

2.6 THE MOBILE TELEPHONE SYSTEM

The traditional telephone system (even if it some day gets multigigabit end-to-end fiber) will still not be able to satisfy a growing group of users: people on the go. People now expect to make phone calls from airplanes, cars, swimming pools, and while jogging in the park. Within a few years they will also expect to send e-mail and surf the Web from all these locations and more. Consequently, there is a tremendous amount of interest in wireless telephony. In the following sections we will study this topic in some detail.

Wireless telephones come in two basic varieties: cordless phones and mobile phones (sometimes called **cell phones**). **Cordless phones** are devices consisting of a base station and a handset sold as a set for use within the home. These are never used for networking, so we will not examine them further. Instead we will concentrate on the mobile system, which is used for wide area voice and data communication.

Mobile phones have gone through three distinct generations, with different technologies:

1. Analog voice.
2. Digital voice.
3. Digital voice and data (Internet, e-mail, etc.).

Although most of our discussion will be about the technology of these systems, it is interesting to note how political and tiny marketing decisions can have a huge impact. The first mobile system was devised in the U.S. by AT&T and mandated for the whole country by the FCC. As a result, the entire U.S. had a single (analog) system and a mobile phone purchased in California also worked in New York. In contrast, when mobile came to Europe, every country devised its own system, which resulted in a fiasco.

Europe learned from its mistake and when digital came around, the government-run PTTs got together and standardized on a single system (GSM), so any European mobile phone will work anywhere in Europe. By then, the U.S. had decided that government should not be in the standardization business, so it left digital to the marketplace. This decision resulted in different equipment manufacturers producing different kinds of mobile phones. As a consequence, the U.S. now has two major incompatible digital mobile phone systems in operation (plus one minor one).

Despite an initial lead by the U.S., mobile phone ownership and usage in Europe is now far greater than in the U.S. Having a single system for all of Europe is part of the reason, but there is more. A second area where the U.S. and Europe differed is in the humble matter of phone numbers. In the U.S. mobile phones are mixed in with regular (fixed) telephones. Thus, there is no way for a

caller to see if, say, (212) 234-5678 is a fixed telephone (cheap or free call) or a mobile phone (expensive call). To keep people from getting nervous about using the telephone, the telephone companies decided to make the mobile phone owner pay for incoming calls. As a consequence, many people hesitated to buy a mobile phone for fear of running up a big bill by just receiving calls. In Europe, mobile phones have a special area code (analogous to 800 and 900 numbers) so they are instantly recognizable. Consequently, the usual rule of “caller pays” also applies to mobile phones in Europe (except for international calls where costs are split).

A third issue that has had a large impact on adoption is the widespread use of prepaid mobile phones in Europe (up to 75% in some areas). These can be purchased in many stores with no more formality than buying a radio. You pay and you go. They are preloaded with, for example, 20 or 50 euro and can be recharged (using a secret PIN code) when the balance drops to zero. As a consequence, practically every teenager and many small children in Europe have (usually prepaid) mobile phones so their parents can locate them, without the danger of the child running up a huge bill. If the mobile phone is used only occasionally, its use is essentially free since there is no monthly charge or charge for incoming calls.

2.6.1 First-Generation Mobile Phones: Analog Voice

Enough about the politics and marketing aspects of mobile phones. Now let us look at the technology, starting with the earliest system. Mobile radiotelephones were used sporadically for maritime and military communication during the early decades of the 20th century. In 1946, the first system for car-based telephones was set up in St. Louis. This system used a single large transmitter on top of a tall building and had a single channel, used for both sending and receiving. To talk, the user had to push a button that enabled the transmitter and disabled the receiver. Such systems, known as **push-to-talk systems**, were installed in several cities beginning in the late 1950s. CB-radio, taxis, and police cars on television programs often use this technology.

In the 1960s, **IMTS (Improved Mobile Telephone System)** was installed. It, too, used a high-powered (200-watt) transmitter, on top of a hill, but now had two frequencies, one for sending and one for receiving, so the push-to-talk button was no longer needed. Since all communication from the mobile telephones went inbound on a different channel than the outbound signals, the mobile users could not hear each other (unlike the push-to-talk system used in taxis).

IMTS supported 23 channels spread out from 150 MHz to 450 MHz. Due to the small number of channels, users often had to wait a long time before getting a dial tone. Also, due to the large power of the hilltop transmitter, adjacent systems had to be several hundred kilometers apart to avoid interference. All in all, the limited capacity made the system impractical.

Advanced Mobile Phone System

All that changed with **AMPS (Advanced Mobile Phone System)**, invented by Bell Labs and first installed in the United States in 1982. It was also used in England, where it was called TACS, and in Japan, where it was called MCS-L1. Although no longer state of the art, we will look at it in some detail because many of its fundamental properties have been directly inherited by its digital successor, D-AMPS, in order to achieve backward compatibility.

In all mobile phone systems, a geographic region is divided up into **cells**, which is why the devices are sometimes called cell phones. In AMPS, the cells are typically 10 to 20 km across; in digital systems, the cells are smaller. Each cell uses some set of frequencies not used by any of its neighbors. The key idea that gives cellular systems far more capacity than previous systems is the use of relatively small cells and the reuse of transmission frequencies in nearby (but not adjacent) cells. Whereas an IMTS system 100 km across can have one call on each frequency, an AMPS system might have 100 10-km cells in the same area and be able to have 10 to 15 calls on each frequency, in widely separated cells. Thus, the cellular design increases the system capacity by at least an order of magnitude, more as the cells get smaller. Furthermore, smaller cells mean that less power is needed, which leads to smaller and cheaper transmitters and handsets. Hand-held telephones put out 0.6 watts; transmitters in cars are 3 watts, the maximum allowed by the FCC.

The idea of frequency reuse is illustrated in Fig. 2-41(a). The cells are normally roughly circular, but they are easier to model as hexagons. In Fig. 2-41(a), the cells are all the same size. They are grouped in units of seven cells. Each letter indicates a group of frequencies. Notice that for each frequency set, there is a buffer about two cells wide where that frequency is not reused, providing for good separation and low interference.

Finding locations high in the air to place base station antennas is a major issue. This problem has led some telecommunication carriers to forge alliances with the Roman Catholic Church, since the latter owns a substantial number of exalted potential antenna sites worldwide, all conveniently under a single management.

In an area where the number of users has grown to the point that the system is overloaded, the power is reduced, and the overloaded cells are split into smaller **microcells** to permit more frequency reuse, as shown in Fig. 2-41(b). Telephone companies sometimes create temporary microcells, using portable towers with satellite links at sporting events, rock concerts, and other places where large numbers of mobile users congregate for a few hours. How big the cells should be is a complex matter, which is treated in (Hac, 1995).

At the center of each cell is a base station to which all the telephones in the cell transmit. The base station consists of a computer and transmitter/receiver connected to an antenna. In a small system, all the base stations are connected to

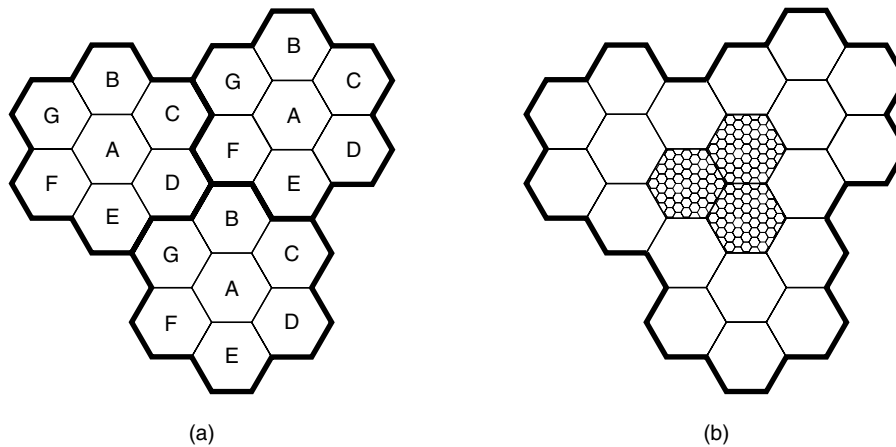


Figure 2-41. (a) Frequencies are not reused in adjacent cells. (b) To add more users, smaller cells can be used.

a single device called an **MTSO (Mobile Telephone Switching Office)** or **MSC (Mobile Switching Center)**. In a larger one, several MTSOs may be needed, all of which are connected to a second-level MTSO, and so on. The MTSOs are essentially end offices as in the telephone system, and are, in fact, connected to at least one telephone system end office. The MTSOs communicate with the base stations, each other, and the PSTN using a packet-switching network.

At any instant, each mobile telephone is logically in one specific cell and under the control of that cell's base station. When a mobile telephone physically leaves a cell, its base station notices the telephone's signal fading away and asks all the surrounding base stations how much power they are getting from it. The base station then transfers ownership to the cell getting the strongest signal, that is, the cell where the telephone is now located. The telephone is then informed of its new boss, and if a call is in progress, it will be asked to switch to a new channel (because the old one is not reused in any of the adjacent cells). This process, called **handoff**, takes about 300 msec. Channel assignment is done by the MTSO, the nerve center of the system. The base stations are really just radio relays.

Handoffs can be done in two ways. In a **soft handoff**, the telephone is acquired by the new base station before the previous one signs off. In this way there is no loss of continuity. The downside here is that the telephone needs to be able to tune to two frequencies at the same time (the old one and the new one). Neither first nor second generation devices can do this.

In a **hard handoff**, the old base station drops the telephone before the new one acquires it. If the new one is unable to acquire it (e.g., because there is no available frequency), the call is disconnected abruptly. Users tend to notice this, but it is inevitable occasionally with the current design.

Channels

The AMPS system uses 832 full-duplex channels, each consisting of a pair of simplex channels. There are 832 simplex transmission channels from 824 to 849 MHz and 832 simplex receive channels from 869 to 894 MHz. Each of these simplex channels is 30 kHz wide. Thus, AMPS uses FDM to separate the channels.

In the 800-MHz band, radio waves are about 40 cm long and travel in straight lines. They are absorbed by trees and plants and bounce off the ground and buildings. It is possible that a signal sent by a mobile telephone will reach the base station by the direct path, but also slightly later after bouncing off the ground or a building. This may lead to an echo or signal distortion (multipath fading). Sometimes, it is even possible to hear a distant conversation that has bounced several times.

The 832 channels are divided into four categories:

1. Control (base to mobile) to manage the system.
2. Paging (base to mobile) to alert mobile users to calls for them.
3. Access (bidirectional) for call setup and channel assignment.
4. Data (bidirectional) for voice, fax, or data.

Twenty-one of the channels are reserved for control, and these are wired into a PROM in each telephone. Since the same frequencies cannot be reused in nearby cells, the actual number of voice channels available per cell is much smaller than 832, typically about 45.

Call Management

Each mobile telephone in AMPS has a 32-bit serial number and a 10-digit telephone number in its PROM. The telephone number is represented as a 3-digit area code in 10 bits, and a 7-digit subscriber number in 24 bits. When a phone is switched on, it scans a preprogrammed list of 21 control channels to find the most powerful signal.

The phone then broadcasts its 32-bit serial number and 34-bit telephone number. Like all the control information in AMPS, this packet is sent in digital form, multiple times, and with an error-correcting code, even though the voice channels themselves are analog.

When the base station hears the announcement, it tells the MTSO, which records the existence of its new customer and also informs the customer's home MTSO of his current location. During normal operation, the mobile telephone re-registers about once every 15 minutes.

To make a call, a mobile user switches on the phone, enters the number to be called on the keypad, and hits the SEND button. The phone then transmits the

number to be called and its own identity on the access channel. If a collision occurs there, it tries again later. When the base station gets the request, it informs the MTSO. If the caller is a customer of the MTSO's company (or one of its partners), the MTSO looks for an idle channel for the call. If one is found, the channel number is sent back on the control channel. The mobile phone then automatically switches to the selected voice channel and waits until the called party picks up the phone.

Incoming calls work differently. To start with, all idle phones continuously listen to the paging channel to detect messages directed at them. When a call is placed to a mobile phone (either from a fixed phone or another mobile phone), a packet is sent to the callee's home MTSO to find out where it is. A packet is then sent to the base station in its current cell, which then sends a broadcast on the paging channel of the form "Unit 14, are you there?" The called phone then responds with "Yes" on the access channel. The base then says something like: "Unit 14, call for you on channel 3." At this point, the called phone switches to channel 3 and starts making ringing sounds (or playing some melody the owner was given as a birthday present).

2.6.2 Second-Generation Mobile Phones: Digital Voice

The first generation of mobile phones was analog; the second generation was digital. Just as there was no worldwide standardization during the first generation, there was also no standardization during the second, either. Four systems are in use now: D-AMPS, GSM, CDMA, and PDC. Below we will discuss the first three. PDC is used only in Japan and is basically D-AMPS modified for backward compatibility with the first-generation Japanese analog system. The name **PCS (Personal Communications Services)** is sometimes used in the marketing literature to indicate a second-generation (i.e., digital) system. Originally it meant a mobile phone using the 1900 MHz band, but that distinction is rarely made now.

D-AMPS—The Digital Advanced Mobile Phone System

The second generation of the AMPS systems is **D-AMPS** and is fully digital. It is described in International Standard IS-54 and its successor IS-136. D-AMPS was carefully designed to co-exist with AMPS so that both first- and second-generation mobile phones could operate simultaneously in the same cell. In particular, D-AMPS uses the same 30 kHz channels as AMPS and at the same frequencies so that one channel can be analog and the adjacent ones can be digital. Depending on the mix of phones in a cell, the cell's MTSO determines which channels are analog and which are digital, and it can change channel types dynamically as the mix of phones in a cell changes.

When D-AMPS was introduced as a service, a new frequency band was made available to handle the expected increased load. The upstream channels were in

Using better compression algorithms, it is possible to get the speech down to 4 kbps, in which case six users can be stuffed into a frame, as illustrated in Fig. 2-42(b). From the operator's perspective, being able to squeeze three to six times as many D-AMPS users into the same spectrum as one AMPS user is a huge win and explains much of the popularity of PCS. Of course, the quality of speech at 4 kbps is not comparable to what can be achieved at 56 kbps, but few PCS operators advertise their hi-fi sound quality. It should also be clear that for data, an 8 kbps channel is not even as good as an ancient 9600-bps modem.

The control structure of D-AMPS is fairly complicated. Briefly summarized, groups of 16 frames form a superframe, with certain control information present in each superframe a limited number of times. Six main control channels are used: system configuration, real-time and nonreal-time control, paging, access response, and short messages. But conceptually, it works like AMPS. When a mobile is switched on, it makes contact with the base station to announce itself and then listens on a control channel for incoming calls. Having picked up a new mobile, the MTSO informs the user's home base where he is, so calls can be routed correctly.

One difference between AMPS and D-AMPS is how handoff is handled. In AMPS, the MTSO manages it completely without help from the mobile devices. As can be seen from Fig. 2-42, in D-AMPS, 1/3 of the time a mobile is neither sending nor receiving. It uses these idle slots to measure the line quality. When it discovers that the signal is waning, it complains to the MTSO, which can then break the connection, at which time the mobile can try to tune to a stronger signal from another base station. As in AMPS, it still takes about 300 msec to do the handoff. This technique is called **MAHO (Mobile Assisted HandOff)**.

GSM—The Global System for Mobile Communications

D-AMPS is widely used in the U.S. and (in modified form) in Japan. Virtually everywhere else in the world, a system called **GSM (Global System for Mobile communications)** is used, and it is even starting to be used in the U.S. on a limited scale. To a first approximation, GSM is similar to D-AMPS. Both are cellular systems. In both systems, frequency division multiplexing is used, with each mobile transmitting on one frequency and receiving on a higher frequency (80 MHz higher for D-AMPS, 55 MHz higher for GSM). Also in both systems, a single frequency pair is split by time-division multiplexing into time slots shared by multiple mobiles. However, the GSM channels are much wider than the AMPS channels (200 kHz versus 30 kHz) and hold relatively few additional users (8 versus 3), giving GSM a much higher data rate per user than D-AMPS.

Below we will briefly discuss some of the main properties of GSM. However, the printed GSM standard is over 5000 [sic] pages long. A large fraction of this material relates to engineering aspects of the system, especially the design of

receivers to handle multipath signal propagation, and synchronizing transmitters and receivers. None of this will be even mentioned below.

Each frequency band is 200 kHz wide, as shown in Fig. 2-43. A GSM system has 124 pairs of simplex channels. Each simplex channel is 200 kHz wide and supports eight separate connections on it, using time division multiplexing. Each currently active station is assigned one time slot on one channel pair. Theoretically, 992 channels can be supported in each cell, but many of them are not available, to avoid frequency conflicts with neighboring cells. In Fig. 2-43, the eight shaded time slots all belong to the same connection, four of them in each direction. Transmitting and receiving does not happen in the same time slot because the GSM radios cannot transmit and receive at the same time and it takes time to switch from one to the other. If the mobile station assigned to 890.4/935.4 MHz and time slot 2 wanted to transmit to the base station, it would use the lower four shaded slots (and the ones following them in time), putting some data in each slot until all the data had been sent.

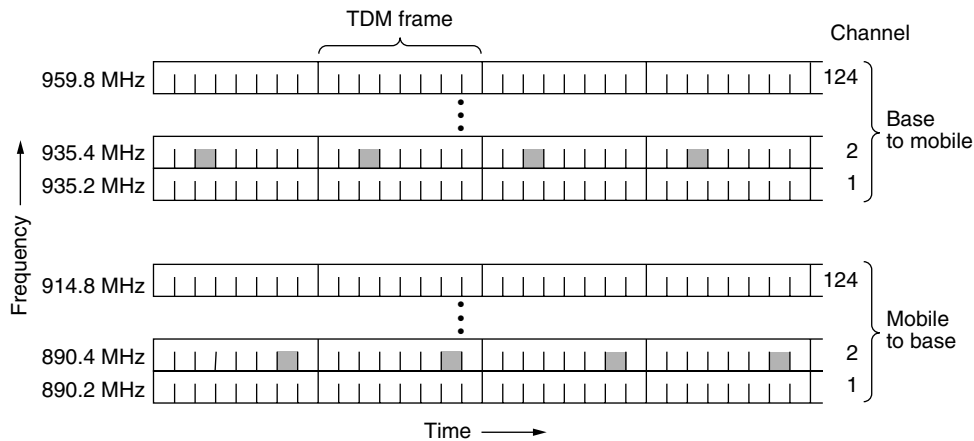


Figure 2-43. GSM uses 124 frequency channels, each of which uses an eight-slot TDM system.

The TDM slots shown in Fig. 2-43 are part of a complex framing hierarchy. Each TDM slot has a specific structure, and groups of TDM slots form multi-frames, also with a specific structure. A simplified version of this hierarchy is shown in Fig. 2-44. Here we can see that each TDM slot consists of a 148-bit data frame that occupies the channel for 577 μsec (including a 30- μsec guard time after each slot). Each data frame starts and ends with three 0 bits, for frame delineation purposes. It also contains two 57-bit *Information* fields, each one having a control bit that indicates whether the following *Information* field is for voice or data. Between the *Information* fields is a 26-bit *Sync* (training) field that is used by the receiver to synchronize to the sender's frame boundaries.

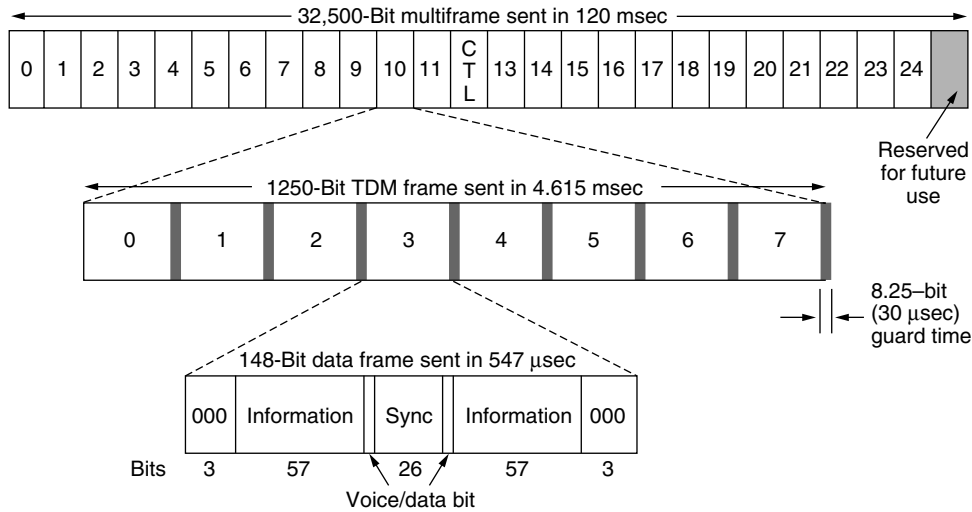


Figure 2-44. A portion of the GSM framing structure.

A data frame is transmitted in 547 μsec , but a transmitter is only allowed to send one data frame every 4.615 msec, since it is sharing the channel with seven other stations. The gross rate of each channel is 270,833 bps, divided among eight users. This gives 33.854 kbps gross, more than double D-AMPS' 324 bits 50 times per second for 16.2 kbps. However, as with AMPS, the overhead eats up a large fraction of the bandwidth, ultimately leaving 24.7 kbps worth of payload per user before error correction. After error correction, 13 kbps is left for speech, giving substantially better voice quality than D-AMPS (at the cost of using correspondingly more bandwidth).

As can be seen from Fig. 2-44, eight data frames make up a TDM frame and 26 TDM frames make up a 120-msec multiframe. Of the 26 TDM frames in a multiframe, slot 12 is used for control and slot 25 is reserved for future use, so only 24 are available for user traffic.

However, in addition to the 26-slot multiframe shown in Fig. 2-44, a 51-slot multiframe (not shown) is also used. Some of these slots are used to hold several control channels used to manage the system. The **broadcast control channel** is a continuous stream of output from the base station containing the base station's identity and the channel status. All mobile stations monitor their signal strength to see when they have moved into a new cell.

The **dedicated control channel** is used for location updating, registration, and call setup. In particular, each base station maintains a database of mobile stations currently under its jurisdiction. Information needed to maintain this database is sent on the dedicated control channel.

Finally, there is the **common control channel**, which is split up into three logical subchannels. The first of these subchannels is the **paging channel**, which

the base station uses to announce incoming calls. Each mobile station monitors it continuously to watch for calls it should answer. The second is the **random access channel**, which allows users to request a slot on the dedicated control channel. If two requests collide, they are garbled and have to be retried later. Using the dedicated control channel slot, the station can set up a call. The assigned slot is announced on the third subchannel, the **access grant channel**.

CDMA—Code Division Multiple Access

D-AMPS and GSM are fairly conventional systems. They use both FDM and TDM to divide the spectrum into channels and the channels into time slots. However, there is a third kid on the block, **CDMA (Code Division Multiple Access)**, which works completely differently. When CDMA was first proposed, the industry gave it approximately the same reaction that Columbus first got from Queen Isabella when he proposed reaching India by sailing in the wrong direction. However, through the persistence of a single company, Qualcomm, CDMA has matured to the point where it is not only acceptable, it is now viewed as the best technical solution around and the basis for the third-generation mobile systems. It is also widely used in the U.S. in second-generation mobile systems, competing head-on with D-AMPS. For example, Sprint PCS uses CDMA, whereas AT&T Wireless uses D-AMPS. CDMA is described in International Standard IS-95 and is sometimes referred to by that name. The brand name **cdmaOne** is also used.

CDMA is completely different from AMPS, D-AMPS, and GSM. Instead of dividing the allowed frequency range into a few hundred narrow channels, CDMA allows each station to transmit over the entire frequency spectrum all the time. Multiple simultaneous transmissions are separated using coding theory. CDMA also relaxes the assumption that colliding frames are totally garbled. Instead, it assumes that multiple signals add linearly.

Before getting into the algorithm, let us consider an analogy: an airport lounge with many pairs of people conversing. TDM is comparable to all the people being in the middle of the room but taking turns speaking. FDM is comparable to the people being in widely separated clumps, each clump holding its own conversation at the same time as, but still independent of, the others. CDMA is comparable to everybody being in the middle of the room talking at once, but with each pair in a different language. The French-speaking couple just hones in on the French, rejecting everything that is not French as noise. Thus, the key to CDMA is to be able to extract the desired signal while rejecting everything else as random noise. A somewhat simplified description of CDMA follows.

In CDMA, each bit time is subdivided into m short intervals called **chips**. Typically, there are 64 or 128 chips per bit, but in the example given below we will use 8 chips/bit for simplicity.

Each station is assigned a unique m -bit code called a **chip sequence**. To transmit a 1 bit, a station sends its chip sequence. To transmit a 0 bit, it sends the

one's complement of its chip sequence. No other patterns are permitted. Thus, for $m = 8$, if station *A* is assigned the chip sequence 00011011, it sends a 1 bit by sending 00011011 and a 0 bit by sending 11100100.

Increasing the amount of information to be sent from b bits/sec to mb chips/sec can only be done if the bandwidth available is increased by a factor of m , making CDMA a form of spread spectrum communication (assuming no changes in the modulation or encoding techniques). If we have a 1-MHz band available for 100 stations, with FDM each one would have 10 kHz and could send at 10 kbps (assuming 1 bit per Hz). With CDMA, each station uses the full 1 MHz, so the chip rate is 1 megachip per second. With fewer than 100 chips per bit, the effective bandwidth per station is higher for CDMA than FDM, and the channel allocation problem is also solved.

For pedagogical purposes, it is more convenient to use a bipolar notation, with binary 0 being -1 and binary 1 being $+1$. We will show chip sequences in parentheses, so a 1 bit for station *A* now becomes $(-1 -1 -1 +1 +1 -1 +1 +1)$. In Fig. 2-45(a) we show the binary chip sequences assigned to four example stations. In Fig. 2-45(b) we show them in our bipolar notation.

A: 0 0 0 1 1 0 1 1	A: (-1 -1 -1 +1 +1 -1 +1 +1)
B: 0 0 1 0 1 1 1 0	B: (-1 -1 +1 -1 +1 +1 +1 -1)
C: 0 1 0 1 1 1 0 0	C: (-1 +1 -1 +1 +1 +1 -1 -1)
D: 0 1 0 0 0 0 1 0	D: (-1 +1 -1 -1 -1 -1 +1 -1)
(a)	(b)

Six examples:

-- 1 - C	S ₁ = (-1 +1 -1 +1 +1 +1 -1 -1)
- 1 1 - B + C	S ₂ = (-2 0 0 0 +2 +2 0 -2)
1 0 - - A + B	S ₃ = (0 0 -2 +2 0 -2 0 +2)
1 0 1 - A + B + C	S ₄ = (-1 +1 -3 +3 +1 -1 -1 +1)
1 1 1 1 A + B + C + D	S ₅ = (-4 0 -2 0 +2 0 +2 -2)
1 1 0 1 A + B + C + D	S ₆ = (-2 -2 0 -2 0 -2 +4 0)
(c)	

S ₁ • C = (1 +1 +1 +1 +1 +1 +1 +1)/8 = 1
S ₂ • C = (2 +0 +0 +0 +2 +2 +0 +2)/8 = 1
S ₃ • C = (0 +0 +2 +2 +0 -2 +0 -2)/8 = 0
S ₄ • C = (1 +1 +3 +3 +1 -1 +1 -1)/8 = 1
S ₅ • C = (4 +0 +2 +0 +2 +0 -2 +2)/8 = 1
S ₆ • C = (2 -2 +0 -2 +0 -2 -4 +0)/8 = -1
(d)

Figure 2-45. (a) Binary chip sequences for four stations. (b) Bipolar chip sequences. (c) Six examples of transmissions. (d) Recovery of station *C*'s signal.

Each station has its own unique chip sequence. Let us use the symbol **S** to indicate the m -chip vector for station *S*, and $\bar{\mathbf{S}}$ for its negation. All chip sequences

are pairwise **orthogonal**, by which we mean that the normalized inner product of any two distinct chip sequences, \mathbf{S} and \mathbf{T} (written as $\mathbf{S} \bullet \mathbf{T}$), is 0. It is known how to generate such orthogonal chip sequences using a method known as **Walsh codes**. In mathematical terms, orthogonality of the chip sequences can be expressed as follows:

$$\mathbf{S} \bullet \mathbf{T} \equiv \frac{1}{m} \sum_{i=1}^m S_i T_i = 0 \quad (2-4)$$

In plain English, as many pairs are the same as are different. This orthogonality property will prove crucial later on. Note that if $\mathbf{S} \bullet \mathbf{T} = 0$, then $\mathbf{S} \bullet \overline{\mathbf{T}}$ is also 0. The normalized inner product of any chip sequence with itself is 1:

$$\mathbf{S} \bullet \mathbf{S} = \frac{1}{m} \sum_{i=1}^m S_i S_i = \frac{1}{m} \sum_{i=1}^m S_i^2 = \frac{1}{m} \sum_{i=1}^m (\pm 1)^2 = 1$$

This follows because each of the m terms in the inner product is 1, so the sum is m . Also note that $\mathbf{S} \bullet \overline{\mathbf{S}} = -1$.

During each bit time, a station can transmit a 1 by sending its chip sequence, it can transmit a 0 by sending the negative of its chip sequence, or it can be silent and transmit nothing. For the moment, we assume that all stations are synchronized in time, so all chip sequences begin at the same instant.

When two or more stations transmit simultaneously, their bipolar signals add linearly. For example, if in one chip period three stations output +1 and one station outputs -1, the result is +2. One can think of this as adding voltages: three stations outputting +1 volts and 1 station outputting -1 volts gives 2 volts.

In Fig. 2-45(c) we see six examples of one or more stations transmitting at the same time. In the first example, C transmits a 1 bit, so we just get C 's chip sequence. In the second example, both B and C transmit 1 bits, so we get the sum of their bipolar chip sequences, namely:

$$(-1 -1 +1 -1 +1 +1 +1 -1) + (-1 +1 -1 +1 +1 +1 -1 -1) = (-2 0 0 0 +2 +2 0 -2)$$

In the third example, station A sends a 1 and station B sends a 0. The others are silent. In the fourth example, A and C send a 1 bit while B sends a 0 bit. In the fifth example, all four stations send a 1 bit. Finally, in the last example, A , B , and D send a 1 bit, while C sends a 0 bit. Note that each of the six sequences S_1 through S_6 given in Fig. 2-45(c) represents only one bit time.

To recover the bit stream of an individual station, the receiver must know that station's chip sequence in advance. It does the recovery by computing the normalized inner product of the received chip sequence (the linear sum of all the stations that transmitted) and the chip sequence of the station whose bit stream it is trying to recover. If the received chip sequence is \mathbf{S} and the receiver is trying to listen to a station whose chip sequence is \mathbf{C} , it just computes the normalized inner product, $\mathbf{S} \bullet \mathbf{C}$.

To see why this works, just imagine that two stations, A and C , both transmit a 1 bit at the same time that B transmits a 0 bit. The receiver sees the sum, $\mathbf{S} = \mathbf{A} + \bar{\mathbf{B}} + \mathbf{C}$ and computes

$$\mathbf{S} \cdot \mathbf{C} = (\mathbf{A} + \bar{\mathbf{B}} + \mathbf{C}) \cdot \mathbf{C} = \mathbf{A} \cdot \mathbf{C} + \bar{\mathbf{B}} \cdot \mathbf{C} + \mathbf{C} \cdot \mathbf{C} = 0 + 0 + 1 = 1$$

The first two terms vanish because all pairs of chip sequences have been carefully chosen to be orthogonal, as shown in Eq. (2-4). Now it should be clear why this property must be imposed on the chip sequences.

An alternative way of thinking about this situation is to imagine that the three chip sequences all came in separately, rather than summed. Then, the receiver would compute the inner product with each one separately and add the results. Due to the orthogonality property, all the inner products except $\mathbf{C} \cdot \mathbf{C}$ would be 0. Adding them and then doing the inner product is in fact the same as doing the inner products and then adding those.

To make the decoding process more concrete, let us consider the six examples of Fig. 2-45(c) again as illustrated in Fig. 2-45(d). Suppose that the receiver is interested in extracting the bit sent by station C from each of the six sums S_1 through S_6 . It calculates the bit by summing the pairwise products of the received \mathbf{S} and the \mathbf{C} vector of Fig. 2-45(b) and then taking $1/8$ of the result (since $m = 8$ here). As shown, the correct bit is decoded each time. It is just like speaking French.

In an ideal, noiseless CDMA system, the capacity (i.e., number of stations) can be made arbitrarily large, just as the capacity of a noiseless Nyquist channel can be made arbitrarily large by using more and more bits per sample. In practice, physical limitations reduce the capacity considerably. First, we have assumed that all the chips are synchronized in time. In reality, such synchronization is impossible. What can be done is that the sender and receiver synchronize by having the sender transmit a predefined chip sequence that is long enough for the receiver to lock onto. All the other (unsynchronized) transmissions are then seen as random noise. If there are not too many of them, however, the basic decoding algorithm still works fairly well. A large body of theory exists relating the superposition of chip sequences to noise level (Pickholtz et al., 1982). As one might expect, the longer the chip sequence, the higher the probability of detecting it correctly in the presence of noise. For extra reliability, the bit sequence can use an error-correcting code. Chip sequences never use error-correcting codes.

An implicit assumption in our discussion is that the power levels of all stations are the same as perceived by the receiver. CDMA is typically used for wireless systems with a fixed base station and many mobile stations at varying distances from it. The power levels received at the base station depend on how far away the transmitters are. A good heuristic here is for each mobile station to transmit to the base station at the inverse of the power level it receives from the base station. In other words, a mobile station receiving a weak signal from the will use more power than one getting a strong signal. The base station can also

give explicit commands to the mobile stations to increase or decrease their transmission power.

We have also assumed that the receiver knows who the sender is. In principle, given enough computing capacity, the receiver can listen to all the senders at once by running the decoding algorithm for each of them in parallel. In real life, suffice it to say that this is easier said than done. CDMA also has many other complicating factors that have been glossed over in this brief introduction. Nevertheless, CDMA is a clever scheme that is being rapidly introduced for wireless mobile communication. It normally operates in a band of 1.25 MHz (versus 30 kHz for D-AMPS and 200 kHz for GSM), but it supports many more users in that band than either of the other systems. In practice, the bandwidth available to each user is at least as good as GSM and often much better.

Engineers who want to gain a very deep understanding of CDMA should read (Lee and Miller, 1998). An alternative spreading scheme, in which the spreading is over time rather than frequency, is described in (Crespo et al., 1995). Yet another scheme is described in (Sari et al., 2000). All of these references require quite a bit of background in communication engineering.

2.6.3 Third-Generation Mobile Phones: Digital Voice and Data

What is the future of mobile telephony? Let us take a quick look. A number of factors are driving the industry. First, data traffic already exceeds voice traffic on the fixed network and is growing exponentially, whereas voice traffic is essentially flat. Many industry experts expect data traffic to dominate voice on mobile devices as well soon. Second, the telephone, entertainment, and computer industries have all gone digital and are rapidly converging. Many people are drooling over a lightweight, portable device that acts as a telephone, CD player, DVD player, e-mail terminal, Web interface, gaming machine, word processor, and more, all with worldwide wireless connectivity to the Internet at high bandwidth. This device and how to connect it is what third generation mobile telephony is all about. For more information, see (Huber et al., 2000; and Sarikaya, 2000).

Back in 1992, ITU tried to get a bit more specific about this dream and issued a blueprint for getting there called **IMT-2000**, where IMT stood for **International Mobile Telecommunications**. The number 2000 stood for three things: (1) the year it was supposed to go into service, (2) the frequency it was supposed to operate at (in MHz), and (3) the bandwidth the service should have (in kHz).

It did not make it on any of the three counts. Nothing was implemented by 2000. ITU recommended that all governments reserve spectrum at 2 GHz so devices could roam seamlessly from country to country. China reserved the required bandwidth but nobody else did. Finally, it was recognized that 2 Mbps is not currently feasible for users who are *too* mobile (due to the difficulty of performing handoffs quickly enough). More realistic is 2 Mbps for stationary indoor users (which will compete head-on with ADSL), 384 kbps for people walking,

and 144 kbps for connections in cars. Nevertheless, the whole area of **3G**, as it is called, is one great cauldron of activity. The third generation may be a bit less than originally hoped for and a bit late, but it will surely happen.

The basic services that the IMT-2000 network is supposed to provide to its users are:

1. High-quality voice transmission.
2. Messaging (replacing e-mail, fax, SMS, chat, etc.).
3. Multimedia (playing music, viewing videos, films, television, etc.).
4. Internet access (Web surfing, including pages with audio and video).

Additional services might be video conferencing, telepresence, group game playing, and m-commerce (waving your telephone at the cashier to pay in a store). Furthermore, all these services are supposed to be available worldwide (with automatic connection via a satellite when no terrestrial network can be located), instantly (always on), and with quality-of-service guarantees.

ITU envisioned a single worldwide technology for IMT-2000, so that manufacturers could build a single device that could be sold and used anywhere in the world (like CD players and computers and unlike mobile phones and televisions). Having a single technology would also make life much simpler for network operators and would encourage more people to use the services. Format wars, such as the Betamax versus VHS battle when videorecorders first came out, are not good for business.

Several proposals were made, and after some winnowing, it came down to two main ones. The first one, **W-CDMA (Wideband CDMA)**, was proposed by Ericsson. This system uses direct sequence spread spectrum of the type we described above. It runs in a 5 MHz bandwidth and has been designed to interwork with GSM networks although it is not backward compatible with GSM. It does, however, have the property that a caller can leave a W-CDMA cell and enter a GSM cell without losing the call. This system was pushed hard by the European Union, which called it **UMTS (Universal Mobile Telecommunications System)**.

The other contender was **CDMA2000**, proposed by Qualcomm. It, too, is a direct sequence spread spectrum design, basically an extension of IS-95 and backward compatible with it. It also uses a 5-MHz bandwidth, but it has not been designed to interwork with GSM and cannot hand off calls to a GSM cell (or a D-AMPS cell, for that matter). Other technical differences with W-CDMA include a different chip rate, different frame time, different spectrum used, and a different way to do time synchronization.

If the Ericsson and Qualcomm engineers were put in a room and told to come to a common design, they probably could. After all, the basic principle behind both systems is CDMA in a 5 MHz channel and nobody is willing to die for his

preferred chip rate. The trouble is that the real problem is not engineering, but politics (as usual). Europe wanted a system that interworked with GSM; the U.S. wanted a system that was compatible with one already widely deployed in the U.S. (IS-95). Each side also supported its local company (Ericsson is based in Sweden; Qualcomm is in California). Finally, Ericsson and Qualcomm were involved in numerous lawsuits over their respective CDMA patents.

In March 1999, the two companies settled the lawsuits when Ericsson agreed to buy Qualcomm's infrastructure. They also agreed to a single 3G standard, but one with multiple incompatible options, which to a large extent just papers over the technical differences. These disputes notwithstanding, 3G devices and services are likely to start appearing in the coming years.

Much has been written about 3G systems, most of it praising it as the greatest thing since sliced bread. Some references are (Collins and Smith, 2001; De Vriendt et al., 2002; Harte et al., 2002; Lu, 2002; and Sarikaya, 2000). However, some dissenters think that the industry is pointed in the wrong direction (Garber, 2002; and Goodman, 2000).

While waiting for the fighting over 3G to stop, some operators are gingerly taking a cautious small step in the direction of 3G by going to what is sometimes called **2.5G**, although 2.1G might be more accurate. One such system is **EDGE (Enhanced Data rates for GSM Evolution)**, which is just GSM with more bits per baud. The trouble is, more bits per baud also means more errors per baud, so EDGE has nine different schemes for modulation and error correction, differing on how much of the bandwidth is devoted to fixing the errors introduced by the higher speed.

Another 2.5G scheme is **GPRS (General Packet Radio Service)**, which is an overlay packet network on top of D-AMPS or GSM. It allows mobile stations to send and receive IP packets in a cell running a voice system. When GPRS is in operation, some time slots on some frequencies are reserved for packet traffic. The number and location of the time slots can be dynamically managed by the base station, depending on the ratio of voice to data traffic in the cell.

The available time slots are divided into several logical channels, used for different purposes. The base station determines which logical channels are mapped onto which time slots. One logical channel is for downloading packets from the base station to some mobile station, with each packet indicating who it is destined for. To send an IP packet, a mobile station requests one or more time slots by sending a request to the base station. If the request arrives without damage, the base station announces the frequency and time slots allocated to the mobile for sending the packet. Once the packet has arrived at the base station, it is transferred to the Internet by a wired connection. Since GPRS is just an overlay over the existing voice system, it is at best a stop-gap measure until 3G arrives.

Even though 3G networks are not fully deployed yet, some researchers regard 3G as a done deal and thus not interesting any more. These people are already working on 4G systems (Berezdivin et al., 2002; Guo and Chaskar, 2002; Huang

and Zhuang, 2002; Kellerer et al., 2002; and Misra et al., 2002). Some of the proposed features of 4G systems include high bandwidth, ubiquity (connectivity everywhere), seamless integration with wired networks and especially IP, adaptive resource and spectrum management, software radios, and high quality of service for multimedia.

Then on the other hand, so many 802.11 wireless LAN access points are being set up all over the place, that some people think 3G is not only not a done deal, it is doomed. In this vision, people will just wander from one 802.11 access point to another to stay connected. To say the industry is in a state of enormous flux is a huge understatement. Check back in about 5 years to see what happens.