40
500	27	15	16	42
398	26	18	25	44
316	25	21	39.8	46
250	24	24	63.1	48
200	23	27	100	50
157	22	30	157	52

Source: [4].

no such restriction. However, Part 15.407 effectively restricts the EIRP to 53 dBm as shown in Table 10.4.

Limitations in the AP

Most of the industry is now focused on enterprise applications where Vo802.11 applications allow the employee to roam the premises accessing the wireless

Table 10.4
Point-to-Point Operation in the 5.8-GHz U-NII Band

Power at Antenna (mW)	Power at Antenna (dBm)	Antenna Gain (dBi)	EIRP (W)	EIRP (dBm)
1,000	30	6	4	36
1,000	30	9	8	39
1,000	30	12	16	42
1,000	30	15	316	45
1,000	30	18	63.1	48
1,000	30	21	125	51
1,000	30	23	250	53

Source: [4].

LAN with an 802.11 phone, laptop, or PDA. The question then becomes one of how many wireless VoIP users can access a given access point at a given time? The limitations to scalability are twofold: the bandwidth available and the ability of the access point to process simultaneous sessions.

As a result of these two factors, vendors of Vo802.11 platforms include APs that are specifically designed to handle voice separate from APs that are part of the data network. Some of these voice-specific APs can process about 10 to 12 simultaneous conversations. Once it appears that there is a demand for more than 12 conversations per AP, then the enterprise can add another AP to cover that area (break room, conference room, and so on). At the time this book was written, high-capacity, phased-array access points were coming on the market that allowed hundreds of simultaneous conversations. However, these products have yet to be rigorously tested in the marketplace.

Scalability in VoIP Switching

On the wire-line side of a Vo802.11 architecture, the next bottleneck with regard to scaling is the bandwidth of the connection to the AP or wireless switch. In most cases the constraint will occur at the AP. However, if, for example, a wireless switch capable of processing hundreds of simultaneous wireless sessions is connected to a T1 (1.54 Mbps), then there will not be enough bandwidth to transport the voice and data sessions from that wireless switch to the IP network.

The most relevant restriction on the wire-line side of a wireless VoIP network is the call processing capability of the VoIP softswitch. Some Vo802.11 vendors offer only an interface to a legacy TDM PBX. In this case, scalability is a function of blocking on the TDM PBX. Other vendors (Cisco and Vocera, for example) offer an IP-PBX to perform call setup and teardown. The Cisco CallManager, for example, can handle 512 simultaneous calls.

If a service provider were to contemplate offering Vo802.11 to a mass market, they would need to implement a carrier grade softswitch or IP Centrex with the ability to process millions of BHCAs. This is also expressed in *calls per second* (CPS). The chief limitation on call processing is the computing power of the server on which the softswitch is hosted. The flagship of Class 5 switches, for example, is the Lucent #5ESS switch, which can process 800,000 BHCAs. New softswitches on the market can exceed 5 million BHCAs. Ergo, it can be argued that from a switching perspective, wireless VoIP is much more scalable than the TDM switching technology used in the PSTN.

What is missing from this discussion is the fact that an enterprise 802.11 AP would ordinarily be handling data applications in addition to wireless VoIP sessions. Most APs were designed to handle no more than a dozen simultaneous sessions. The main limitation is the processing capabilities of the AP. The processing capability of the AP is a function of its processing power and its associated TCP/IP stack.

Conclusion

This chapter explored the question of scalability of wireless VoIP applications. Potential bottlenecks exist with regard to bandwidth, spectrum allocation, the capacity of the AP or wireless switch to process multiple sessions, and, finally, the call processing capability of the softswitch.

References

[1] Trapeze Networks, "The Top 10 Myths About Wireless LANs," 2003, http://www.trapezenetworks.com/solutions/market_myths.asp.

[2] Jordan, E. C., *Reference Data for Engineers: Radio, Electronics, Computer, and Communications,* Indianapolis, IN: Howard W. Sams and Co., 1986.

[3] Prasad, N., and A. Prasad, (eds.), *WLAN Systems and Wireless IP for Next Generation Communications,* Norwood, MA: Artech House, 2002.

[4] The FCC Web site, http://www.fcc.gov, has a lot of material. Part 15 in its entirety can be found at http://www.access.gpo.gov/nara/cfr/waisidx_01/47cfr15_01.html.

11

Vo802.11 Reliability

A recurring objection to VoIP, softswitch, and 802.11 solutions is the perception that these solutions do not match the "five 9s" of reliability of the PSTN. This chapter posits that a Vo802.11 solution is just as reliable (actually "available") as legacy TDM solutions and is potentially more reliable. This chapter will explore what is meant by the "five 9s" and why engineering a network to deliver that level of reliability is only a matter of good engineering not limited to Bell Labs designs. "Five 9s" applies to data networks as well as to telephone switches. Data networks have long been engineered to achieve "five 9s." VoIP is voice over a data network. Ergo, a Vo802.11 network can be engineered to deliver "five 9s" of reliability. A wireless VoIP network can be just as, if not more, reliable than the PSTN (Figure 11.1).

Understanding Reliability

Availability is often expressed numerically as a percentage of uninterrupted productive time containing from one to five nines. For instance, 99% availability, or "two 9s," equates to a certain amount of availability versus downtime, as does 99.9% (three 9s), and so on. The downtime calculations shown in Table 11.1 for each of the five 9s are based on 24-hour, year-round operation.

The terms *reliability* and *availability* are often used interchangeably but they are two distinct measures of quality. Reliability refers to component failure rates measured over time, usually a year. Common reliability measures of components are *annual failure rate, failures in time, mean time between failure, mean time to repair,* and *single point of failure* (SPOF), as described in Table 11.2.

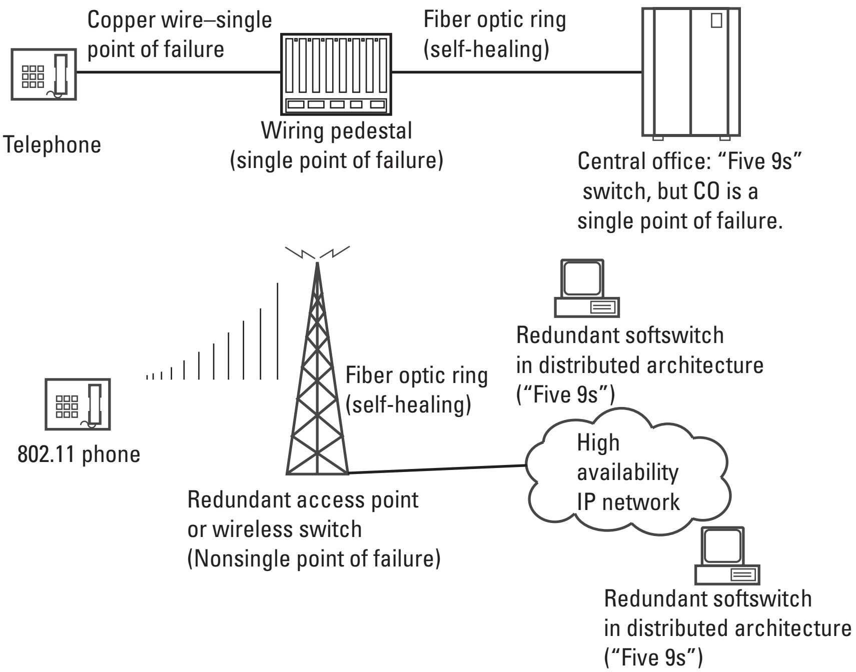

Copper wire–single point of failure
Fiber optic ring (self-healing)
Telephone
Wiring pedestal (single point of failure)
Central office: "Five 9s" switch, but CO is a single point of failure.
Redundant softswitch in distributed architecture ("Five 9s")
Fiber optic ring (self-healing)
802.11 phone
High availability IP network
Redundant access point or wireless switch (Nonsingle point of failure)
Redundant softswitch in distributed architecture ("Five 9s")

Figure 11.1 Properly engineered, wireless VoIP networks can be more reliable than the PSTN.

How Availability Is Calculated

Availability measures reliability and indicates system "uptime" from an operation perspective. System availability is a function of aggregate component reliability, thus availability is likewise measured in terms of time. Availability of a hardware/software module can be obtained by the formula given below for calculating availability:

Table 11.1
Availability and Downtime: How the "Five 9s" Figure Is Calculated

Availability	Downtime
90% (one 9)	36.5 days per year
99% (two 9s)	3.65 days per year
99.9% (three 9s)	8.76 hours per year
99.99% (four 9s)	52.55 minutes per year
99.999% (five 9s)	5.25 minutes per year

Table 11.2
Terms and Definitions Related to Availability

Term	**Definition**
Annual failure rate (AFR)	AFR is the amount of downtime expressed as the relationship between the MTBF and the number of hours in a year (8,760).
Failure in time (FIT)	FIT is the total number of failures of a module in a billion hours (1,000,000,000 hours).
Mean time between failures (MTBF)	MTBF is the average time a manufacturer estimates before a failure occurs in a component or complete system. MTBF is an average and half of the components are expected to fail before that figure and half after.
Mean time to repair (MTTR)	MTTR is an estimate on the part of the vendor as to the average time necessary to do repairs on equipment.
Single point of failure (SPOF)	SPOF refers to a single point or network element at which failure could bring down a network or subnetwork.

Source: [1].

$$A = \text{MTBF}/\text{MTBF} + \text{MTTR}$$

where A is availability and the other abbreviations are as explained earlier. We can also calculate unavailability:

$$U = \text{MTTR}/\text{MTBF} + \text{MTTR}$$

where U is unavailability. In addition,

$$\text{Availability} = 1 - \text{Unavailability}$$

The annual failure rate, using 8,760 hours per year, is calculated as follows:

$$\text{AFR} = 8{,}760/\text{MTBF}$$

The greatest requirement for availability is for network elements (a circuit switch, for example), which are generally required to provide 99.999% availability or 0.001% unavailability [2]. Such a system is termed *highly available* (HA).

Reliability in Wireless Access in a Vo802.11 Network

A common perception is that a wireless network cannot be as reliable as a wired network because the airwaves cannot be as "solid" as a wire. In reality, a wireless form of access in the enterprise or "last mile" offers more forms of backup or redundancy than that found in wired networks.

Redundancy in Vo802.11 Networks

Manufacturers have been designing redundancy into their products for years in the form of redundant power supplies, multiple processors, segmented memory, and redundant disks. A Vo802.11 network can incorporate redundancy in the form of multiple channels to back up those channels that fail or become congested (Figure 11.2). Chapter 10 provides a description of these redundant channels in 802.11a and 802.11b.

HA is enhanced when each component is replicated in a system. This is called *redundancy.* If one unit fails, its replicated unit takes over. Redundant configurations are expressed by the notation *m:n,* where *m* represents the number of standby unit(s) and *n* represents the number of active unit(s) supported by the standby unit(s). A typical configuration is 1:1 where there is one active unit for every active unit or 1:6 where there is one standby unit for six active units. Usually, the smaller the *n*, the greater the protection and cost. Given the highly reliable nature of today's components, a carrier may determine that configurations greater than 1:1 provide sufficient availability. Class 4/5 switches are more likely to use a 1:1 redundancy model because the effect of a failure is more expensive. Moore's law, which states that computing power doubles while computing cost halves every 18 months, has the effect of making redundancy less expensive as time goes by [2].

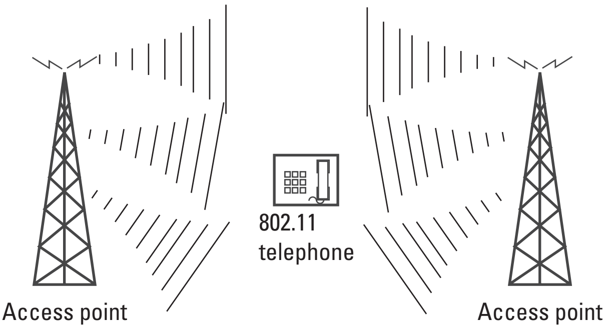


Figure 11.2 A Vo802.11 network requires planning for redundancy to avoid any single point of failure.

A Vo802.11 network can deploy redundant access points to cover for access points that fail. Network planners can also plan for overlapping cells of access point coverage. In this way, when one AP becomes inoperable, another AP whose cell covers that of the failed AP can cover those subscribers served by the failed AP. Given the declining price of access points, it is becoming increasingly cheaper to provide high levels of reliability by simply building in redundancy in a Vo802.11 network with redundant access points.

Repairability

Repairability is the relative ease with which service technicians can resolve or replace failing components. Two common metrics used to evaluate this trait are how long it takes to do the actual repair and how often the repair work needs to be repeated. In more sophisticated systems, this can be done from remote network operations centers, where failures are detected and circumvented and arrangements are made for permanent resolution with little or no involvement of operations personnel. The market now has a number of network monitoring tools for 802.11 networks that allow a network manager to quickly determine if and where network degradation is taking place and what components need to be put in standby. These tools also allow a network manager to determine where the voice quality in the wireless side of the network is being degraded and take steps to correct it.

Recoverability

Recoverability refers to the ability to overcome a momentary failure in such a way that there is no impact on end-user availability. It could be as small as a portion of main memory recovering from a single-bit memory error or as large as having a redundant AP switch over to its standby system with no loss of data or transactions. By hot swapping redundant APs to cover those that have failed, a Vo802.11 network planners can maintain a high degree of reliability in their networks.

Achieving the “Five 9s” with a Vo802.11 Softswitch

A call across a Vo802.11 network requires a switch to perform the call setup and teardown activities—no switch, no call. When PSTN engineers refer to “five 9s,” they are referring only to the switch and not to the network as a whole. Achieving “five 9s” is mostly a matter of engineering . If softswitch vendors can engineer out single points of failure and engineer in redundancy and other measures that Class 4/5 switch vendors have used for years to ensure reliability, then they, like the Class 4/5 switch vendors, can also advertise “five 9s” of reliability. Softswitch vendors can also engineer their platforms to be NEBS compliant. These measures allow softswitch to match the "five 9s" of reliability demonstrated by Class 4. A number of softswitch solutions have achieved "five 9s."

Building high availability into the switching component of a Vo802.11 network is simply a matter of good engineering. The main components of engineering a Class 4 or Class 5 switch for HA are redundancy, no SPOF, hot switchover, preservation of calls, in-service upgrades, component reliability, reproducible quality, and NEBS compliance. The same is true for a softswitched network.

In an HA system, two or more systems are loosely coupled to each other with the help of redundancy software. The reliability provided can be further classified as asymmetric or symmetric based on whether the systems act as active/standby (idle) or run in a parallel load sharing/balancing mode. An active/standby type of a system has further categories such as 1+1 redundancy or N+K redundancy based on number of active nodes and the number of standby nodes that are available. Cluster mode is another such HA architecture in which applications can run in either load-sharing or failover mode. The reliability of HA systems can be further enhanced by hardening some of the hardware components of the individual system constituting an HA system. Typical candidates for such treatment are network interface cards, disk controllers, disks, and power supply [2].

HA computing utilizes the redundant resources of clustered (two or more) processors (Figure 11.3). Such solutions address redundancy for all components of a system, processors, main and secondary memory, network interfaces, and power/cooling. Although hardware redundancy may be effectively addressed by clustered (redundant) hardware resources, the class of errors detected is less comprehensive and the time required to recover from errors is much longer than that of fault-tolerant machines. Still, fault recovery in tens of seconds, from most common equipment failures, can be achieved at less than half the cost of traditional fault-tolerant computer systems. HA systems are often configured as dual-redundant pairs [3].

Carriers require high system availability and are concerned with the effects of possible softswitch downtime. Carriers demand low MTBF and employ traffic overload control, the shedding of call processing capacity in the event of component failures, and quick failure detection and recovery mechanisms. The softswitch answer is to architect redundant softswitch hardware nodes at different locations throughout the network, which contributes to the overall network reliability.

In addition to COs posing a SPOF, the copper wires that go to the wiring pedestal in a residential neighborhood are also a SPOF. The wiring pedestal is a SPOF. The fiber-optic cable that runs from the wiring pedestal to the CO is also

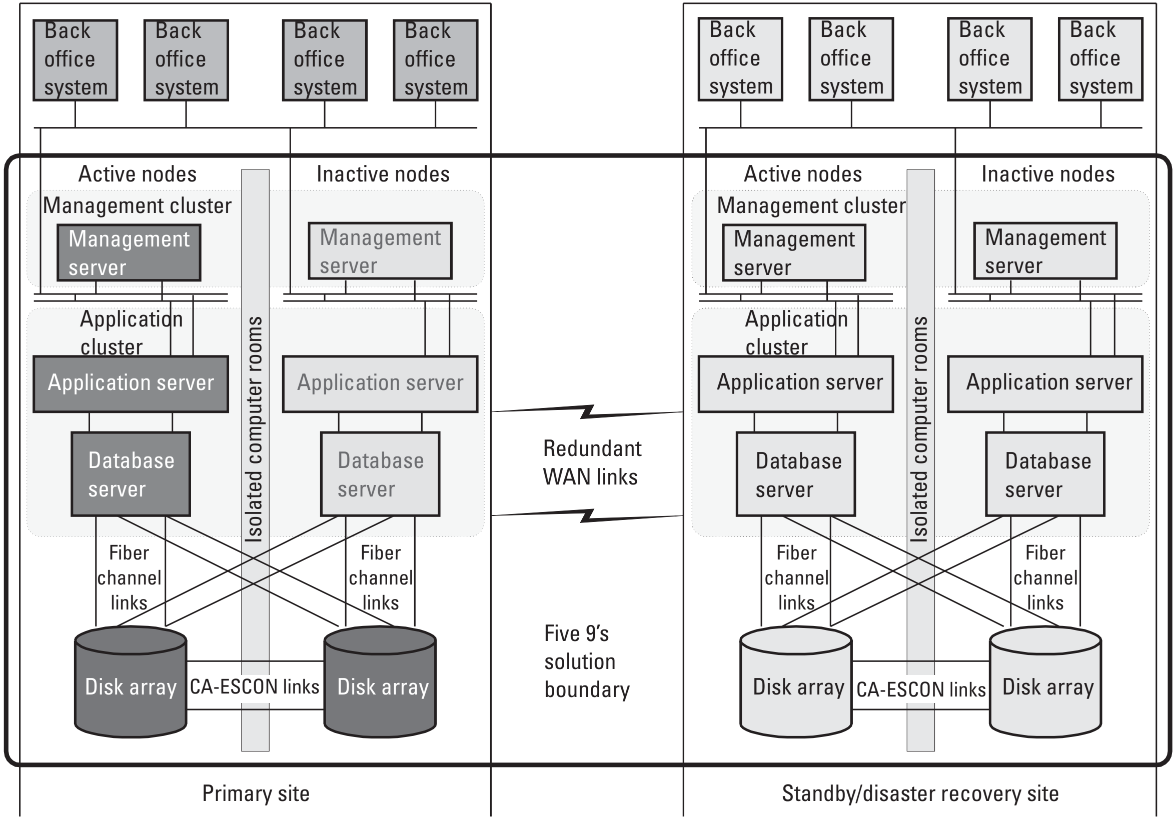


Figure 11.3 HA systems replicate all network elements resulting in no single point of failure. (*From:* [4]. © 2002 Hewlett-Packard Development Company, L.P. Reprinted with permission.)

a SPOF. If any of these items fails or is destroyed, the corresponding subscribers are without service [1].

NEBS

In addition to "five 9s," the other buzzword for reliability in the Class 4 market is the *Network Equipment Building Standards.* NEBS addresses the physical reliability of a switch. It is contained in Telcordia specification SR 3580, an extensive set of rigid performance, quality, safety, and environmental requirements applicable to network equipment installed in a carrier's CO. Nearly all major carriers in North America require that equipment in their COs or switching locations undergo rigorous NEBS testing. Tests include electrical safety, immunity from electromagnetic emissions, lightning and power faulting, and bonding and grounding evaluations. Equipment must also pass a series of physical standards including temperature, humidity, and altitude testing, fire resistance (usually by destructive burning), earthquake vibration resistance, and a battery of other rigid tests. As a final check of NEBS compliance, service providers also examine backup and disaster recovery strategies. Such strategies include: (1) ensuring access to mirror sites and fire and waterproof storage facilities for critical database and configuration backup information and (2) backing up electrical power using diesel-powered generators to prevent network outages in the event of power failures [5].

By using many of the mechanisms that Class 4/5 switches have utilized over the years (redundancy, fault tolerance, NEBS) to achieve the "five 9s" of reliability, softswitch is achieving the same levels of reliability. Given the declining costs of computing power, it is possible that softswitch may even exceed the "five 9s" of reliability while remaining economically competitive to a Class 5 solution. As a result, a wireless VoIP network can be engineered as an HA network to match or exceed the reliability or availability of the PSTN.

Distributed architecture can also improve the reliability of a softswitch solution. With distributed architecture there is no single point of failure on a network. Any redundant component on the IP network can pick up where the primary component failed. If a media gateway controller in Denver is destroyed in a force majeure, for instance, another media gateway controller can pick up where the Denver media gateway controller failed [5].

Power Availability

Power and environment also have the potential to impact the overall availability of a telephone network. Power is also unique in that it does not impact one device at a time as does software or hardware. Its affects, or can affect, an entire building or multiple buildings at a time. This can impact all devices in the network availability definition including distribution, core, gateway, and softswitch components all at once. The calculations, therefore, change from device-based calculations to entire network-based ones, which creates a significant impact to theoretical availability depending on the power protection strategy used for the network. Consider these statistics:

- The average number of outages sufficient to cause system malfunction per year at a typical site is approximately 15.
- Ninety percent of the outages are less than 5 minutes in duration.
- Ninety-nine percent of the outages are less than 1 hour in duration.
- Total cumulative outage duration is approximately 100 minutes per year.

Availability levels of “five 9s” or higher require a UPS system with a minimum of 1-hour battery life or a generator with an onsite service contract or 4-hour response for UPS system failures or problems. A recommended HA solution requires additional support to achieve “five 9s” overall. The HA softswitch solution must include UPS and generator backup for all distribution, core, gateway and softswitch devices. In addition, the organization should have UPS systems that have auto-restart capabilities and a service contract for 4-hour response to support the UPS/generator. Given the following recommended core infrastructure for softswitch HA, it is estimated that nonavailability (downtime due to power failure) will be 2 minutes per year:

- UPS system and generator backup;
- UPS systems with auto-restart capabilities;
- UPS system monitoring;
- Four-hour service response contract for UPS system problems;
- Maintenance of recommended equipment operating temperatures 24 hours per day, seven days per week.

Overall power availability using the above suggestions is estimated to be 99.99962%. This impacts overall availability, 99.99993%, in the same way a new module would affect a device in a serial system. The calculation used to determine overall estimated availability is then (0.9999962) × (0.9999993) or 99.99955% [6].

Conclusion

This chapter covered the argument regarding matching the “five 9s” of reliability currently enjoyed by PSTN CO switches. A Vo802.11 network that utilizes 802.11 as access and softswitch for switching can match or exceed the “five 9s” of reliability with engineering that parallels that of PSTN switches (for the PSTN, its only the switch and not the network as a whole that is “five 9s” compliant). That is, by architecting a Vo802.11 network to contain no single points of failure, being NEBS compliant while utilizing an HA network design, a wireless VoIP network can be just as—if not more—reliable than the PSTN.

References

[1] Ohrtman, F., *Softswitch: Architecture for VoIP,* New York: McGraw-Hill, 2002.

[2] Convergent Networks, “Understanding Carrier-Grade Reliability and Availability,” white paper, http://www.convergentnetworks.com.

[3] Kehret, W., “High Availability Requires Redundancy and Fault Tolerance,” *RTC Magazine,* Vol. VI, No. 2, February 1998, http://www.themis.com/new/pubs/ha_article.html.

[4] Cook, N., “The Cost of 99.999% Availability,” *Hewlett-Packard Company Presentation*, Colorado University, February 5, 2002.

[5] Eline, J. C., and M. Pyykkonen, *New Softswitch Technology: The Next Evolution for Today’s Public Telephone Network,* Towbin: C. E. Unterberg, 2001, pp. 27–28.

[6] Cisco, “IP Telephony: The Five Nines Story,” white paper, http://www.cisco.com/warp/public/cc/so/neso/vvda/iptl/5nine_wp.pdf.

12

Vo802.11 Features and Applications

Imagine installing a Vo802.11 application in your corporate WLAN. How does the caller get voice mail, call forwarding, conferencing, three-way calling, last call returned, and other calling features? Is the Vo802.11 network the equivalent of a walkie-talkie? Deploying voice-only over a Vo802.11 network is not commercially viable in an industrialized economy. If a Vo802.11 network is to replace an enterprise PBX or the local PSTN, it must offer features comparable to that of the network it replaces. One often-heard objection is the perception that any alternative to the PSTN or corporate PBX cannot perform E911 or CALEA. Vo802.11 products are now on the market that perform these functions. In addition, the means are available for wireless VoIP networks to provide all the features of the network it replaces. Enterprise users deploying wireless VoIP must also take into consideration which features their staff will need and how they will be delivered. Will the firm continue to use its existing TDM PBX or will it upgrade to an IP-PBX or outsource to an IP Centrex for switching and features?

One objection many service providers and subscribers alike have regarding Vo802.11 technologies is the perception that they do not duplicate the 3,500 features of a Class 5 switch or the 21 or so features available from a PBX or Centrex service. The architecture described in this chapter applies both to enterprise and carrier grade applications.

A softswitch in a Vo802.11 network that replaces a Class 5 switch or PBX utilizes application and media servers that replicate the features found on a Class 5 or PBX. That same softswitch also potentially offers features not found on a Class 5 switch or not even possible with a Class 5 switch. These new Vo802.11 features are usually written in text-based languages using open standards. It is possible that, given the flexibility in creating new features, some softswitch solutions that replace Class 5 switches may eventually offer *more* than 3,500 features. This chapter provides an overview of how those features are delivered over a softswitched Vo802.11 network.

Features in the Legacy PSTN

CLASS features are the basic services available in each LATA. Features and the services they enable are a function of Class 5 switches and SS7 networks. Features often allow service provider systems to offer high margins that result in stronger revenue streams.

Examples of features offered through a CLASS switch can be grouped under two major portfolios: basic services and enhanced services. Examples of basic services include 1+, 800/900 service, travel cards, account codes, PINs, operator access, speed dialing, hotline service, *automatic number identification* (ANI) screening, VPNs, calling cards, and call detail recording.

Enhanced services include information database services (NXX number service, authorization codes, calling card authorization, debit/prepaid card), routing and screening services (includes CIC routing, time-of-day screening, ANI screening, class-of-service Screening), enterprise networks, data and video services (dedicated access lines, ISDN PRI services, dialable wideband services, switched 56-Kbps service), and multiple dialing plans (full 10-digit routing, 7-digit VPN routing, 150-digit international dialing, speed dialing, and hotline dialing). Most of the features mentioned have been standard on CLASS switches for many years.

The long list of features above is evidence of the importance of features in the legacy market in which they were developed. Service providers will not give up these features and the margins they generate. In the converging market, features are equally important to reliability because service providers do not want to offer fewer features to their customers and they will want to continue to offer high-margin features.

A softswitch solution emphasizes open standards as opposed to the legacy Class 4/5 switches that historically offered a proprietary and closed environment. Softswitch vendors stress that their open standards are aimed at freeing service providers from vendor dependence and the long and expensive service development cycles of legacy switch manufacturers.

Features and Signaling

Features are a function of the Class 5 switch and the SS7 network. So how are the features of the PSTN transferred to a converging market where the softswitch competes with the Class 4/5 switches? First, SS7 information generated in the PSTN must be transported transparently across the IP network. The preceding chapter described SS7 over IP. It is possible that with mechanisms that make it easier for service providers to quickly roll out new, high-margin services that the voice market will shift in favor of the service provider that can deliver those features quickly.

Note that softswitch could allow a service provider to offer expanded features that are made available via the subscriber's PC or IP handset. The PC offers greater flexibility than a telephone handset in range of communications between the subscriber and the switch. For example, *11, or any other handset input is limited relative to all that can be offered on a Web page. In theory, if the 3,500 or so Class 5 features were cataloged on a Web site and made more obvious to and convenient for a subscriber to use, then the service provider could better capitalize on those features. To date, Class 4/5 switches offer only the telephone handset as a user interface to features.

The *service creation environment* (SCE) of softswitch is almost unlimited and has the potential to change much of the switch market in the converging market in the favor of softswitch. Softswitch has the capability to relay existing features of the Class 5 and SS7 network. In addition, it can offer a variety of new features potentially leading to a total count of features in excess of the 3,500 features of the Class 5 switch [1, p. 198].

SCE

The simple transport of voice (Vo802.11, VoIP, TDM) is highly commoditized and offers low margins to service providers. Without services, softswitches in a Vo802.11 network would not be able to generate the voice revenue that currently provides 80% of overall service provider revenue [2]. The key to high margins in the converging market is rapid service creation. The service providers that can differentiate themselves from the competition with unique services will win market share and profitable revenue margins from those services. In the legacy market, the offering of such services was dictated by the switch vendors (no more than two to three in the North American market). The switch vendors had little incentive to enable their customers to roll out new services in a rapid and flexible manner. Softswitch changes this scenario.

By virtue of its open standards interface, softswitch enables service providers to quickly create and roll out new services. The structure of the softswitch architecture makes it possible for service providers to integrate applications from the softswitch vendor or third-party vendors or even to develop their own in-house applications. The inclusion of the option to write custom APIs on a softswitch allows a service provider to write only those custom APIs that are applicable to a given service at a given location. In addition to interoperability between gateways and softswitches, the APIs are standardized to allow any third-party developer the ability to create applications on top of the softswitch.

Service creation allows the development of new service building blocks and the assembly of services from these building blocks, typically using one or more commercially available, off-the-shelf tools such as an *integrated development environment* (IDE). The concept here is *modularity,* in which a service provider can "mix-and-match" components in their network regardless of the vendors involved. If the voice mail platform of a Vo802.11 service provider is not working to the satisfaction of that service provider, it can be replaced by a competing voice mail platform. Vo802.11 network services can be assembled on the fly in a plug-and-play fashion, drastically reducing the time and effort required to develop services [3].

APIs

Features reside at the application layer in a softswitched Vo802.11 architecture. The interface between the call control layer and specific applications is the application program interface. Writing and interfacing an application with the rest of the softswitch architecture occurs in the SCE. Those open standards include APIs known as Parlay, JAIN, CORBA, and XML. Figure 12.1 details the relationship of APIs in the SCE.

APIs and Services

To compete with incumbent service providers, Vo802.11 service providers must be able to provide existing services while introducing advanced services quickly and inexpensively. By moving responsibility for advanced services out of call control entities, APIs resident on an application server, such as JAIN, Parlay, CORBA, CPL, CGI, and Servlets, provide Vo802.11 service providers with an environment where differentiating services can be deployed rapidly. The open,

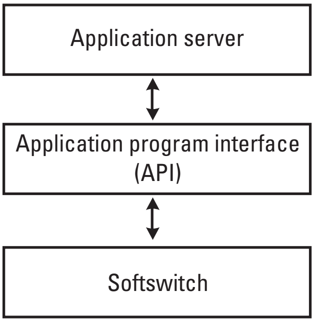


Figure 12.1 Relationship of APIs in the SCE.

as opposed to proprietary, nature of these APIs allow services to be created, managed, and deployed without requiring new or upgraded network infrastructure functionality.

One of the major advantages of using softswitched Vo802.11 is the fact that the underlying transport protocol (IP) is also used for a variety of other services, including Web access, e-mail, presence, and instant messaging. Enhanced services in Vo802.11 networks will enable integration of these services, bringing together telephony and Internet services. Application servers are required to interwork with a variety of nontelephony protocols and APIs in order to provide these services. Because of the complexity and overhead, service APIs reside on an application server to provide enhanced services processing.

The application server model provides an architecture for enhanced services within a Vo802.11 network and within the business domain of the network provider. APIs such as Parlay, JAIN, and CORBA focus on hybrid networks and provide for services both inside and outside of the network provider's business and technology domains.

XML

Extensible Markup Language (XML), the next generation of HTML, is now viewed as the standard way in which information will be exchanged in environments that do not share common platforms. XML 1.0 was released in February 1998. XML-based network management uses the World Wide Web Consortium's (W3C's) XML to encode communications data, providing an excellent mechanism for transmitting the complex data that are used to manage networking gear. Building an API around an XML-based *remote procedure call* (RPC) gives a simple, extensible way to exchange these data with a device. Receiving data in XML opens options for handling the data using standards-based tools. XML is widely accepted in other problem domains, and free and commercial tools are emerging at an accelerating rate [4].

SIP: Architecture for Enhanced Services in Softswitched Vo802.11 Networks

Two components in a softswitched Vo802.11 network, the application server and the media server, are introduced into the architecture to provide support for enhanced service logic, management functions, and specialized media resources (Figure 12.2). Application servers are intended to host a variety of enhanced services. An application server provides a framework for the execution and management of enhanced services. These services make use of the call control

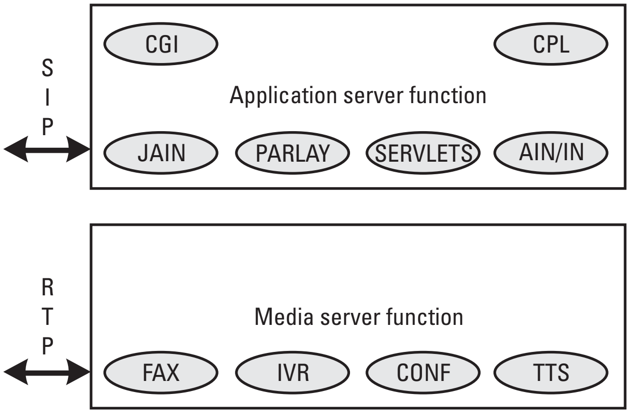


Figure 12.2 Functions of application and media servers. (*From:* [5]. © 2001 BroadSoft, Inc. Reprinted with permission.)

functions provided by the underlying infrastructure and also allow the integration of messaging, presence, and Web services.

Media Servers

A media server provides specialized resources such as *interactive voice response* (IVR), conferencing, and facsimile functions. Media servers and application servers are independent and can be deployed on separate physical platforms or on the same platform. An application server can utilize resources on a media server for enhanced services, which require access to the media stream.

Application Servers

APIs, as described earlier in this chapter, are hosted on the application server and provide access to underlying service and switching functions. Using these APIs, services can be easily developed and deployed. In addition, traditional telephony call models and TCAP/INAP protocols can potentially reside on an application server, providing access to standard AIN/IN telephony services.

Call control entities in a Vo802.11 network, such as gateways, softswitches, and IP phones, act as SIP user agents. Application servers may act as SIP entities: user agent, redirect server, proxy server, or third-party call controller (back-to-back user agents). All entities may communicate directly or through proxy servers. A register mechanism is required to allow an application server to inform a call control entity of its availability. Alternatively, a call control entity may be statically configured with the address information of an application server.

Architecture

A functional view of the Vo802.11 softswitch architecture is shown in Figure 12.3, and the function of each entity shown there is described in Table 12.1.

The introduction of an application server to a Vo802.11 network provides for platforms specifically designed for enhanced services. By using SIP as the application interface between a call control entity (softswitch) and an application server, a common, standard protocol can be used for communication between all call control and service execution entities. These new services can be deployed quickly with little or no impact on network rollouts.

An application server utilizes a SIP interface, which provides access to the signaling aspects of a call, while a media server utilizes an RTP interface, which provides access to the media stream. Services on application servers make use of the resources provided by media servers. With an application server and media server, services that require access to both the signaling path and media stream can be provided.

Interface Between Call Control and Application Server

SIP is used as the interface between call control entities (softswitch) and application servers because of its general acceptance, availability, and ability to set up, tear down, and manage sessions between endpoints. In this context, a call control entity has the ability to set up and take down a call signaling path to an application server, and for an application server to set up and take down a call signaling path to a call control entity. It also includes the ability to convey

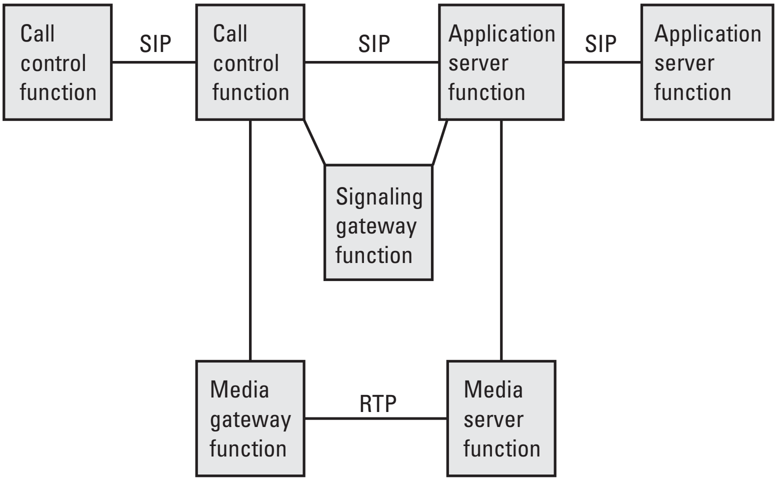


Figure 12.3 Softswitch enhanced service architecture. (*From:* [5]. © 2001 BroadSoft, Inc. Reprinted with permission.)

Table 12.1
Functions of Enhanced Services Architecture Entities

Function	Description
Call control function	May provide connection control, translations and routing, gateway management, call control, bandwidth management, signaling, provisioning, security, and call detail record generation.
Media gateway function	May provide conversion between circuit-switched resources (lines, trunks) and the packet network (IP, ATM), including voice compression, fax relay, echo cancellation, and digit detection.
Signaling gateway function	May provide conversion between the SS7 signaling network (SS7 links) and the packet network, including protocols such as ISUP and TCAP.
Application server function	May provide for the execution and management of enhanced services, handling the signaling interface to a call control function. It also provides APIs for creating and deploying services.
Media server function	May provide for specialized media resources (IVR, conferencing, facsimile, announcements, speech recognition), and handling the bearer interface to a media gateway function.

calling and called party information, hold and resume connections, transfer sessions, and establish multiple-party connections. SIP is used only for signaling; RTP is used to carry the media. SIP relays the information necessary to establish RTP communication between endpoints.

Application Server Interactions

In addition to using SIP as the interface between application servers and call control entities, it can also be used as the interface between application servers. This allows application servers to interact, providing the ability for two or more enhanced services on the same or different application servers to be linked. With the ability for an application server to delegate control to other application servers, a mechanism of managing feature interactions can be introduced. This becomes critical as more enhanced services are introduced into a network.

The introduction of applications into Vo802.11 networks provides a vehicle for the deployment of enhanced services. Utilizing a standard protocol between call control entities and application servers, enhanced services can be quickly introduced on application servers by application experts. Media servers can be used to provide specialized bearer resources, required by many enhanced services. Standardized service APIs and AIN/IN call models, which reside on an application server, can be used to allow service developers access to underlying telephony network functions. In addition, application servers can also utilize Internet protocols and APIs to provide truly converged services [6].

Vo802.11 Networks and E911 and CALEA Requirements

E911

One obstacle to wide deployment of Vo802.11 is accommodating the demands of *Enhanced 911* (E911) service. Although 911 services allow users to quickly request service from emergency personnel, E911 goes a step further by using ANI to relate location information to that number and determine which *public service answering point* (PSAP) should handle the call.

In an enterprise setting, it is not only important for E911 to direct emergency services to a particular building, but to indicate from where in the building the 911 call originated. Typically, when someone dials 911 from a desk, the ANI information that reaches the PSAP is based on the company's general, seven-digit phone number associated with the PBX, not the user's extension. In a life-and-death circumstance, that incorrect information can create harmful delays. As a result, some local and state governments make incorporation of E911 services a legal obligation. To that end, many private phone system vendors have incorporated ways for PBXs to fulfill these E911 requirements.

Different vendors offer different strategies toward accommodating the demands of E911. The most straightforward way is to relate circuit terminations to an extension number. In a traditional, PBX environment, each phone jack is associated with a particular extension. When an E911 call is placed from a particular location, the PBX sends a seven-digit number associated with that extension to a central office router, which can then use ANI to route the call to the correct PSAP. The PSAP can then furnish emergency personnel with that special number's location information.

However, in a Vo802.11 network, the circuit-switched world's E911 fix does not apply to SIP (or any other VoIP protocol). Because SIP phones and other VoIP devices become "floating" network appliances that can attach to the network from any point, their location is not static. Just like a laptop, a SIP phone can connect anywhere on a company's network—all it needs is a valid IP address. If the phone can be anywhere, then how can a company's phone system serve the E911 network with the necessary information?

Several fixes to this problem have arisen from within the SIP community:

- Install GPS chipsets in SIP end-user devices that can provide geographic location information.
- Triangulate the source of the call by Vo802.11 access points.
- Simply have users "log on" to their devices when moving them from one location to another.
- If a SIP deployment makes use of preexisting voice cabling to create a second, "VoIP-only" network, the IP PBX can relate circuits

terminations to *direct inward delivery* (DID) lines in much the same way as a circuit-switched PBX.

Whatever the solution, "fixes" to the E911 issue will have to be standardized across SIP products from all vendors, since the E911 space is regulated [7].

CALEA

What is CALEA? The *Communications Assistance for Law Enforcement Act* (CALEA) is a U.S. law providing for wiretapping (intercept) of information from a telecommunications network. CALEA dates from the 1990s and is not a traditional requirement of the PSTN. Another concept related to CALEA is lawful intercept, which happens when a judge has issued a court order allowing the wiretapping of a given telephone number (not a person). The court order is awarded to a law enforcement agency who in turn produces the court order to a telecommunications service provider. There may be multiple court orders by different law enforcement agencies for a given telephone number.

Wiretaps fall into two categories: call detail and call content. In a call detail wiretap, information regarding the various calls sent and received from that phone number are compiled. That is, the wiretap records what numbers were called or from what numbers calls were received, the date and time of each call, and the duration of each call. Call content information is recordings of the calls themselves revealing the actual content of the call. The suspect must not detect the wiretap. The tap must then occur within the network and not at the subscriber gateway or customer premise. The wiretap must not be detectable by any change in timing, feature availability, or operation.

In the PSTN, lawful intercept occurs in the Class 5 switch because it avoids any contact with the subscriber gateway or customer premise. The concern in the industry is that, in a VoIP network that does not have a Class 5 switch, there will be no easy mechanism for installing wiretaps.

It is possible to perform lawful intercept in a VoIP network. VoIP networks contain separate call agents and media gateways. The call agent is responsible for all call control and is the element that collects all details about the calls required in a call detail tap. Hence, a softswitch can provide a call detail wiretap solution. The call agent does not see the call content contained in the RTP media stream so call content must be collected elsewhere in the network.

CALEA applies to those service providers offering their services as primary line. CALEA does not apply to service providers offering their voice services as secondary line. This applies to service providers in the United States. Not all markets in the world have similar requirements. CALEA is a valid concern for vendors wishing to sell their Vo802.11 solutions to a U.S.-based RBOC (there are four) as a Class 5 replacement. Given the variety and flexibility of softswitch solutions, compliance with CALEA is neither impossible nor as clear-cut as it is in a TDM PSTN setting [1, pp. 220–221].

Vo802.11 Applications Made Possible by Softswitch Features

Web Provisioning

Web provisioning enables customers to do their own provisioning for their Vo802.11 service via a Web site. Customers can choose their own product mix and when individual services can be turned on or off. Other features available on softswitched Vo802.11 networks include voice mail to e-mail, voice e-mail browsing, voice calendar (in which the phone calls to alert you about upcoming events), voice-enabled dialing, very flexible call routing based on external events, click to dial on PCs (in which the phone rings and connects to party), remote call control (in which you can forward your phone including feature via a Web page), and Web phone control. In short, using VoiceXML programming a new feature can be written in 20 minutes to a few days [Nathan Stratton, CTO, Exario, telephone interview with author, October 28, 2001].

Voice-Activated Web Interface

Many cellular service providers already offer voice-activated dialing in which the subscriber's voice interfaces with a database of telephone numbers and selects one to be dialed. The same application can apply to Vo802.11 networks. Vo802.11 vendor Vocera offers this feature. There is no dial pad on Vocera Vo802.11 handsets. Rather, "dialing" is done by voice command, known as *voice dialing.*

That technology can be taken a few steps further to voice-activated Web interfacing. Instead of needing a computer to interface with Web sites or e-mail, a subscriber could access those resources via a voice-activated interface. In short, a subscriber could retrieve driving instructions via their cell phones or obtain stock quotes, the latest news, and weather reports, as well as their e-mail.

The Big "So What!?" of Enhanced Features in Vo802.11 Networks

At the time of this writing, it is still far too early to determine what the "killer app" will be with regard to applications for softswitch that cannot be duplicated on a Class 4/5 switch. Applications such as follow me, Web provisioning, and voice-activated Web pages may look very promising at this time, but it should be stressed that it is not the specific application that constitutes a "killer app" in context, but rather the infrastructure that makes any new and convenient application possible. Perhaps the best definition of what new services can be made possible by softswitch architecture and applications is provided by Ike Elliott, former chairman of the International Softswitch Consortium: "that feature or set of features that makes a worker more efficient." In other words, this architecture offers a business or industry the flexibility to write features and applications that are specific to those businesses or industries [personal communications, Ike Elliott, senior vice-president, Global Softswtich Services, Level 3 Communications, multiple interviews with author, May and June 2002].

Example of a Wireless Killer App: I-Mode

The arrival of Vo802.11 in the marketplace is not as simple as substituting Vo802.11 for TDM voice. It is about a greatly expanded service set that includes voice and a growing array of data services made possible by the added bandwidth available via 802.11. A good example of the possibilities of converged voice and data via wireless delivery is the Japanese cell phone service provider DoCoMo's *i-mode,* which provides HTML-based (text-based language delivered over a packet network) information to DoCoMo's cell phone-enabled subscribers. DoCoMo subscribers can access graphic information from their cell phones. Some 46,000 unregulated sites can be reached by typing in a URL, but the 1,800 official i-mode offerings are constantly monitored by DoCoMo to make sure they are current and easy to use. Official i-mode sites are allowed to charge ¥100 to ¥300 a month ($0.85 to $2.50), which DoCoMo collects for them in exchange for a 9% fee, and although these sites are created with a compact version of HTML, the lingua franca of the Web, terms like *HTML* and even *Web* never appear in i-mode ads.

I-mode, introduced with minimal expectations in February 1999, has attracted more than 25 million subscribers—one-fifth of Japan's population. New subscribers are still signing on at the rate of 43,000 a day, 1.3 million a month. I-mode now has 39 million cell phone subscribers and generated revenues last year of $39 billion with 180 employees. That translates to $216 million per employee per year. Compare this to Qwest and WorldCom at $295,000 and $229,000, respectively, per employee per year (before restatements for that year). Vo802.11 service providers could duplicate this success [8].

Conclusion

For Vo802.11 to replace the PSTN or a TDM enterprise telephony solution, it must be able to offer the same set of features available in the PSTN or the enterprise PBX. This is possible with softswitched architecture and programming languages such as VoiceXML. Given a flexibility relative to Class 4 and 5 switches, a Vo802.11 network can offer just as many of the necessary features that the PSTN switches offer. Given the ease of writing and deploying new features relative to legacy networks, it is possible that softswitch solutions potentially offer more than the 3,500 features available on a Class 5 switch. The ease of deployment of these new feature sets makes it easy for Vo802.11 service providers seeking to compete with PSTN service providers to roll out services competitive to the PSTN. It is also a component in lowering the barriers to entry for competitors to the PSTN.

References

[1] Ohrtman, F., *Softswitch: Architecture for VoIP,* New York: McGraw-Hill, 2002.

[2] Telica, "Accelerating the Deployment of Voice over IP (VoIP) and Voice over ATM (VoATM)," white paper, June 2001, posted by International Engineering Consortium at http://www.iec.org.

[3] SUN Microsystems, "The JAIN APIs: Integrated Network APIs for the JAVA Platform," white paper, May 2002, p. 7, http://www.java.sun.com/products/jain.

[4] Shafer, P., "XML-Based Network Management," Juniper Networks white paper, August 2001.

[5] Hoffpauir, S., "Softswitch & Third Party APIs," *Softswitch Expo,* San Diego, CA, December 5, 2001, http://www.broadsoft.com/pdf/softswitch%20Expo%202001.pdf.

[6] Hoffpauir, S., "Enhanced Services Framework," International Softswitch Consortium, white paper.

[7] Mitel, "SIP at the Desktop Intelligence to the Edge," white paper, September 2001, p. 7.

[8] Rose, R., "Pocket Monster," *Wired Magazine,* September 2001, http://www.wired.com/wired/archive/9.09/ docomo.html.

13

Regulatory Considerations for Vo802.11 Networks

An objection often raised about Vo802.11 applications is that because the spectrum used by 802.11 (2.4 GHz for 802.11b and 5.8 GHz for 802.11a) is unlicensed, it will inevitably become overused (called "tragedy of the commons") to the point of being unusable at which time the government (U.S. government or other) will step in to control the spectrum making it "not free," thus costing the service provider his or her profit margin and relegating the market to the formerly deep-pocketed monopolists.

This chapter first explores the considerations wireless service providers should take into account when deploying Vo802.11 service on unlicensed 802.11 bands. Next, the chapter covers regulatory concerns for voice over IP, regardless of whether it is in the wired or wireless environment. Liberalizing spectrum policy will inevitably encourage the use of Vo802.11 as a means of bypassing incumbent telephone service providers.

Current Regulatory Environment for 802.11

The Vo802.11 specification is VoIP over 802.11. Even though 802.11 operates in unlicensed spectrum, service providers must know a number of things in order to stay out trouble with state and federal authorities. The following sections outlines the most prominent problem areas.

Spectrum is managed by a number of different organizations. The most visible to the general public is the Federal Communications Commission. The FCC manages civilian, state, and local government usage of the radio spectrum. The FCC regulations are contained in the *Code of Federal Regulations, Title 47.*

At the time of this writing, the FCC has very limited resources for enforcement, because the trend for the last couple of decades has been toward deregulation and the reduction of staffing in the enforcement bureaus. There is also the *National Telecommunications and Information Administration* (NTIA), which works with the *Interdepartmental Radio Advisory Committee* (IRAC) to manage federal use of the spectrum.

We provide a brief overview of what a service provider needs to be concerned about when operating in an unlicensed spectrum. This synopsis was provided by Tim Pozar of the Bay Area Wireless Users Group and is based on many years of experience advising friends and clients on what they can and cannot do with unlicensed spectrum.

Power Limits

Ideally, a well-engineered path will have just the amount of power required to get from point A to point B with good reliability. Good engineering will limit the signal to only the area being served. This has the effect of reducing interference and providing a more efficient use of the spectrum. Using too much power will cover more area than is needed and has the potential to interfere with other users of the band. Because 802.11 is designed for short range use such as offices and homes, it is limited to very low power.

802.11b and Its Relationship to FCC Part 15, Section 247

Point-to-Multipoint Links

The 802.11 service providers are allowed up to 30 dBm or 1W of *transmitter power output* (TPO) with a 6-dBi antenna or 36 dBm or 4W of EIRP. The TPO needs to be reduced 1 dB for every decibel of antenna gain over 6 dBi.

Point-to-Point Links

The FCC encourages directional antennas to minimize interference to other users. The FCC in fact is more lenient with point-to-point links than point-to-multipoint links by requiring only the TPO to be reduced by one-third of a decibel instead of the full decibel required for point-to-multipoint links. More specifically, for every 3 dB of antenna gain over a 6-dBi antenna, a WISP must reduce the TPO 1 dB below 1W. For example, a 24-dBi antenna is 18 dB over a 6-dBi antenna. This requires lowering a 1W (30-dBm) transmitter 18/3 or 6 dB to 24 dBm or 0.25W.

802.11a and Its Relationship to FCC Part 15, Section 407

Point-to-Multipoint Links

As described earlier, the U-NII band is chopped into three sections. The “low” band runs from 5.15 to 5.25 GHz and has a maximum power of 50 mW (TPO). This band is meant to be used within buildings only as defined by the FCC’s Rules and Regulations Part 15.407(d) and (e):

> (d) Any U-NII device that operates in the 5.15-5.25 GHz band shall use a transmitting antenna that is an integral part of the device.
> (e) Within the 5.15-5.25 GHz band, U-NII devices will be restricted to indoor operations to reduce any potential for harmful interference to co-channel MSS operations.

The “middle” band runs from 5.25 to 5.35 GHz, with a maximum power limit of 250 mW. Finally, the “high” band runs from 5.725 to 5.825 GHz, with a maximum transmitter power of 1W and antenna gain of 6 dBi or 36 dBm or 4W EIRP.

Point-to-Point Links

As with 802.11b, the FCC does give some latitude to point-to-point links in 15.407(a)(3). For the 5.725- to 5.825-GHz band, the FCC allows a TPO of 1W and up to a 23-dBi gain antenna without reducing the TPO 1 dB for every 1 dB of gain over 23 dBi.

Part 15.247(b)(3)(ii) does allow the use of any gain antenna for point-to-point operations without having to reduce the TPO for the 5.725- to 5.825-GHz band.

Interference

Interference is typically the state of the signal you are interested in while it is being destructively overpowered by a signal in which you are not interested. The FCC has a specific definition of “harmful interference”:

> Part 15.3(m) **Harmful interference**.
> Any emission, radiation or induction that endangers the functioning of a radio navigation service or of other safety services or seriously degrades, obstructs or repeatedly interrupts a radiocommunications service operating in accordance with this chapter.

Because there are other users of the band, interference will be a factor in 802.11 deployments. The 2.4-GHz band is a bit more congested than the

Table 13.1
Spectrum Allocation for 802.11 and Cousers

Part/Use	Starting Gigahertz	Ending Gigahertz
Part 87	0.4700	10.5000
Part 97	2.3900	2.4500
Part 15	2.4000	2.4830
Fusion lighting	2.4000	2.4835
Part 18	2.4000	2.5000
Part 80	2.4000	9.6000
ISM–802.11b	2.4010	2.4730
Part 74	2.4500	2.4835
Part 101	2.4500	2.5000
Part 90	2.4500	2.8350
Part 25	5.0910	5.2500
U-NII–low	5.1500	5.2500
U-NII–middle	5.2500	5.3500
Part 97	5.6500	5.9250
U-NII–high	5.7250	5.8500
ISM	5.7250	5.8500
Part 18	5.7250	5.8750

Source: [1].

5.8-GHz band, but both have their cousers (Table 13.1). The following sections describe the other users of this spectrum and what interference mitigation may be possible for each.

Devices That Fall into Part 15

The 2,400- to 2,483-MHz Range

Table 13.2 lists the 802.11b spectrum bands. The 2,400- to 2,483-MHz range includes unlicensed telecommunications devices such as cordless phones, home spy cameras, and FHSS and DSSS LAN transceivers. Operators have no priority over or parity with any of these users, and any device that falls into Part 15 must not cause harmful interference to any of the licensed and legally operating Part 15 users and must accept interference from all licensed and all legally operating Part 15 users. Table 13.2 lists the 802.11b spectrum bands. This is stated in Part 15.5 (b) and (c).

Table 13.2
Spectrum Bands for 802.11b

Channel	Bottom (GHz)	Center (GHz)	Top (GHz)
1	2.491	2.412	2.423
2	2.406	2.417	2.428
3	2.411	2.422	2.433
4	2.416	2.427	2.438
5	2.421	2.432	2.443
6	2.426	2.437	2.448
7	2.431	2.442	2.453
8	2.436	2.447	2.458
9	2.441	2.452	2.463
10	2.446	2.457	2.468
11	2.451	2.462	2.473

Source: [1].

> (b) Operation of an intentional, unintentional, or incidental radiator is subject to the conditions that no harmful interference is caused and that interference must be accepted that may be caused by the operation of an authorized radio station, by another intentional or unintentional radiator, by industrial, scientific and medical (ISM) equipment, or by an incidental radiator [or basically everything].
>
> (c) The operator of a radio frequency device shall be required to cease operating the device upon notification by a Commission representative that the device is causing harmful interference. Operation shall not resume until the condition causing the harmful interference has been corrected.

Operators of other licensed and unlicensed devices can inform you of interference and require that you terminate operation. It doesn't have to be a "commission representative." Using 802.11b you can interfere even if you are on different channels, because the channels are 22 MHz wide and are only spaced 5 MHz apart. Channels 1, 6, and 11 are the only channels that don't interfere with each other (see Table 13.2).

Devices That Fall into the U-NII Band

Unlike the 2.4-GHz band, this band does not have overlapping channels. For the lower U-NII band, there are eight 20-MHz-wide channels. You can use any of the channels without interfering with other radios on other channels that are within "earshot." Ideally, it would be good to know what other Part 15 users are out there. Looking into groups under the banner of “Freenetworks” is a good place to start.

ISM—Part 18

This is an unlicensed service. Typical ISM applications are the production of physical, biological, or chemical effects such as heating, ionization of gases, mechanical vibrations, hair removal, and acceleration of charged particles. Users are ultrasonic devices such as jewelry cleaners and ultrasonic humidifiers, microwave ovens, medical devices such as diathermy equipment and magnetic resonance imaging equipment (MRI), and industrial uses such as paint dryers (Part 18.107). RF should be contained within the devices but other users must accept interference from these devices. Part 18 frequencies that could affect 802.11 devices are 2.400 to 2.500 GHz and 5.725 to 5.875 GHz. Because Part 18 devices are unlicensed and operators are likely to be clueless about their impact, it will be difficult to coordinate with them. Fusion lighting is covered by Part 18.

Satellite Communications—Part 25

This part of the FCC’s rules is used for the uplink or downlink of data, video, and so on to or from satellites in Earth orbit. One band that overlaps the U-NII band is reserved for Earth-to-space communications at 5.091 to 5.25 GHz. Within this spectrum, 5.091 to 5.150 GHz is also allocated to the fixed-satellite service (Earth-to-space) for nongeostationary satellites on a primary basis. The FCC is trying to decommission this band for “feeder” use to satellites as “after 01 January 2010, the fixed-satellite service will become secondary to the aeronautical radionavigation service” (see Part 97 below). A note in Part 2.106 [S5.446] also allocates 5.150 through 5.216 GHz for a similar use, except it is for space-to-Earth communications. There is a high chance of interfering with these installations, because Earth stations are dealing with very low signal levels from distance satellites.

Broadcast Auxiliary—Part 74

Normally the traffic is *electronic news gathering* (ENG) video links going back to studios or television transmitters. These remote vehicles such as helicopters and trucks need to be licensed. Only Part 74 eligibles such as TV stations and networks can hold these licenses (Part 74.600). Typically these transmitters are scattered all around an area, because TV remote trucks can go anywhere. This can cause interference to 802.11 gear such as access points deployed with omnidirectional antennas servicing an area. Also the “receiving” points for ENG are often mountaintops and towers. Depending on how 802.11 transmitters are deployed at these same locations, they could cause interference to these links. Wireless providers should consider contacting a local frequency coordinator for Part 74 frequencies that would be affected. At this point, there have been reports of FHSS devices interfering with these transmissions because the dwell time for FHSS tends to punch holes in the video links. DHSS is less likely to cause interference to ENG users, but their links can cause problems with your 802.11 deployment. ENG frequencies that overlap 802.11 devices are 2.450 to 2.467 GHz (channel A08) and 2.467 to 2.4835 GHz (channel A09) (Part 74.602).

Land Mobile Radio Services—Part 90

For subpart C of this part, users can be anyone engaged in a commercial activity. They can use from 2.450 to 2.835 GHz, but can only license 2.450 to 2.483 GHz. Users in subpart B would be local government. This would include organizations such as law enforcement and fire departments. Some uses may be video downlinks for flying platforms such as helicopters, also known as terrestrial surveillance. Depending on the commercial or government agency, coordination goes through different groups such as the Association of Public Safety Communications Officials. Consider going to their conferences. You can also try to network with engineering companies to which the government outsources for their frequency coordination.

Amateur Radio—Part 97

Amateur radio frequencies that overlap 802.11b are 2.390 to 2.450 GHz and 5.650 to 5.925 GHz for 802.11a. They are primary from 2.402 to 2.417 GHz and secondary at 2.400 to 2.402 GHz. There is a *Notice of Proposed Rule Making* (NPRM) in with the FCC to change the 2.400 to 2.402 to primary. Amateurs are very protective about their spectrum.

Federal Usage (NTIA/IRAC)

The federal government uses this band for “radiolocation” or “radionavigation.” There are several warnings in the FCC’s Rules and Regulations that disclose this fact. In the case of 802.11b, a note in the Rules warns:

> 15.247(h) Spread spectrum systems are sharing these bands on a noninterference basis with systems supporting critical Government requirements that have been allocated the usage of these bands, secondary only to ISM equipment operated under the provisions of Part 18 of this chapter. Many of these Government systems are airborne radiolocation systems that emit a high EIRP which can cause interference to other users.

In the case of 802.11a, the FCC has a note in Part 15.407 stating that:

> Commission strongly recommends that parties employing U-NII devices to provide critical communications services should determine if there are any nearby Government radar systems that could affect their operation.

Laws on Antennas and Towers

FCC Preemption of Local Law

The installation of antennas may run counter to local ordinances and homeowner agreements that would prevent installations. Thanks to the Satellite Broadcasting and Communications Association, which lobbied the FCC, the FCC has stepped in and overruled these ordinances and agreements.

This rule should only apply to broadcast signals such as TV, DBS, or MMDS. It could be argued that the provision for MMDS could cover wireless data deployment.

Height Limitations

Local Ordinances

Most if not all cities regulate the construction of towers. There will be maximum height zoning of the antenna/tower (residential or commercial) and construction and aesthetic (e.g., what color, how hidden) regulations.

FAA and the FCC Tower Registration

The FAA is very concerned about things that airplanes might bump into. Part 17.7(a) of the FCC R&R describes "Any construction or alteration of more than 60.96 meters (200 feet) in height above ground level at its site" [1].

Regulatory Issues Concerning VoIP

Vo802.11 is VoIP over 802.11. Therefore, regulations that apply to VoIP can apply to Vo802.11. In its April 10, 1998, Report to Congress, the FCC determined that "phone-to-phone" IP telephony is an enhanced service and not a telecommunications service. The important distinction here is that telecommunications service providers are liable for access charges to local service providers both at the originating and terminating end of a long-distance call. A telecommunications service provider must also pay into the Universal Service Fund. Long-distance providers using VoIP (and, by inference, Vo802.11) avoid paying access and universal service fees. Given thin margins on domestic long distance, this poses a significant advantage for phone-to-phone IP telephony service providers [2].

The possibility that the FCC may rule differently in the future cannot be discounted. Having to pay access fees to local carriers to originate and terminate a call coupled with having to pay into the Universal Service Fund would pose a significant financial risk to the business plan of a softswitch-equipped Vo802.11 service provider. Just as international long-distance bypass providers used VoIP to bypass international accounting rates and make themselves more competitive than circuit-switched carriers, Vo802.11 carriers can make themselves more competitive in the domestic market by bypassing access charges and avoiding paying into the Universal Service Fund. The service provision model set forth below is strongly affected by the possibility of the FCC reversing itself on phone-to-phone IP telephony.

Access fees in North American markets run from about 1 cent per minute for origination and termination fees to upwards of 5 cents per minute in some rural areas. That is, a call originating in Chicago, for example, would generate an origination fee of 1 cent per minute. If the call terminated in Plentywood, Montana, it may generate a 5-cent-per-minute termination fee. This call would generate a total of 6 cents per minute in access fees. If the carrier can only charge 10 cents per minute, it will reap only 4 cents per minute for this call after paying access fees to the generating and terminating local phone service providers. Table 13.3 illustrates the impact on profit and loss for a long-distance service provider that must pay access fees. It is possible that the FCC at some point could reverse this ruling and make VOIP carriers pay access fees

Conclusion

This chapter outlined the current regulatory regime for Vo802.11 operators. The chapter attacks the objection that there is too little spectrum available for a mass market deployment of Vo802.11 and that government interference in this area will only lead to a stultifying regulatory regime that could kill Vo802.11 as a promising last mile solution to telephone and cable TV companies. Recent studies and pronouncements by the FCC and members of the U.S. Senate indicate support for reforming the spectrum policy in promoting the deployment of 802.11 and its related technologies [3, 4].

Regulation of VoIP does not loom ominously on the horizon. Note also that it is difficult to regulate the flow of packets across an IP background,

Table 13.3
Impact of Access Fees on Long-Distance Revenues

Retail Price (Cents Per Minute)	Origination Fee	Termination Fee	Revenue After Access Fees	Access Fees as a Percentage of Retail Price
5	1	1	3	40
5	1	5	–1	120
5	5	5	–5	200
10	1	1	8	20

particularly because those packets flow unconnected across state and national borders. Another question arises as to how and at what expense the flow of those voice packets could be regulated.

References

[1] Pozar, T., "Regulations Affecting 802.11 Deployment," Bay Area Wireless Users Group white paper, pp. 2–7, 10–11, http://www.lns.com/papers/part15.

[2] Federal Communications Commission Report to Congress, April 10, 1998, paragraphs 88–93, http:// www.fcc.gov.

[3] Long, J., "Senators Boxer and Allen to Introduce Broadband Legislation," *Phone+,* November 22, 2002, http://www.phoneplusmag.com/hotnews/2bh2214134.html.

[4] Powell, M., "Broadband Migration—New Directions in Wireless Policy," *Silicon Flatirons Conference,* University of Colorado, Boulder, CO, October 30, 2002.

14

Economics of Vo802.11 Networks

Why deploy a Vo802.11 network? Most Vo802.11 technology is currently focused on the enterprise market. This chapter looks at economic aspects of Vo802.11 in the enterprise. The enterprise market for Vo802.11 is driven by two factors: (1) It works and provides a number of mechanisms that make its users more efficient as evidenced by numerous case studies and (2) it saves the enterprise money on telecommunications.

Vo802.11 Works: Case Studies

Vo802.11 found its niche markets early in certain vertical markets where the conveniences of mobility were obvious to network planners. These markets include medical, education, financial services, and manufacturing and warehousing, as discussed in the following sections.

Medical

Sint-Annendael Hospital

Sint-Annendael Hospital of Diest, Belgium, deployed a Vo802.11 telephone system from SpectraLink of Boulder, Colorado. The psychiatric hospital has equipped 70 of its doctors, nurses, and staff with the SpectraLink NetLink telephones, allowing them to use the handsets throughout critical medical departments. As a result, health care professionals in the 168-bed hospital are more responsive to the needs of patients, their families, and other staff throughout the entire facility. Sint-Annendael selected Vo802.11 telephones because the system was compatible with their newly purchased TDM PBX and their network of WLAN access points.

The Vo802.11 telephones bring newfound mobility to Sint-Annendael's medical staff, who previously were limited to using beepers and returning pages on wired telephones at nurses' stations, which was time consuming, lacked privacy, and caused response-time delays. The wireless telephones integrate with the best selling PBX systems and support proprietary IP protocols as well as H.323 for enterprise Vo802.11 solutions [1].

Mercy Medical Center, Roseburg, Oregon

Mercy Medical Center in Roseburg, Oregon, could have given nurses regular phones or new pagers, but opted for wireless VoIP devices instead. The hospital uses Vo802.11 communications badge appliances from Vocera Communications.

Nurses wear the gadgets, about the size of TV remotes, around their necks with lanyards or pinned to their shirts. To reach someone, a nurse presses a button on the badge and scrolls through names in the system, and then presses another button to talk. The voice signal travels to the recipient over the hospital's Vo802.11 infrastructure.

Mercy Medical installed 10 Cisco 802.11b access points throughout the facility to support the Vo802.11 infrastructure, which replaces an outdated pager system. The Vo802.11 lets nurses contact each other faster and more efficiently than previous pager systems or with telephones.

A Windows server running Vo802.11 management software and user database controls the Vo802.11 network on the back end. The software lets administrators add and remove users from the system and customize individual calling features. They can track users on the system using an open source database. The hospital uses IVR software to process voice-activated commands.

During emergency situations when nurses do not have time to scroll through names, they can use the IVR feature by voice prompt. Speaking a person's name or the name of a group or for all nurses on a certain team will initiate the call. Users can also determine where someone is through the Vo802.11 system. They say "find" and the name of the person, and the IVR software responds with the location of the requested user. To make this possible, Mercy Medical assigned all of the Wi-Fi access points in the hospital a name based on their location, such as "Emergency," "OR," or "Cafeteria" and entered them into the Vocera database. The hospital is even attaching Vocera badges to frequently used pieces of equipment, such as EKG machines or defibrillators so nurses can find these devices quickly.

The Vo802.11 system is slightly more expensive than the pager system it replaced, but the hospital managers expect to save money ultimately because it is giving Vo802.11 access to other groups, such as doctors, maintenance workers, and cleaning staff. These employees had used pager systems or walkie-talkies [2].

University of Southern California University Hospital

At University of Southern California University Hospital (USCUH) in Los Angeles, nurses and doctors also use Vo802.11 phones. A total of 273 wireless IP handsets are in use at the hospital. Wireless IP phones are now a single source of communication for all staff and replace a mix of communication methods used in the past such as nurse call buttons, a public address paging system, and cordless telephones, which were inefficient.

The Vo802.11 decision came after two separate infrastructure projects USCUH undertook last year. In the first, the hospital built an 802.11b network to support mobile devices, such as laptops. In the second project, the hospital installed an IP PBX to connect some remote facilities to a legacy PBX over IP. USCUH brought those projects together when it chose to give the medical staff Vo802.11 phones. To ensure voice quality, the hospital relies on the Vo802.11 server, which provides a proprietary QoS feature for giving voice calls priority over data. When the voice packets hit the wired network, they are placed into the first of eight priority queues on LAN switches. The Vo802.11 users can place external calls, too. The Vo802.11 traffic converts to regular voice as it moves from the IP PBX to an ISDN line, connecting that IP PBX to the legacy TDM PBX. From there, the call can be passed on to the PSTN.

Education

Richardson (Texas) Independent School District

Teachers need to be able to communicate with the outside world, but they are often hampered by the unusual demands of their workplace and the limited technology solutions available to them. Due to their range and accessibility limitations, traditional telephones are unable to respond adequately to the physical and logistical challenges faced by teachers every day. Constantly on the go and often working out of different classrooms throughout the day, teachers are frequently unable to get to a central area to make calls or check messages via standard means of communications. In addition, traditional telephones are also ineffective in the classroom environment. With limited coverage and disruptive ringing, traditional wired phone systems are proving to be inadequate. For example, wired phones do not provide a method for tracking down a teacher who is between classes, on another area of the school grounds, or in a different classroom than usual.

The Richardson Independent School District (Richardson ISD) in Richardson, Texas, has solved this communications problem with SpectraLink's Link Wireless Telephone System (Link WTS). The Link WTS allows teachers and campus administrators to be in touch and available wherever they are on school grounds. Richardson uses approximately 2,100 of the lightweight, durable handsets, providing one for every teacher in each of its 54 schools.

Vo802.11 as Part of Complete Wireless Coverage

Richardson ISD had previously used another wireless communications solution, but it had outlived its useful life. The system did not provide 100% coverage of all school properties. The Link WTS provides Richardson ISD with 100% coverage, including the playing fields and the parking lots.

A common misconception is that only classrooms need to be accessible during the school day. However, according to a 1999 Harris teachers' poll, 89% of school accidents take place outside of the classroom. Thus, adequate communications coverage of all of the school grounds is critical. Teachers need to be able to summon a school nurse or resource officer from anywhere a crisis may occur. They may also need to call for emergency personnel outside the school. With Vo802.11, calls can be made and received from the basement, stairwells, or hallways, as well as outside the school campus. Without complete coverage, a teacher might have to send a student runner to deliver a message if an incident occurred in a "dead zone" or, worse yet, be forced to leave an injured student's side in order to summon help.

The Link WTS integrates with each school's existing PBX, saving the school the cost of an extensive telecommunications upgrade. The Link WTS consists of a network of base stations, about the size of smoke detectors, that are strategically located throughout the campus to relay calls between the Vo802.11 telephones and the PBXs. Each base station handles multiple simultaneous telephone calls, which are handed off from one unit to another as a user walks throughout the facility and school grounds. Because all calls are routed through the existing PBXs, no airtime charges are incurred when using the Link WTS wireless telephones.

Anglia Polytechnic University

The Michael A. Ashcroft Building, which houses Anglia Polytechnic University's (APU's) Ashcroft International Business School, boasts a 95% wireless environment with 45 SpectraLink NetLink wireless telephones providing Vo802.11 within the building.

Professors in the building carry NetLink Vo802.11 telephones while attending classes, meeting with students, and conducting research in the four-story building, which is dotted with 26 Proxim Orinoco 802.11b access points. The university has plotted 20 additional access points throughout the university campus. The installation of wireless voice communications followed that of the university's data communications, but wireless voice was always part of what architects envisioned in their new high-tech building. Today, staff and students are able to use their laptops wirelessly throughout the building and staff members carry their NetLink Vo802.11 telephones with them everywhere, never having to worry about missing an important call while they are away from their offices. The university has also made several NetLink wireless telephones available to students in the building's conference center. SpectraLink was selected for this project largely because of its interoperability with the university's existing telecom system and Proxim WLAN.

Financial Services

Bear Stearns

Enterprises today rely on real-time information to successfully run their businesses—nowhere more so than in the financial industry. Constant accessibility is critical in order to stay on top of perpetually changing stock prices, currency fluctuations, breaking news, and the latest rumors. Being out of touch can mean missing an important piece of information, but more importantly risks loss of considerable money and client relationships. Due to these critical day-to-day factors, financial professionals are always looking to improve their ability to stay in touch, no matter where they are in the office or what they are doing during the workday.

Bear Stearns, a leading investment banking, securities, and brokerage firm, has taken a major step toward keeping its people accessible throughout the workday. Bear Stearns has installed SpectraLink's Link WTS throughout its new world headquarters in New York City, allowing its employees to communicate over Vo802.11 telephones anywhere in the 45-story building. Now employees can stay in touch on their Vo802.11 telephones, whether at their desks, in meetings, or between floors, ensuring that the lines of communication are always open. Bear Stearns eliminated fruitless rounds of telephone tag, enabling their business to run at greater speeds and with more efficiency than ever before.

In contrast to cellular phones, the Vo802.11 telephone connects directly into Bear Stearns' existing PBX through a series of base stations that are strategically placed throughout the facility. This means that the handsets do not accrue expensive airtime charges, and that every area in the facility is covered. Whether an employee is in a hallway, stairwell, or basement, she or he is fully capable of making and receiving calls on a Vo802.11 telephone.

The Link WTS is tied to Bear Stearns' PBX, enabling employees to maintain the same phone extension and features on their wireless handsets that they have on their desk telephones. Call transferring, speed dial, hold, and extension dialing are all available on the lightweight, durable handsets that are easily carried in a pocket or clipped onto a belt.

Every floor in the building is equipped for the Link WTS, even the four large floors designated as trading floors where various types of securities, futures, and other financial instruments are exchanged. There are more than 300 base stations—each roughly the size of a smoke detector—installed throughout the building to ensure optimal coverage.

The countless multimillion-dollar transactions that financial firms conduct on a daily basis require the utmost attention to detail, focus, and responsibility. A missed opportunity can immediately damage a client's investment portfolio and the firm's reputation. By ensuring that staff is always accessible, financial firms not only reduce the risks inherent in their business, but also increase the overall client experience, which creates longer term relations and greater profitability for everyone involved. By giving their staff and management Vo802.11 telephones, Bear Stearns is working to improve the quality of service for their customers and the quality of work for their employees.

Manufacturing and Warehousing

Kaindl Flooring

Kaindl Flooring GmbH, a leading manufacturer of laminated flooring products, deployed SpectraLink Vo802.11 telephones in its primary warehouse in Salzburg, Austria. The deployment was added to Kaindl Flooring's existing WLAN, which is used for inventory management and scanning, in order to give users in the warehouse the ability to communicate with one another.

Kaindl Flooring worked through IBM Global Services to purchase and install the new system. The Vo802.11 telephones are integrated with a Cisco CallManager IP telephony application, which is tied to the company's existing Alcatel PBX. Calls are relayed over an 802.11b wireless LAN through a network of Cisco Aironet access points installed throughout the 150m by 100m facility. The wireless LAN gives the ability to combine voice and data traffic on a converged wireless infrastructure. This capability will save the flooring manufacturer expenses related to telecommunications resources while making employees more efficient through increased mobility.

Using the NetLink Vo802.11 telephones to leverage its existing wireless LAN to bring voice service into its warehouse, Kaindl was able to save the expense of upgrading their PBX. Kaindl employees report the voice quality is exceptional and that the handsets' rugged design is perfect for their demanding environment [1].

WISPs

AmberWaves: Vo802.11 as Differentiator

AmberWaves is a WISP in northwest Iowa. One of their clients has three offices with 35 employees linked by a point-to-point 802.11 network. The greatest distance between the three offices is 19 miles. Calls between the three offices are long distance, which would retail at 10 cents per minute in this rural community if they did not use Vo802.11. Data circuits are not inexpensive here either.

AmberWaves equipped its customer with 802.11 wireless point-to-point bridges to connect the LANs in the three offices. By adding a VoIP gateway at each office, the company is able to route its interoffice phone traffic on to the 802.11 network between the three offices. This means that all traffic that previously went over the PSTN was diverted to their internal 802.11 network, saving them money by eliminating local phone bills (no need for a large number of lines to serve the 35 employees) and by eliminating their interoffice long-distance expenses.

This wireless network allows the firm to be its own internal data and voice service provider. The use of Vo802.11 frees the firm from local and long-distance telephone bills (Figure 14.1). The end users report that the QoS on the network is better than the frame relay circuit they previously used [personal communications with Brent Bierstedt, CTO of AmberWave Communications, November 20, 2002]. By offering this unique service to its customers, AmberWaves differentiates itself from its competitors and avoids turnover ("churn") among its customers.

Vo802.11 Telephone System Cost Justification in the Workplace

In justifying the deployment of Vo802.11, it is important for enterprise planners to evaluate scenarios in which the use of Vo802.11 will save the company money. Once the questions of "how" Vo802.11 can save the company money are answered, the next question is "how much" money can the deployment of Vo802.11 save the company, that is, what is the return on investment on a

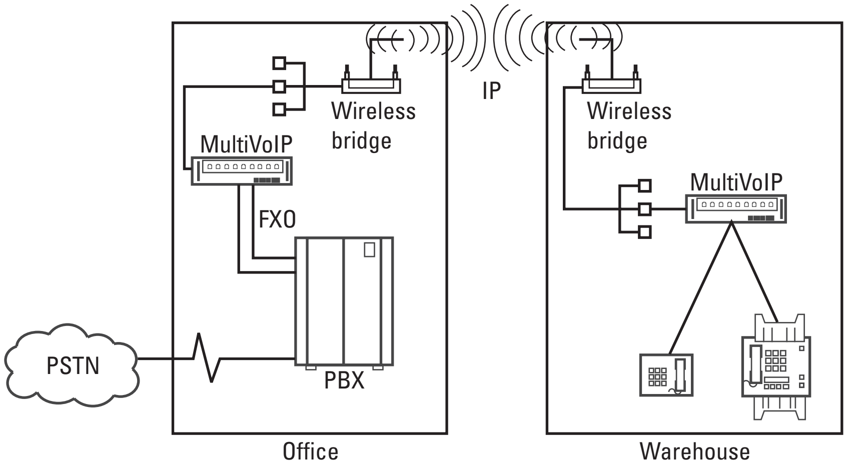


Figure 14.1 Using Vo802.11 in interoffice telephony saves on both local and long-distance phone bills. (*From:* [3]. © 2002 Anixter International. Reprinted with permission.)

Vo802.11 deployment in the enterprise? Some means of measuring "how much" money is saved is to analyze savings in platform costs, *moves, adds, changes* (MAC), supervisor time savings, efficiency in production line maintenance, long-distance savings, and interoffice telephony.

The economics of Vo802.11 in enterprise applications should be assessed in two ways: (1) by comparing applications where the wireless network is simply less expensive to deploy than the wired network where both applications perform the same function, and (2) by examining situations where a wireless network enables employees to be more productive. Money saved is money earned.

Platform Costs

Perhaps the most obvious comparison of Vo802.11 versus VoIP (on a WLAN or WAN) versus a legacy TDM voice infrastructure is to simply compare the costs of each of the networks. Given that a number of different platforms represent each of these technologies with a wide range of pricing, Figure 14.2 gives a broad comparison of the three competing technologies currently available to enterprise telephony planners. Table 14.1 compares the costs of installation and operation of a Vo802.11 networks versus legacy TDM and cell phones. The table indicates that Vo802.11 is the least expensive option both in terms of acquisition and operation. Tables 14.2 and 14.3 offer more comparisons.

MAC

TDM telephone networks in an enterprise are expensive to maintain. The estimated cost for a move, add, or change order in the network is upwards of $150

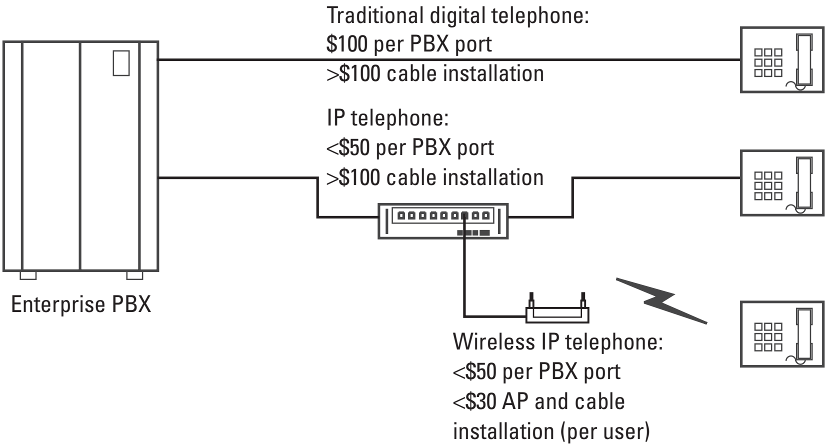


Figure 14.2 Cost comparisons of enterprise telephone infrastructure types: TDM, VoIP, and Vo802.11. (*From:* [1]. SpectraLink Corporation. Reprinted with permission.)

Table 14.1
Cost Comparisons for an Enterprise Telephone Infrastructure:
Cell Phone Versus TDM Desk Phone Versus Vo802.11

Cost Component	Cell Phone	TDM Desk Phone	Vo802.11 Phone
Handset	$100	$350	$400
PBX port	$0	$100	$50
Wireless network	$0	$0	$30
Installation	$0	$100	$0
Air time	$50/month	$0	$0
Total	$1,800 (3 years of service)	$550	$480

Source: [1].

for a computer on the network. When circuit-switched telephony is included, the cost can climb to $500 per seat. Because a WLAN recognizes almost any device at almost any location on a network, the costs associated with MAC orders are largely eliminated.

Saving Time and Money in Health Care

Table 14.4 gives an industry estimate of the time saved per year when nurses can take advantage of Vo802.11 networks in their hospitals. The table demonstrates how the convenience of Vo802.11 can save much valuable nursing time in a

Table 14.2
Cost Comparisons of Cell Phones Versus Vo802.11 Phones

Location	Charges for Cell Phone	Charges for Vo802.11 Phone
Office	Per-minute charge (monthly plan)	No charge
Hot spot (coffee shop/airport)	Per-minute charge (monthly plan)	No charge
Customer site with 802.11	Per-minute charge (monthly plan)	No charge
Home with 802.11	Per-minute charge (monthly plan)	No charge
Car (driving)	Per-minute charge (monthly plan)	No coverage

Source: [1].

Table 14.3
Cost Comparisons of Office Land Lines Versus Vo802.11

Function	Corporate Office Land-Line Costs	Vo802.11 Costs
Local service	$50/month business line/T1	No charge
Interoffice long distance	Per-minute charge	No charge

Table 14.4
Time Savings in Nursing Field with a Vo802.11 Telephone System

Task	Nurse Time Saved Per Nursing Unit
Nurse travel time to answer phone	58 minutes/day
Nurse waiting time at nurse station	88 minutes/day
TOTAL nursing time recovered	888 hours/year
Clerical time to locate nurses	90 minutes/day
TOTAL clerical time recovered	548 hours/year
Hold time for incoming calls	116 minutes/day
Recovered hold time for callers	706 hours/year

Note: Multiply time saved by hourly rate for each nurse to determine return on investment.

hospital. If a nurse costs the hospital $20 per hour and the hospital can save 888 hours per year by issuing the nurse a Vo802.11 device, the hospital has saved $17,760 in 1 year. If the cost per nurse for a Vo802.11 system is $400 (industry norm) and the nurse costs the hospital $20/hour, the device has paid for itself in 20 hours (two and a half 8-hour shifts) [personal communications with Brent Bierstedt, CTO of AmberWave Communications, November 20, 2002].

Supervisor Time Savings

Within large facilities, it takes supervisors a significant amount of time to get to an available wall or desk phone. For example, 20 seconds lost walking both to and from a wall phone is not a lot of time, but when it happens 20 to 30 times per day (not unusual), it adds up. Companies have conducted time and motion studies like those shown in Table 14.5.

Table 14.5
Time and Money Saved with a Vo802.11 Telephone Systems

Task	Time and Money Saved
20 seconds to and from wall phone	40 seconds saved/call
25 calls per day × 40 seconds saved/call	1,000 seconds saved/day
1,000 seconds saved/day/3,600 seconds/hour	278 hours/day/supervisor
0.278 hours/day/supervisor × 8 supervisors	2.22 total hours saved/day
2.22 hours saved/day × $15 salary/hour	$33.30 saved/day
$ 33.30 saved/day × 300 workdays/year	$9,990 is saved per year just by supervisors not having to find an available wall phone.

Source: [1].

Efficiencies in Maintenance of the Production Line

Maintenance personnel are some of the most enthusiastic users of Vo802.11 telephones. The Vo802.11 telephones allow them to be notified and respond to production line malfunctions immediately. They also allow the maintenance supervisor to repair machinery using both hands while receiving instructions on the Vo802.11 telephone directly from the machinery manufacturer's technical specialists. This eliminates errors and wasted time spent walking to and from a wall phone after each repair instruction. The continuous operation of production lines is vital to companies. Most companies know what the value of production is per hour. For example, one manufacturer told us that if the production line goes down, the company loses $1,000 worth of production per hour. To attach specific dollar savings, supervisors asked their maintenance engineers to highlight specific instances in which the Vo802.11 telephones were utilized and then to estimate the number of minutes saved. Their valuation of this benefit is shown in Table 14.6.

Table 14.6
Savings in Production Costs with a Vo802.11 Telephone System

Tasks	Time and Money Saved
Minutes saved per repair = 15 minutes	0.25 hours
Frequency of repairs where wireless telephones made a significant difference	1 time/week
Value of production/hour	$1,000/hour
Dollar value of efficiencies = 25 hours × 1/week × $1,000/hour	$250/week × 52 weeks/year = $ 13,000 savings per year

Source: [1].

Cost Savings with Regard to Long-Distance Customers

When customers or suppliers are statewide or national in scope, their long-distance telephone bills are significant. By immediately answering incoming calls from customers and suppliers, companies save the cost of returning those long distance calls later, as shown in Table 14.7.

Interoffice Telephony

It is useful to apply Figure 14.2 from earlier in the chapter to Table 14.8 to realize how this company has saved money on their telecommunications expenses by bypassing local telephone service providers.

Enterprise Conclusion

There are many ways to save money in an enterprise environment using Vo802.11. While many economists might focus on the hard and fast comparison of wired versus wireless, the real savings would initially appear less tangible and making an ROI or net present value analysis a little more challenging. When employee mobility and efficiency are taken into account, the savings begin to quickly add up. Where Vo802.11 can be employed, an even more tangible financial case for deploying 802.11 can be recognized when comparing the costs of wired versus unwired phones on the corporate LAN or WAN.

Some new technologies allow a voice-enabled PDA to be dual-channel 802.11 and CDMA/GSM. This capability allows an employee to talk over the 802.11 WLAN at the office or 802.11-serviced home or home office when in those offices. Using an 802.11 network allows the user to avoid per-minute charges on their cell phone. Once they leave the office they can switch over to their cell phone service provider should they have to make or receive calls.

Table 14.7
Annual Savings on Long-Distance Calls Made with a Vo802.11 Telephone System

Component	**Savings**
Average dollar cost per long-distance call	$2.50
Number of long-distance calls per day per supervisor not made due to the use of Vo802.11 telephones	2
Number of supervisors	6
Number of workdays per year	300
Dollar savings = $2.50 × 2 × 6 × 300	$ 9,000 per year

Source: [1].

Table 14.8
Potential Savings as a Result of Bypassing Local and Long-Distance Telephone Service Providers

Number of Phone Lines (T1, DS3) Per Office That Could Be Replaced with Vo802.11	Cost Per Month of Phone Lines That Could Be Replaced by Interoffice Vo802.11	Total Savings Per Office Per Month
Office A		
Office B		
Office C		

Lower Barrier to Entry

A Vo802.11 solution is considerably less expensive than a traditional network both in terms of acquisition and operation. This presents a lower barrier to entry and exit for a competitive service provider. A lower barrier to entry and exit allows alternative service providers to enter the market. Some types of service providers that could be encouraged to offer voice services in competition with incumbent telephone service providers (local and long distance) include ISPs, cable TV companies, electric utility companies, *application service providers* (ASPs), municipalities, and wireless service providers.

Considerations in Bypassing the PSTN with Vo802.11

This chapter has focused on cost comparisons for enterprise applications for Vo802.11 systems. At the time of this writing, there are only a few installations where Vo802.11 bypasses or substitutes for the PSTN and cost data are not well defined. However, the major components of the PSTN and cell phone infrastructure can be compared to Vo802.11. Table 14.9 compares those components from a high overview.

One of the major cost components of a TDM PSTN infrastructure is the Class 5 switch. A basic model costs about $10 million with an approximately $1 million per year maintenance contract. New softswitches on the market that can compete favorably with a Class 5 switch start at $150,000 with maintenance contracts in the neighborhood of $15,000 per year. Softswitches scale upward far more economically than Class 5 switches. Most importantly, unlike Class 5 switches, softswitches are geographically independent of their subscribers. The operating costs of a softswitch are a fraction of that of the Class 5 switch. Legacy cell phone infrastructure is based on Class 5 switches.

Access is a component of a Vo802.11 bypass in which Vo802.11 clearly has advantages over both the PSTN and cell phone infrastructure. One of the many reasons the Telecommunications Act of 1996 failed to bring true

Table 14.9
Cost Comparisons to Deploy a Cell Phone, PSTN, or Vo802.11 Infrastructure

Component	Cellular	PSTN	Vo802.11
Switching	Class 4/5: $10 million each	Class 4/5: $10 million each	Softswitch: $150,000 each
Access	Licensed spectrum (expensive depending on market)	Copper wire (ROW difficult/impossible to obtain)	Unlicensed spectrum (free)
Transport	TDM/ATM (expensive per-mile costs)	TDM/ATM (expensive per-mile costs)	IP (inexpensive; charged by bandwidth)
Data rates	128 Kbps (2.5G)	56-Kbps dial-up; 256-Kbps DSL; 1.54 Mbps at $1,000/month	54 Mbps (for 802.11g; given link budget will be less)

competition to the local loop in the North American market is that the cost of duplicating the copper wire "last mile" was prohibitively expensive if not impossible given ROW issues to reach subscribers. With regard to cell phone service providers, the purchase of wireless spectrum at FCC auctions can also be prohibitively expensive depending on the market to be served. Vo802.11 requires no ROW, no copper or coaxial cable, and no spectrum license.

Transport is another component where Vo802.11 enjoys strong advantages over PSTN or cell phone service. Much of the PSTN and most cell phone service providers use ATM for transport. This is charged in an equation that calculates both mileage and bandwidth. A Vo802.11 service uses IP for transport. This is charged only by bandwidth at considerable savings over legacy ATM infrastructures.

Finally, subscriber demands are no longer limited to voice only. Service providers must have a robust data delivery offering for their subscribers in order to stay competitive. The PSTN can offer data at the low end at 56 Kbps of bandwidth with dial-up access. DSL is a little more money per month and delivers about 256 Kbps, but most service providers cannot guarantee a steady data rate. Only 10% of PSTN subscribers have DSL in the U.S. market. Cell phone service providers range from 14 Kbps to as high as 128 Kbps for 2.5G services. An 802.11 service on which Vo802.11 is delivered, per the 802.11g standard, can be as high as 54 Mbps (although the link budget it dictate a slower rate).

Conclusion

Vo802.11 works. It has proven itself in the marketplace as evidenced by the numerous success stories in multiple vertical markets. The Vo802.11 market is validated by the presence of telecommunications vendor industry giants Cisco, Motorola, NEC, and Avaya. Its popularity in certain vertical markets will prove to be a testing ground for other vertical markets, which will inevitably lead to a universal acceptance of the technology based on less tangible selling points such as mobility, efficiency, and convenience.

Once proven in the enterprise market, Vo802.11 will migrate to the service provider industry where it will bypass the PSTN. PSTN service providers who ignore its incredible economic and technical advantages will be severely disrupted by it.

References

[1] SpectraLink case study. Available at http://spectralink.com/solutions/education.html.

[2] Vocera case study. Available at http://vocera.com/products/whitepapers.shtm.

[3] Strange, S., Anixter International, "Wireless: The Next Generation Cabling?" *BiCsi Breakfast Club Meeting,* June 26, 2002, http://www.bicsi.org/Content/Files/Presentations/Europe-breakfast10/Steve%20Strange.ppt.

15

Conclusion: Vo802.11 Is the Future of Voice Communications

The purpose of this book is to overcome objections to using 802.11 as a means of transmitting voice over the Internet Protocol. First, this book sets forth the thesis that Vo802.11 can replace the elements of the PSTN: access, switching, and transport. By utilizing 802.11 as an access medium, the copper wires of the PSTN can be bypassed (but not replaced in the short term). By using VoIP and a softswitch, the Class 4 and Class 5 switches of the PSTN can be bypassed. Underutilized IP backbones replace the long-distance networks of the PSTN to further bypass the PSTN infrastructure. Assuming the PSTN can be physically bypassed, the debate then shifts to dealing with the objections related to 802.11 and VoIP. By meticulously addressing each of the objections, it becomes increasingly clear that Vo802.11 is a viable technology.

Potential for a New Regulatory Regime

FCC New Spectrum Policy

The American spectrum management regime is approximately 90 years old. In the opinion of FCC Chairman Michael Powell, it needs a hard look and a new direction. Historically, four core assumptions have guided spectrum policy: (1) Unregulated radio interference will lead to chaos, (2) spectrum is scarce, (3) government command and control of the scarce spectrum resource is the only way chaos can be avoided, and (4) the public interest centers on government choosing the highest and best use of the spectrum.

Problem Areas in Spectrum Management and Their Solutions

Interference—The Problem

From 1927 through to today, interference protection has always been at the core of federal regulators' spectrum mission. The Radio Act of 1927 empowered the Federal Radio FCC to address interference concerns. While interference protection remains essential to its mission, interference rules that are too strict limit users' ability to offer new services; rules that are too lax may harm existing services. I believe the FCC should continuously examine whether there are market or technological solutions that can—in the long run—replace or supplement pure regulatory solutions to interference.

The FCC's current interference rules were typically developed based on the expected nature of a single service's technical characteristics in a given band. The rules for most services include limits on power and emissions from transmitters. Each time the old service needs to evolve with the demands of its users, the licensee has to come back to the commission for relief from the original rules. This process stymies innovation.

Due to the complexity of interference issues and the RF environment, interference protection solutions may be largely technology driven. Interference is not solely "caused" by transmitters, which many seem to assume—and on which our regulations are almost exclusively based. Instead, interference is often more a product of receivers; that is receivers are too dumb or too sensitive or too cheap to filter out unwanted signals. Yet, the FCC's decades-old rules have generally ignored receivers. Emerging communications technologies are becoming more tolerant of interference through sensory and adaptive capabilities in receivers. That is, receivers can "sense" what type of noise or interference or other signals are operating on a given channel and then "adapt" so that they transmit on a clear channel that allows them to be heard.

Both the complexity of the interference task—and the remarkable ability of technology (rather than regulation) to respond to it—are most clearly demonstrated by the recent success of unlicensed operations. According to the Consumer Electronics Association, a complex variety of unlicensed devices is already in common use, including garage and car door openers, baby monitors, family radios, wireless headphones, and millions of wireless Internet access devices using Wi-Fi technologies. Yet despite the sheer volume of devices and their disparate uses, manufacturers have developed technology that allows receivers to sift through the noise to find the desired signal.

Interference and Interference Protection

The recommendation of the Interference Protection Working Group of the FCC's Spectrum Policy Task Force was that the FCC should consider use of the *interference temperature* metric as a means of quantifying and managing interference: As introduced in this report, *interference temperature* is a measure of the RF power available at a receiving antenna to be delivered to a receiver—power generated by other emitters and noise sources. More specifically, it is the temperature equivalent of the RF power available at a receiving antenna per unit bandwidth, measured in units of degrees Kelvin. As conceptualized by the working group, the terms *interference temperature* and *antenna temperature* are synonymous. The term *interference temperature* is more descriptive for interference management.

Interference temperature can be calculated as the power received by an antenna (watts) divided by the associated RF bandwidth (hertz) and a term known as Boltzmann's constant (equal to 1.3807 W sec/K). Alternatively, it can be calculated as the power flux density available at a receiving antenna (watts per meter squared), multiplied by the effective capture area of the antenna (meter squared), with this quantity divided by the associated RF bandwidth (hertz) and Boltzmann's constant. An *interference temperature density* could also be defined as the interference temperature per unit area, expressed in units of degrees Kelvin per meter squared and calculated as the interference temperature divided by the effective capture area of the receiving antenna—determined by the antenna gain and the received frequency. Interference temperature density could be measured for particular frequencies using a reference antenna with known gain. Thereafter, it could be treated as a signal propagation variable independent of receiving antenna characteristics.

As illustrated in Figure 15.1, interference temperature measurements could be taken at receiver locations throughout the service areas of protected communications systems, thus estimating the real-time conditions of the RF environment.

Like other representations of radio signals, instantaneous values of interference temperature would vary with time and, thus, would need to be treated statistically. The working group envisions that interference "thermometers" could continuously monitor particular frequency bands, measure and record interference temperature values, and compute appropriate aggregate value(s). These real-time values could govern the operation of nearby RF emitters. Measurement devices could be designed with the option to include or exclude the on-channel energy contributions of particular signals with known characteristics such as the emissions of users in geographic areas and bands where spectrum is assigned to licensees for exclusive use.

The FCC could use the interference temperature metric to set maximum acceptable levels of interference, thus establishing a "worst case" environment in which a receiver would operate. Interference temperature thresholds could thus be used, where appropriate, to define interference protection rights.

The time has come to consider an entirely new paradigm for interference protection. A more forward-looking approach requires that there be a clear

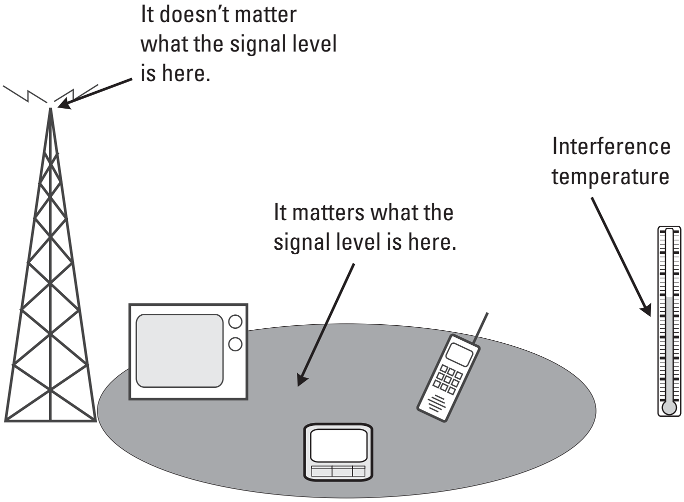


Figure 15.1 Interference temperature illustrated. (*From:* [2]. © 2002 FCC. Reprinted with permission.)

quantitative application of what is acceptable interference for both license holders and the devices that can cause interference. Transmitters would be required to *ensure* that the interference level—or interference temperature—is not exceeded. Receivers would be required to *tolerate* an interference level.

Rather than simply saying your transmitter cannot exceed a certain power, the industry instead would utilize receiver standards and new technologies to ensure that communication occurs without interference and that the spectrum resource is fully utilized. So, for example, perhaps services in rural areas could utilize higher power levels because the adjacent bands are less congested, therefore decreasing the need for interference protection [1].

From a simplistic and physical standpoint, any transmission facility requires a transmitter, a medium for transmission, and a receiver. Focus on receiver characteristics has not been high in past spectrum use concerns, hence, a shift in focus is in order. The FCC believes that receiver reception factors, including sensitivity, selectivity, and interference tolerance, need to play a prominent role in spectrum policy [2].

Spectrum Scarcity—The Problem

Much of the FCC's spectrum policy was driven by the assumption that there is never enough for those who want it. Under this view, spectrum is so scarce that government rather than market forces must determine who gets to use the spectrum and for what. The spectrum scarcity argument shaped the Supreme Court's *Red Lion* decision, which gave the FCC broad discretion to regulate broadcast media on the premise that spectrum is a unique and scarce resource. Indeed most assumptions that underlie the current spectrum model derive from traditional radio broadcasting and are oblivious to wireless broadband Internet applications.

The FCC has recently conducted a series of tests to assess actual spectrum congestion in certain locales. These tests, which were conducted by the FCC's Enforcement Bureau in cooperation with the task force, measured use of the spectrum at five major U.S. cities. The results showed that while some bands were heavily used, others either were not used or were used only part of the time. It appeared that these "holes" in bandwidth or time could be used to provide significant increases in communication capacity, without impacting current users, through use of new technologies. These results call into question the traditional assumptions about congestion. Indeed it appears that most of spectrum is not in use most of the time.

Today's digital migration means that more and more data can be transmitted in less and less bandwidth. Not only is less bandwidth used, but innovative technologies like software-defined radio and adaptive transmitters can bring additional spectrum into the pool of spectrum available for use.

Spectrum Scarcity—The Solution

In analyzing the current use of spectrum, the task force took a unique approach, looking for the first time at the *entire* spectrum, not just one band at a time. This review prompted a major insight: There is a substantial amount of "white space" out there that is not being used by anybody. The ramifications of the insight are significant. It suggests that while spectrum *scarcity* is a problem in some bands some of the time, the larger problem is spectrum *access*—how to get to and use those many areas of the spectrum that are either underutilized or not used at all.

One way the FCC can take advantage of this white space is by facilitating access in the time dimension. Since the beginning of spectrum policy, the government has "parceled" this resource in frequency and in space. The FCC historically permitted use in a particular band over a particular geographic region often with an expectation of perpetual use. The FCC should also look at *time* as an additional dimension for spectrum policy. How well could society use this resource if FCC policies fostered access in frequency, space, and time?

Technology has facilitated (and now it is hoped FCC policy will also facilitate) access to spectrum in the time dimension that will lead to more efficient use of the spectrum resource. For example, a software-defined radio may allow licensees to dynamically "rent" certain spectrum bands when they are not in use by other licensees. Perhaps a mobile wireless service provider with software-defined phones will lease a local business's channels during the hours the business is closed. Similarly sensory and adaptive devices may be able to "find" spectrum open space and utilize it until the licensees need those rights for their own use. In a commercial context, secondary markets can provide a mechanism for licensees to create and provide opportunities for new services in distinct slices of time. By adding another meaningful dimension, spectrum policy can move closer to facilitating consistent availability of spectrum and further diminish the scarcity rationale for intrusive government action.

Government Spectrum Policy—The Problem

The theory back in the 1930s was that only government could be trusted to manage this scarce resource and ensure that no one got too much of it. Unfortunately, spectrum policy is still predominantly a "command and control" process that requires government officials—instead of spectrum users—to determine the best use for spectrum and make value judgments about proposed—and often overhyped—uses and technologies. It is an entirely reactive and too easily politicized process.

In the last 20 years, two alternative models to command and control have developed, and both have flexibility at their core. First, we have the "exclusive use" or quasi-property rights model, which provides exclusive, licensed rights to flexible-use frequencies, subject only to limitations on harmful interference. These rights are freely transferable. Second, the "commons" or "open access" model allows users to share frequencies on an unlicensed basis, with usage rights that are governed by technical standards but with no right to protection from interference. The FCC has employed both models with significant success. Licensees in mobile wireless services have enjoyed quasi-property right interests in their licensees and transformed the communications landscape as a result. In contrast, the unlicensed bands employ a commons model and have enjoyed tremendous success as hotbeds of innovation.

Government Spectrum Policy—The Solution

Historically, the FCC often limited flexibility via command and control regulatory restrictions on which services licensees could provide and who could provide them. Any spectrum users who wanted to change the power of their transmitter, the nature of their service, or the size of an antenna had to come to the FCC to ask for permission, wait the corresponding period of time, and only then, if relief was granted, modify the service. Today's marketplace demands that the FCC provide license holders with greater flexibility to respond to consumer wants, market realities, and national needs without first having to ask for the FCC's permission. License holders should be granted the maximum flexibility to use—or allow others to use—the spectrum, within technical constraints, to provide any services demanded by the public. With this flexibility, service providers can be expected to move spectrum quickly to its highest and best use.

Public Interest—The Problem

The fourth and final element of traditional spectrum policy is the “public interest” standard. The phrase (or something similar), “public interest, convenience or necessity” was a part of the Radio Act of 1927 and likely came from other “utility” regulation statutes. The standard was largely a response to the interference and scarcity concerns that were created in the absence of such a discretionary standard in the 1912 act. The “public interest, convenience and necessity” became a standard by which to judge between competing applicants for a scarce resource—and a tool for ensuring interference did not occur. The public interest under the command and control model often decided which companies or government entities would have access to the spectrum resource. At that time, spectrum was not largely a consumer resource—but rather was accessed by a relatively select few. However, Congress wisely did not create a static public interest standard for spectrum allocation and management.

Serving the Public Interest in Spectrum Policy—The Solution

The FCC should develop policies that avoid interference rules that are barriers to entry, that assume a particular proponent’s business model or technology, and that take the place of marketplace or technical solutions. Such a policy must embody what we have seen benefit the public in every other area of consumer goods and services—choice through competition, and limited, but necessary, government intervention into the marketplace to protect such interests as access to people with disabilities, public health, safety, and welfare [1].

Current State of the Industry

Vo802.11 is achieving a surprising degree of adoption in enterprise markets. The entry into this market by data networking and telecommunications giants such as Cisco, Avaya, and Motorola offers powerful validation of this technology. The reality for Vo802.11 is that it is, at the time of this writing, an enterprise application. According to a recent Cahners’ In-Stat report, additional demand from verticals such as education, health care, retail, and logistics will help the overall voice-over-wireless LAN market expand to more than 80,000 handset shipments in 2002, a significant jump from the 20,000 shipments in 2001. Furthermore, Instat/MDR reports that annual shipments of Vo802.11x handsets are expected to pass half a million units by 2006 [3].

Just like PCs, Web access, and e-mail, Vo802.11 will grow out of the enterprise market and into the residential market. Innovative service providers can overcome a number of shortcomings in the legacy PSTN infrastructure to deliver Vo802.11 in addition to wireless broadband Internet services. In fact, the most likely market driver for Vo802.11 in residential markets is wireless broadband Internet in markets not served by DSL or cable modem. WISPs can add voice as a revenue stream in addition to their broadband Internet offering.

For these markets, 802.11 is described as being a "DSL killer" in that it presents a faster ROI for a WISP than the ROI a telephone company could expect when installing an expensive DSLAM (the device necessary to deliver DSL). Given an unfavorable ROI on the DSLAM, DSL is not available in most rural markets.

Projections: Futurecasting for Vo802.11

Figure 15.2 illustrates the convergence and progression of VoIP and 802.11 technologies. Given that they are "cheaper, simpler, smaller, and more convenient to use," these technologies will gradually replace the legacy network infrastructure. There is no doubt that the "if it ain't broke, don't fix it" mentality will prevail in many legacy telecommunications infrastructures, thus preserving a copper wire PSTN for the foreseeable future. However, the appeal of the efficiencies of VoIP and 802.11 cannot be ignored and the end result may very well be a mix of technologies.

Disruptive Technology

In his Harvard University business book, *The Innovator's Dilemma* [4], author Clayton Christensen describes how disruptive technologies have precipitated the failure of leading products and their associated and well-managed firms.

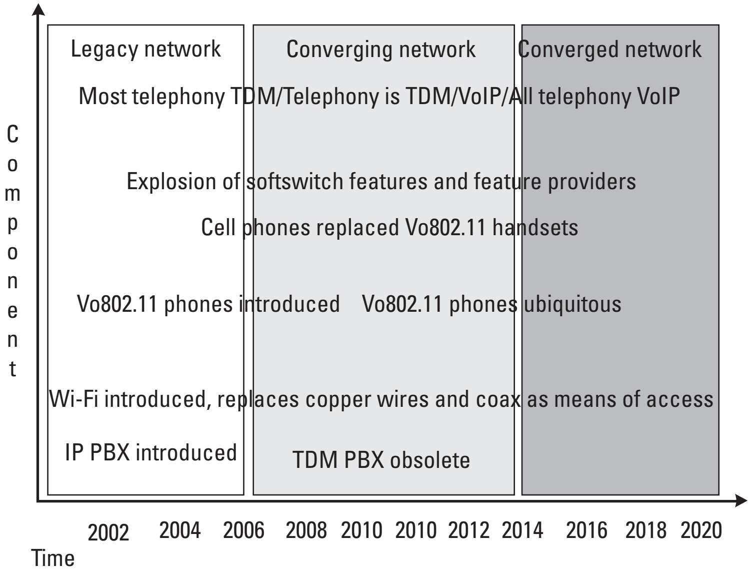

Legacy network
Converging network
Converged network
Most telephony TDM/Telephony is TDM/VoIP/All telephony VoIP
Explosion of softswitch features and feature providers
Cell phones replaced Vo802.11 handsets
Vo802.11 phones introduced
Vo802.11 phones ubiquitous
Wi-Fi introduced, replaces copper wires and coax as means of access
IP PBX introduced
TDM PBX obsolete
Component
2002
2004
2006
2008
2010
2010
2012
2014
2016
2018
2020
Time

Figure 15.2 Adoption of Vo802.11 timeline.

Christensen defines criteria to identify disruptive technologies regardless of their market. These technologies have the potential to replace mainstream technologies and their associated products and principal vendors. Disruptive technologies, abstractly defined by Christensen, are "typically cheaper, simpler, smaller, and, frequently, more convenient" than their mainstream counterparts.

Wireless technologies, relative to incumbent wired networks, are a disruptive technology. For the competitive service provider, 802.11b is "cheaper, simpler, smaller, and frequently, more convenient" than copper wire and its associated infrastructure. In order for a technology to be truly disruptive, it must "disrupt" an incumbent vendor or service provider. Some entity must go out of business before a technology can be considered "disruptive." While it is too early to point out incumbent service providers driven out of business by Vo802.11, its technologies are potentially disruptive to incumbent telephone companies. The migration of wire-line telephone traffic from ILEC to cellular is a powerful example of this trend. The migration to Vo802.11 will certainly mark the disruption of telephone companies as we know them.

How Vo802.11 Will Disrupt the Telephone Industry

Cheaper

Per Table 14.9, a Vo802.11 network is much cheaper to deploy than a comparable TDM-switched, copper wire-based legacy PSTN infrastructure. The Telecommunications Act of 1996 failed to produce any real competition in the local loop because it was economically impossible to build and deploy a network that could compete with an entrenched and financially protected monopoly.

Vo802.11 changes all that. A competitive network can be built for a fraction of the cost of a legacy network. Furthermore, it can be operated for a fraction of the OAM&P of the PSTN. Potentially, it offers more services than the PSTN, generating more revenue than a PSTN voice-based infrastructure.

By virtue of being cheaper to purchase and operate, a Vo802.11 network marks a significant lowering of barriers to market entry. No longer is a voice service the exclusive domain of a century-old protected monopoly. This lowering of the barrier to entry will allow multiple types of service providers to offer voice services in direct competition with the legacy telephone monopoly. This list of service providers could include WISPs, ISPs, power companies, municipalities, cable TV companies, and new market entrants.

Simpler

Given its 100-year evolution, the PSTN is painfully complex. Service providers have melded one technology on top of another during the last century. COs are, in many cases, museums of switching history because operators rarely discard switching equipment that still functions (and enjoys a very generous depreciation schedule).

Vo802.11 service providers will not be burdened by the past. Rather, a Vo802.11 is IP based, meaning it is far more efficient to operate. The key here is open standards as opposed to the closed systems of the legacy PSTN. The open standards allow a service provider to "mix and match" the components of the network. Much of a softswitched voice network is software dependent, which can be upgraded easily and frequently.

Smaller

One recurring excuse for the monopoly of telephone companies is that they posed an "economy of scale" in that something so large, so complex, and so costly could succeed only if it was protected as a monopoly. A Vo802.11 network can be easily deployed as a modular system by even the smallest service providers in rural or developing economies. The same is true of corporate campuses, or multiple-dwelling units. Given that softswitch operations are geographically independent of the subscriber, a service provider can provide switching for widely dispersed subscribers.

The footprint of a softswitch is less than 10% of that of a Class 5 switch handling the same or greater traffic load and does not have to be housed in a telco CO. Access points for PSTN replacement Vo802.11 networks are small (no more than one meter square for many products) and light. This makes deployment fast and inexpensive.

More Convenient to Use

The PSTN may be doomed by the commodity for which it was created: voice. Business and residential markets now demand convenient access to broadband data services. The PSTN does not offer this function efficiently. Vo802.11 networks offer easily deployed and operated broadband data services.

Vo802.11 networks, due to the flexibility of the softswitched infrastructure, offer the subscriber greater convenience because of the vast array of features made available by the softswitch and its associated feature servers.

This marks a high level of convenience for the service provider as well. Rather than wait years and spend millions of dollars to offer a new feature to a set of subscribers, the service providers can often write their own feature(s) in-house and deploy them in a matter of days.

Deconstruction

In their 2000 book entitled *Blown to Bits* [5], Phillip Evans and Thomas Wurster explore how certain industries have been "deconstructed" by the Internet. That is, the emergence of information or services available via the Internet has caused firms to lose sales and market share if not their entire business due to the emergence of new technologies. Examples of those industries include travel agencies, retail banks, and automobile retailers. We now investigate the potential deconstruction of the North American telecom industry by Internet-related telephony applications.

The telecom sector in recent years has been deconstructed, if not by the Internet itself, by technologies that are Internet related. Long-distance bypass using VoIP as described earlier in this book is a good example of such a technology. The delivery of telephony features to a voice service via IP would also be an example of deconstruction of the telecom service provider industry by an internet-related technology.

Deconstruction of Service Providers

Incumbent telephone service providers are deconstructed as their market share shifts from their networks to IP-based networks and applications. As incumbent service providers lose revenue, they have less money to spend on infrastructure. This inevitably impacts the vendors that supply incumbent service providers with legacy platforms (Class 4/5 switches and PBXs). Christensen's *Innovator's Dilemma* [4] describes "value networks" consisting of, in this case, service providers and the vendors that service them. As the service providers see their market shares and resultant revenues fall off, their vendors will also be adversely impacted.[1] Potential deconstruction of the telecom industry by Internet-related technologies focuses on service providers and the vendors that provide their infrastructure. A discussion of carriers is in order. Perhaps the most vulnerable are long-distance providers such as the "Big Three" (WorldCom, AT&T, and Sprint). Seventy percent of corporate telephone traffic is employee to employee, that is, office to office. As this traffic moves to the corporate WAN and the long-distance traffic, for example, becomes almost "free" (expense is bandwidth and the new VoIP infrastructure at the very edge of the network). To dramatize this point, if 70% of corporate long distance migrates away from the "Big Three" and onto the WAN, the "Big Three" will be severely deconstructed by Internet-related technology (VoIP). The new IP PBXs, especially those that are

[1] A value network, as defined by Clayton Christensen [4], incorporates both the physical attributes of a product or system as well as the associated cost structure (as typically measured by gross margin). This cost structure includes all business-related costs (e.g., research, engineering, development, sales, marketing).

SIP based, to quote Christensen, are "cheaper, simpler, smaller, and more convenient to use" than legacy PBXs.

Goetterdaemmerung or Creative Destruction in the Telecommunications Industry

Every month, North American local exchange carriers lose thousands of their TDM line accounts. On top of that, some are deeply in debt. Percentage-wise, this marks the only time since the Great Depression that telephone companies have actually decreased in line count.

How could the telephone company lose business? The answer is simple: Competition is slowly coming *to* as opposed to *in* the local loop. Subscribers are taking their business elsewhere. There are many competing technologies that allow subscribers to divorce themselves from the former monopolies. Many residential subscribers have given all their voice business to their cell phone service provider. Businesses have taken their voice business to data companies that offer VoIP over a data connection (ICG, Vonage). Capital expenditures for telephone companies are at record lows. The near-monopolistic vendors of the past are mired deeply in debt.

Is there no optimism in this market? If one is looking for a "recovery" in the telecommunications market as we know it, there is no cause for optimism. Austrian-born Harvard economist Josef Schumpeter, if he were alive today, would probably refer to the current telecommunications industry as being a good case of *creative destruction.* That is, capitalism is cyclical. Almost all industries grow, mature, and die.

The telecommunications industry as we know it is no exception to this rule of capitalism. Shielded as a quasi-monopoly for most of its life, the North American local exchange carrier had no reason to compete or to innovate. The service it provides, voice, is little changed from more than 100 years ago. The monopolistic protection came to an end with the Telecommunications Act of 1996. The resulting boom in the industry buoyed those incumbent carriers as the "high tide that raises all boats." The telecommunications bust has seen the demise of many competitors in the local loop, but has yet to seriously threaten the survival of the incumbents.

Vo802.11 potentially strikes at the very heart of the incumbent telco business paradigm that relied on a high barrier to entry to the voice market. Technology will inevitably march forward. Vo802.11 technology is "cheaper, simpler, smaller, and more convenient to use." It is disruptive technology that, after matching the incumbent technology, has qualities of its own that will allow it to supersede the incumbent's legacy infrastructure. Vo802.11, unlike incumbent circuit-switched infrastructures, is a technology that can be quickly and cheaply deployed anywhere in the world. The North American telephony market (services) is estimated to do almost $1 trillion in business annually. Service providers, regardless of the technologies they use, will, in a Darwinian struggle, seek to get an ever-increasing larger market share. That market share can only come at the expense of the incumbents.

In summary, there will not be a recovery in the North American telecommunications market. There will be a rebirth. That rebirth will come in the form of new service providers offering new services with new technology. When the exact date of the end of circuit-switched telephony and the century-old PSTN will come is not certain. The best analogy of this passing is in the Wagnerian opera *Goetterdaemmerung* or "twilight of the gods." *Daemmerung* in this case translates into "twilight," which in the German sense of the word can mean either the twilight at both dusk and dawn. In the case of the North American telecommunications market, it is the dusk for the incumbents and their legacy voice-only networks and it is dawn for Vo802.11.

References

[1] Federal Communications FCC Spectrum Policy Task Force, *Report of the Interference Protection Working Group*, November 15, 2002.

[2] Powell, M., "Broadband Migration—New Directions in Wireless Policy," *Silicon Flatirons Conference*, University of Colorado, Boulder, CO, October 30, 2002.

[3] Cahners' In-Stat, "Voice over Wireless LAN: 802.11x Hears the Call for Wireless VoIP," April 2002, http://www.instat.com/newmk.csp?ID=187.

[4] Christensen, C., *The Innovator's Dilemma,* Boston, MA: Harvard Business School Press, 1997.

[5] Evans, P., and T. S. Wurster, *Blown to Bits: How the New Economics of Information Transforms Strategy,* Boston, MA: Harvard Business School Press, 2000.

About the Author

Frank Ohrtman has many years of experience in VoIP and wireless applications. Mr. Ohrtman learned to perform in-depth research and write succinct analyses during his years as a Navy Intelligence Officer (1981–1991) where he specialized in electronic intelligence and electronic warfare. He is a veteran of U.S. Navy actions in Lebanon (where he was awarded the Navy Expeditionary Medal), Grenada, Libya (where he was awarded the Joint Service Commendation Medal) and the Gulf War (where he was awarded the National Defense Service Medal).

His career in VoIP began with selling VoIP gateway switches for Netrix Corporation to long-distance bypass carriers. He went on to promote softswitch solutions for Lucent Technologies (as a Qwest account manager) and Vsys (western region sales manager). Mr. Ohrtman is the author of *Softswitch: Architecture for Voice over IP* (McGraw-Hill, 2003), a number-one bestseller on USTA Bookstore's bestseller list, and *Wi-Fi Handbook: Building 802.11b Wireless Networks* (McGraw-Hill, 2003). He holds an M.S. in telecommunications from the University of Colorado's College of Engineering (his master's thesis was "Softswitch as Class 4 Replacement—A Disruptive Technology"), an M.A. in international relations from Boston University, and a B.A. in political science from the University of Iowa. Mr. Ohrtman lives in Denver, Colorado, where he is the president of Softswitch Consulting (http://www.softswitchconsulting.com).

Index

Recent Titles in the Artech House Telecommunications Library

Vinton G. Cerf, Senior Series Editor

Access Networks: Technology and V5 Interfacing, Alex Gillespie

Achieving Global Information Networking, Eve L. Varma et al.

Advanced High-Frequency Radio Communications, Eric E. Johnson et al.

ATM Interworking in Broadband Wireless Applications, M. Sreetharan and S. Subramaniam

ATM Switches, Edwin R. Coover

ATM Switching Systems, Thomas M. Chen and Stephen S. Liu

Broadband Access Technology, Interfaces, and Management, Alex Gillespie

Broadband Local Loops for High-Speed Internet Access, Maurice Gagnaire

Broadband Networking: ATM, SDH, and SONET, Mike Sexton and Andy Reid

Broadband Telecommunications Technology, Second Edition, Byeong Lee, Minho Kang, and Jonghee Lee

The Business Case for Web-Based Training, Tammy Whalen and David Wright

Centrex or PBX: The Impact of IP, John R. Abrahams and Mauro Lollo

Chinese Telecommunications Policy, Xu Yan and Douglas Pitt

Communication and Computing for Distributed Multimedia Systems, Guojun Lu

Communications Technology Guide for Business, Richard Downey, Seán Boland, and Phillip Walsh

Community Networks: Lessons from Blacksburg, Virginia, Second Edition, Andrew M. Cohill and Andrea Kavanaugh, editors

Component-Based Network System Engineering, Mark Norris, Rob Davis, and Alan Pengelly

Computer Telephony Integration, Second Edition, Rob Walters

Deploying and Managing IP over WDM Networks, Joan Serrat and Alex Galis, editors

Desktop Encyclopedia of the Internet, Nathan J. Muller

Digital Clocks for Synchronization and Communications, Masami Kihara, Sadayasu Ono, and Pekka Eskelinen

Digital Modulation Techniques, Fuqin Xiong

E-Commerce Systems Architecture and Applications, Wasim E. Rajput

Engineering Internet QoS, Sanjay Jha and Mahbub Hassan

Error-Control Block Codes for Communications Engineers, L. H. Charles Lee

FAX: Facsimile Technology and Systems, Third Edition,
Kenneth R. McConnell, Dennis Bodson, and Stephen Urban

Fundamentals of Network Security, John E. Canavan

Gigabit Ethernet Technology and Applications, Mark Norris

Guide to ATM Systems and Technology, Mohammad A. Rahman

A Guide to the TCP/IP Protocol Suite, Floyd Wilder

Home Networking Technologies and Standards, Theodore B. Zahariadis

Information Superhighways Revisited: The Economics of Multimedia,
Bruce Egan

Installation and Maintenance of SDH/SONET, ATM, xDSL, and Synchronization Networks, José M. Caballero et al.

Integrated Broadband Networks: TCP/IP, ATM, SDH/SONET, and WDM/Optics, Byeong Gi Lee and Woojune Kim

Internet E-mail: Protocols, Standards, and Implementation,
Lawrence Hughes

Introduction to Telecommunications Network Engineering, Second Edition,
Tarmo Anttalainen

Introduction to Telephones and Telephone Systems, Third Edition,
A. Michael Noll

An Introduction to U.S. Telecommunications Law, Second Edition,
Charles H. Kennedy

IP Convergence: The Next Revolution in Telecommunications,
Nathan J. Muller

LANs to WANs: The Complete Management Guide, Nathan J. Muller

The Law and Regulation of Telecommunications Carriers, Henk Brands and Evan T. Leo

Managing Internet-Driven Change in International Telecommunications,
Rob Frieden

Marketing Telecommunications Services: New Approaches for a Changing Environment, Karen G. Strouse

Mission-Critical Network Planning, Matthew Liotine

Multimedia Communications Networks: Technologies and Services,
Mallikarjun Tatipamula and Bhumip Khashnabish, editors

Next Generation Intelligent Networks, Johan Zuidweg

Open Source Software Law, Rod Dixon

Performance Evaluation of Communication Networks, Gary N. Higginbottom

Performance of TCP/IP over ATM Networks, Mahbub Hassan and Mohammed Atiquzzaman

Practical Guide for Implementing Secure Intranets and Extranets, Kaustubh M. Phaltankar

Practical Internet Law for Business, Kurt M. Saunders

Practical Multiservice LANs: ATM and RF Broadband, Ernest O. Tunmann

Principles of Modern Communications Technology, A. Michael Noll

Protocol Management in Computer Networking, Philippe Byrnes

Pulse Code Modulation Systems Design, William N. Waggener

Security, Rights, and Liabilities in E-Commerce, Jeffrey H. Matsuura

Service Level Management for Enterprise Networks, Lundy Lewis

SIP: Understanding the Session Initiation Protocol, Second Edition, Alan B. Johnston

Smart Card Security and Applications, Second Edition, Mike Hendry

SNMP-Based ATM Network Management, Heng Pan

Spectrum Wars: The Policy and Technology Debate, Jennifer A. Manner

Strategic Management in Telecommunications, James K. Shaw

Strategies for Success in the New Telecommunications Marketplace, Karen G. Strouse

Successful Business Strategies Using Telecommunications Services, Martin F. Bartholomew

Telecommunications Cost Management, S. C. Strother

Telecommunications Department Management, Robert A. Gable

Telecommunications Deregulation and the Information Economy, Second Edition, James K. Shaw

Telecommunications Technology Handbook, Second Edition, Daniel Minoli

Telemetry Systems Engineering, Frank Carden, Russell Jedlicka, and Robert Henry

Telephone Switching Systems, Richard A. Thompson

Understanding Modern Telecommunications and the Information Superhighway, John G. Nellist and Elliott M. Gilbert

Understanding Networking Technology: Concepts, Terms, and Trends, Second Edition, Mark Norris

Videoconferencing and Videotelephony: Technology and Standards, Second Edition, Richard Schaphorst

Visual Telephony, Edward A. Daly and Kathleen J. Hansell

Voice over 802.11, Frank Ohrtman

Wide-Area Data Network Performance Engineering, Robert G. Cole and Ravi Ramaswamy

Winning Telco Customers Using Marketing Databases, Rob Mattison

WLANs and WPANs towards 4G Wireless, Ramjee Prasad and Luis Muñoz

World-Class Telecommunications Service Development, Ellen P. Ward

For further information on these and other Artech House titles,
including previously considered out-of-print books now available through our
In-Print-Forever® (IPF®) program, contact:

Artech House
685 Canton Street
Norwood, MA 02062
Phone: 781-769-9750
Fax: 781-769-6334
e-mail: artech@artechhouse.com

Artech House
46 Gillingham Street
London SW1V 1AH UK
Phone: +44 (0)20 7596-8750
Fax: +44 (0)20 7630-0166
e-mail: artech-uk@artechhouse.com

Find us on the World Wide Web at:
www.artechhouse.com