MOBILE SATELLITE COMMUNICATION

Mobile satellite systems provide communications services to mobile and portable terminals using a radio transmission path between the terminal and the satellite. An example of such a system, illustrating its typical components, is shown in Fig. 1. The mobile terminal may be installed in any one of a number of platforms including cars, trucks, rail cars, aircraft, and ships. Alternatively, it could be a portable terminal with a size ranging from that of a hand-held unit up to that of a briefcase, depending upon the system and the provided service. Yet a third class could be small but fixed remote terminals serving functions such as seismic data collection and pipeline monitoring and control. A mobile satellite system requires one or more satellites with connectivity to the terrestrial infrastructure (e.g., the public switched telephone network and to the various digital networks) being supplied by one or more earth stations. Typically, most of the communications traffic is between the mobile terminal and another terminal or application outside of the mobile satellite system. However, most mobile satellite systems allow for mobile-to-mobile communications within the system. The earth stations

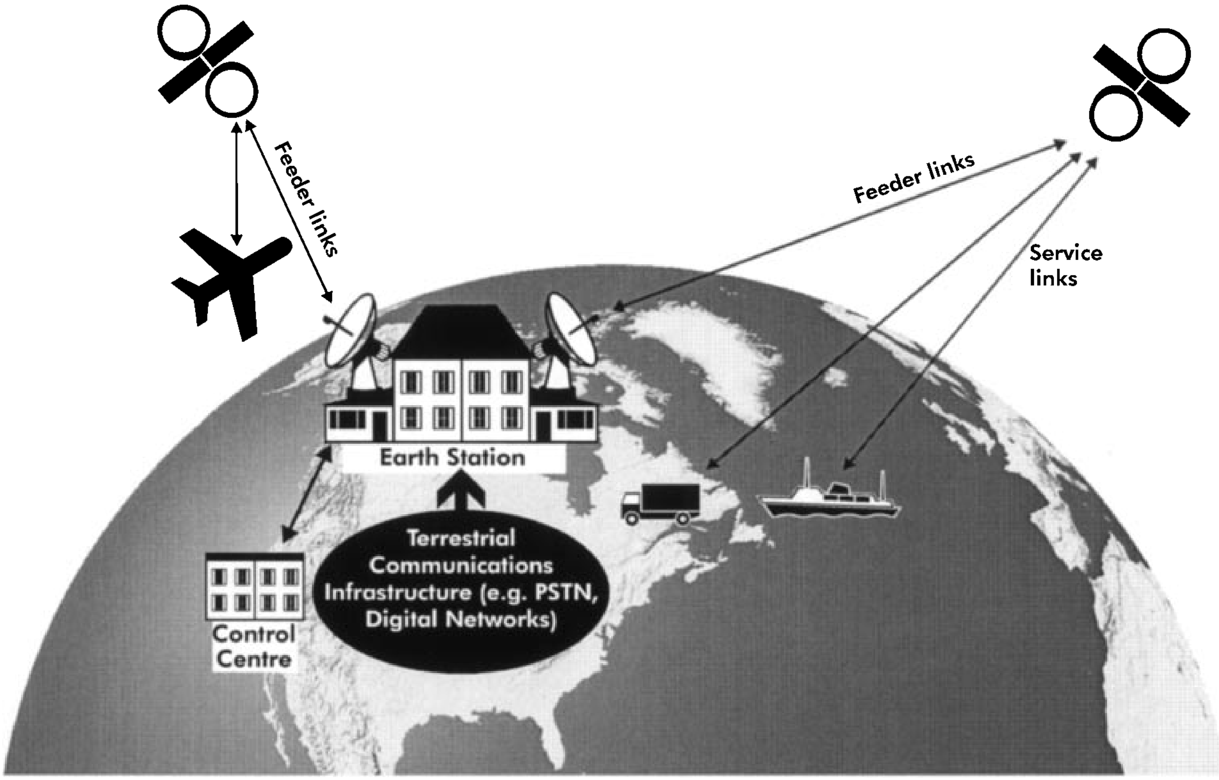

Figure 1. The major components of a mobile satellite system. Lines terminated with arrowheads indicate communications links.

are coordinated by a control center in a way that shares the satellite transmission resources efficiently. Also, the control center may issue commands to the satellites via the earth stations.

A number of radio links are required for such a system. Communication from the earth station to the mobile terminal is said to be in the forward direction, whereas communication from the mobile terminal to the earth station is said to be in the return direction. In both the forward and return directions, an up-link to the satellite and a down-link from the satellite are required, for a total of four radio links. The links between the earth station and the satellite are sometimes referred to as feeder links, whereas the links between the mobile terminal and the satellite are typically referred to as service links or mobile links. In some of the more advanced satellite systems with multiple satellites, there are radio links between adjacent satellites called intersatellite links.

A wide variety of services and applications are being supported by mobile satellite systems, with many more being proposed. First- and second-generation systems are limited to data rates ranging from a few hundred bits per second (bps) to several tens of kilobits per second (kbps) and have concentrated their efforts on providing services that fall within categories such as telephone-quality speech, packet data communications, facsimile, generic asynchronous stream data, and paging. Third-generation systems are expected to be capable of transmission at rates up to several hundred kilobits per second and will be capable of delivering moderate-quality video and high-quality audio services. Increasingly, the services delivered by these systems will appear to be an extension of those available to users over the converging terrestrial systems.

Many satellites isolate selected frequency bands from the composite up-link signal using filtering, translate these selected bands to their down-link frequency band, amplify them, and then transmit them toward the earth in the appropriate antenna beam. The term *transparent* satellite is used in this case. As an extension of this concept, some of the newer satellites use digital processing to select the up-link signal in a given frequency band, time slot, and antenna beam, and then "switch" it to the desired down-link frequency band, time slot, and antenna beam. The most sophisticated satellites demodulate the up-link transmissions and then process the resulting data signals in the same manner as a digital switch prior to modulation for down-link transmission. This type of satellite is sometimes referred to as a regenerative satellite.

A wide variety of mobile satellite terminals is commercially available. Here, we give only a few examples. Figure 2 shows a receive-only unit, manufactured by Skywave Mobile Communications Inc., that can be used to receive alphanumeric messages sent to a personal computer-based terminal, over the Inmarsat-D system. This system is a high-penetration system and can receive messages even when moderate blockage of the satellite signal is occurring. The receiver is the small black rectangular object beside the laptop computer. The white disk-shaped object is the antenna, which has a magnetic base allowing it to be temporarily mounted on the roof of a vehicle. At other times, any flat surface will suffice.

A Mitsubishi MSAT telephone transceiver, mounted on the front wall of the trunk of a car, is shown in Fig. 3. The corre-

Figure 2. A receiver and antenna for the Inmarsat-D high-penetration messaging system. The receiver is shown connected to a laptop computer. Reprinted with permission from Skywave Mobile Communications, Inc.

sponding antenna subsystem, mounted on the car's roof, is shown in Fig. 4. A third subsystem, which is not shown, is the users interface unit in the passenger compartment, including the telephone handset.

The major subsystems of the CAL Corporation's satellite telephone terminal, for telephone communications to aircraft via MSAT, are shown in Fig. 5. Most of the terminal's electronics are contained in the black box on the right-hand side. This box would normally be mounted inside the pressurized cabin of the aircraft. The antenna subsystem is shown on the left-hand side, with its radome placed behind it. For this particular antenna, two short helices are used as the transducing

Figure 3. A Mitsubishi MSAT telephone transceiver mounted on the front wall of the trunk of a car.

Figure 4. The antenna for a Mitsubishi MSAT telephone terminal, mounted on the roof of a car.

elements in order to achieve the required amount of antenna gain while keeping the profile of the antenna low. The antenna is often mounted on the top of the fuselage, as is shown in Fig. 6. However, on some aircraft the top of the tail fin is a preferred location for antenna mounting.

MOBILE SATELLITE LINKS

We will start by considering the path of the radio signal as it travels from the satellite to the mobile terminal, that is the down-link in the forward direction. The detailed discussion of

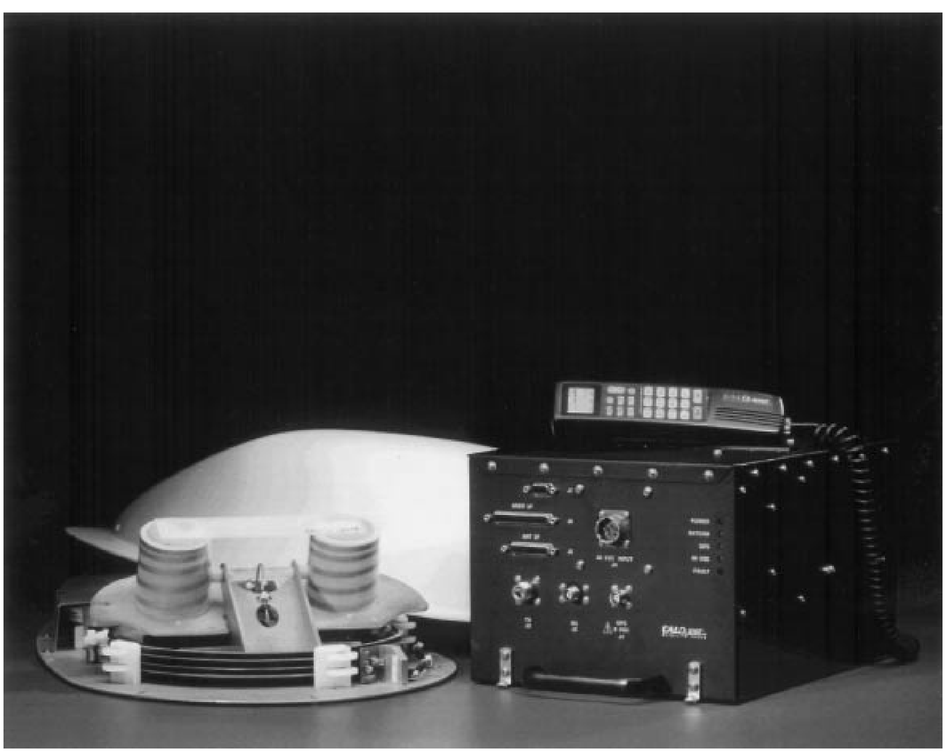

Figure 5. The major subsystems of the satellite telephone terminal, intended for use by aircraft with the MSAT system. Reprinted with permission from CAL Corporation.

Figure 6. A Cessna Citation jet aircraft, operated by the Ontario Air Ambulance Service, equipped with a mobile satellite communications terminal. The antenna subsystem can be seen mounted on the top of the fuselage.

this link will introduce the concepts necessary to understand more concise discussions pertaining to the other links of interest in a mobile satellite system. Radio frequency bandwidth and electrical power are two scarce resources that tend to constrain the design of mobile satellite systems. In this section, the focus is primarily on power, with efficient bandwidth utilization being partially addressed in subsequent sections. Clearly, down-link power will be limited because most satellites use solar power as their primary source of electrical power. Also, up-link power from the mobile terminal tends to be limited because such a terminal receives electrical power from either its own battery or that of the vehicle.

Line-of-Sight Transmission

At the satellite, the signal is amplified so that its average signal power is P_t dBW, at the input to the transmitting antenna. It is the transmitting antenna's function to spread that signal power as uniformly as possible over the desired coverage area on the Earth's surface, while wasting as little power as possible outside this coverage area. This is directly analogous to the ability of the reflecting surface of a flashlight to focus the light from the bulb into a beam of light. A measure of the ability of the antenna to focus the radiation is its gain, which is the ratio of the flux density at the center of the coverage area to that value that would occur if the power had been radiated equally in all directions (i.e., isotropic radiation). This gain is a function of the size of the antenna, and for a circular parabolic antenna it is given by

$$G = 10\log_{10}(\Omega\pi^2 D^2/\lambda^2)\,\text{dBi} \tag{1}$$

where Ω is the efficiency of the antenna (typically between 50% and 70%), D is the diameter in meters, and λ is the wavelength of the radio frequency signal in meters. Other types of antennas will have differing gains, but Eq. (1) provides an order of magnitude estimate of the required antenna size to achieve a prescribed gain. This discussion assumes that a single beam is used to cover the desired area. For reasons that will be discussed later, it may be advantageous to cover the desired area with multiple overlapping beams, but using the same antenna superstructure. An example of one way to achieve this is to use a single large reflector with multiple feeds (i.e., source transducers) in different locations near the focal point of the reflector. Of course, increasing the number of beams increases the complexity of the satellite. The size and weight of the satellite's antennas is constrained by the need to maintain reasonable costs for the satellite and its launch. Nevertheless, advanced technology allows for surprisingly large antennas to be deployed in space. For example, the North American MSAT satellites have two elliptical antennas, measuring 6 m by 5 m, and provide five beams covering all continental North America, the Caribbean Sea, and Hawaii. Some later systems have significantly larger antennas and can support more than 100 beams.

As the signal travels from the satellite to the earth, its flux density decreases as the square of the distance traveled. This power loss is referred to as the free space path loss and is given by

$$L_\text{p} = 10\log_{10}[(4\pi d)^2/\lambda^2]\,\text{dB} \tag{2}$$

where d is the distance traveled between the satellite and the mobile terminal. A geostationary orbit is a circular orbit for which the orbital radius, position, and velocity are such that the satellite remains in approximately the same location above the equator as the earth rotates. For a geostationary orbit, like that of MSAT, the radius is about 42,163 km resulting in a typical propagation delay of greater than an eighth of a second to traverse from the satellite to the surface of the earth. At MSAT frequencies, the corresponding path loss is about 188 dB! The great altitude of a geostationary satellite allows it to view about a third of the surface of the earth. Consequently, global coverage (with the exception of the polar regions) is possible with only three satellites. A larger number of satellites, in circular orbits at lower altitudes, can be used to provide global service with the advantages of lower path loss, shorter propagation delay, and cheaper launch costs on a per satellite basis. For reasons of satellite longevity, altitudes that avoid the Van Allan radiation belts are usually selected. The low earth orbits (LEO) are located beneath the primary belt and have altitudes between 500 km and 2000 km. Similarly, the medium Earth orbits (MEO) are located between the primary and secondary belts and have altitudes between 9000 km and 14,000 km. The medium earth orbits are sometimes referred to as intermediate circular orbits (ICO). Unlike systems that use geostationary orbits, these other systems typically use several distinct orbital planes, each of which is inclined with respect to the equator. A number of proposed systems have planned to use highly elliptical orbits (HEO) instead of circular ones. The potential advantage of a HEO-based system is that it can provide high angle-of-elevation coverage to selected areas in the temperate zones (i.e., those parts of the world for which the demand for communications services is the greatest) with a moderate number of satellites. Despite this advantage, it does not appear that HEO systems will play a significant role in mobile satellite communications.

Upon reaching the terminal, the signal energy is collected by the receiving antenna and is converted by a transducer to an electrical signal. A typical example of a mobile satellite antenna designed for the MSAT system is shown in Fig. 7. Here, the transducing element is a short helical structure, similar to the element used for the land mobile satellite terminal and to each of the two elements for the aircraft mobile satellite terminal shown previously in this article. The white dome is a radome that is placed over the antenna to protect it. This antenna must be steered in azimuth but has a wide enough beam that steering in elevation is not necessary. Many mobile terminals use closed-loop antenna steering mechanisms, based upon the received signal strength. The gray box shown in Fig. 7 contains a self-calibrating electronic compass that can be used to improve the antenna steering achievable using signal strength alone. Because of the fact that the physical rules describing the propagation of transmitting and receiving display a reciprocal relationship, an appropriate measure of the antenna's ability to collect the energy is the antenna gain, as described in the text near Eq. (1). Therefore, the average power of the received signal from the line-of-sight propagation path at the output of the receiving antenna is given by

$$P_r = 10\log_{10} C = P_t + G_t - L_p + G_r \,\text{dBW} \tag{3}$$

where G_t is the gain of the transmitting antenna and G_r is the gain of the receiving antenna.

The signal's radio frequency plays a major role in Eq. (3), with G_t, L_p, and G_r increasing with the square of the frequency. The net result is that the received power also in-

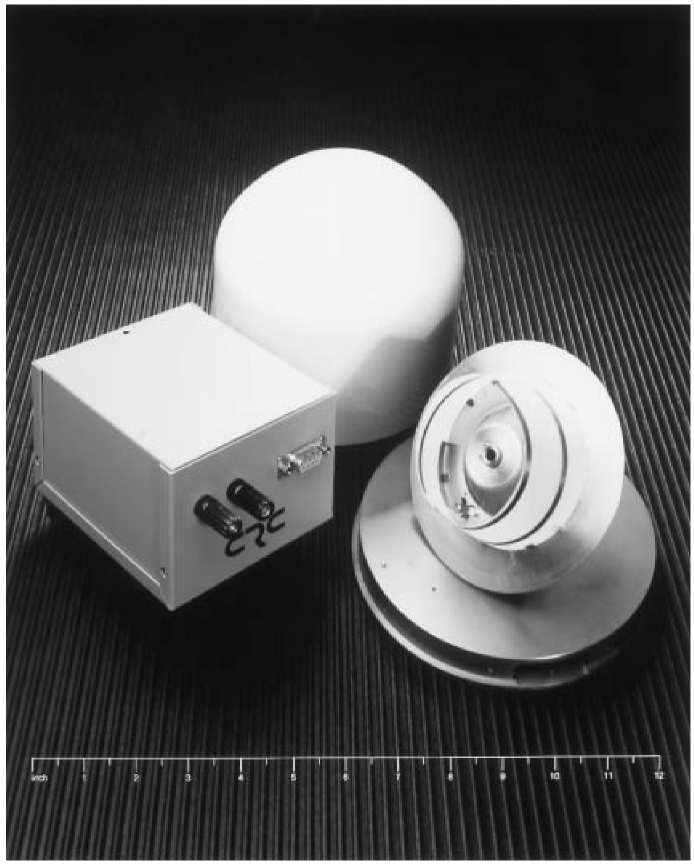

Figure 7. A prototype antenna system designed for the North American MSAT system. On the right-hand side of the foreground is the short helical antenna element. It can be steered in azimuth but is fixed in elevation. The box on the left-hand side is the antenna steering unit, and the radome is shown in the background.

creases with the square of the frequency. Alternatively, if the received power is treated as the fixed parameter, smaller antennas could be used at higher frequencies. Some of this benefit for higher frequencies is offset by other propagation effects. For example, the lower frequencies (i.e., longer wavelengths) are more robust in the presence of blockage by collections of small obstacles such as foliage and rain. A second factor that is very important is the availability of an otherwise unused radio spectrum. At the international level, spectrum usage is determined by the International Telecommunications Union (ITU) at an on-going series of World Administrative Radio Conferences (WARC). Then national bodies, such as the Federal Communications Commission (FCC) in the United States, license specific service providers to offer the corresponding services within each country. A wide variety of frequency bands have been allocated for mobile satellite systems, typically with the larger allocations being at the higher carrier frequencies as a result of availability. Consequently, the systems that offer low data rate services are generally allocated lower frequency bands than those offering high data rate services. For example, a number of systems offering low-rate store-and-forward messaging services communicate between the satellite and the mobile terminal in frequency bands between 100 MHz and 400 MHz, although most of the systems offering medium-rate mobile satellite telephone services use bands between 1.5 GHz and 2.5 GHz, and many of the proposed systems for providing high-rate multimedia services plan to operate in bands between 20 GHz and 30 GHz.

Because of the large path loss that is typical of satellite transmissions, the received power is very low. In fact, it is so low that the thermal noise in the receiving antenna and front end of the receiver must be accounted for. The resulting carrier-to-noise-spectral-density ratio is given by

$$10\log_{10}(C/N_0) = P_r - T_r - k \text{ dB-Hz} \quad (4)$$

where T_r is the composite noise temperature of the receiver expressed (dBK) and k is Boltzmann's constant (−228.6 dBW/K-Hz). If the transmission is digital with a rate of R bps, the energy-per-bit-to-noise-spectral-density ratio is given by

$$E_b/N_0 = C/(N_0 \cdot R) \quad (5)$$

Of course, thermal noise is not the only impairment that needs to be considered. Some of the other common impairments that are encountered by mobile satellite transmissions will be addressed in the following sections.

Multipath Propagation and Shadowing

In addition to the line-of-sight path, the signal can reach the receiving antenna by reflected paths from objects that are usually located nearby. Often, several distinct reflecting objects are in the field of view of the receiving antenna. If the differences in the propagation times for the various propagation paths (reflected and line-of-sight) are much less than the reciprocal of the bandwidth of the transmitted signal, the effect of the multipath propagation can be viewed as non-time-dispersive. This type of multipath propagation will affect the power and carrier phase of the received signal according to the nature of the superposition of the paths, but it will not distort its frequency content or introduce intersymbol interference in the case of a digital transmission. For land mobile satellite applications, measurements (1) taken in a frequency band near 1.8 GHz indicate that the difference in propagation times rarely exceeds 600 ns. Consequently, for signal bandwidths up to several hundred kilohertz, the multipath propagation can be considered non-time-dispersive. The following discussion is based on this assumption being valid. If the geometry of the paths change with time as a result of terminal motion, satellite motion, or motion of the reflecting objects, the power and carrier phase of the received signal will vary with time. This time-varying phenomenon is referred to as fading, or more specifically as flat fading for the non-time-dispersive case.

For the purpose of evaluating the performance of candidate transmission techniques, it is frequently desirable to model the propagation environment in a way that is suitable for numerical analysis and simulation. An approximation that is often made is to assume that the reflecting objects are adequately numerous and independent in nature for the central limit theorem to apply. Consequently the fading can be represented by a Gaussian process that is completely statistically characterized by its power spectral density. The power spectral density will be nonzero only over a bandwidth equal to the difference in frequency between the path with the greatest Doppler frequency shift and that with the least (2). This type of fading model is referred to as Rayleigh fading. The combination of the line-of-sight path with the Rayleigh fading reflected path is referred to as Rician fading, which has the additional parameter called the carrier-to-multipath ratio (C/M), defined to be the ratio of the average signal power received over the line-of-sight path to that received over the reflected paths.

Another effect that can greatly affect the availability and performance of a mobile communications link is shadowing, the term given to blockage of the line-of-sight path. Such blockage occurs naturally in terrestrial mobile satellite environments as the moving vehicle passes by obstacles such as buildings, trees, and bridges. Many obstacles result in such severe attenuation of the line-of-sight signal that it is weaker than the reflected paths and can be ignored. A useful but simple model for shadowing is to switch between a good state (unshadowed) and a bad state (shadowed) with the typical time period for enduring each state being determined by the parameters of a two-state Markov model (3,4). A transmission model corresponding to this discussion is shown in Fig. 8. A simple shadowing model is to apply a fixed attenuation selected on a shadowing event by shadowing event basis, using a lognormal distribution. If the shadowing is predominantly caused by foliage, the line-of-sight path may also be included with it being subjected to attenuation according to another lognormal distribution (5). Of course the values selected for the model's parameters depend upon many issues, including angle of elevation to the satellite, type of terminal (e.g., land mobile, aircraft, marine, hand-held), antenna gain pattern, vehicular velocity, satellite velocity, environment (e.g. urban, suburban, highway), and terrain.

Other Sources of Degradation

Degradation to the received signal caused by thermal noise, multipath propagation, and shadowing have already been discussed. In many systems, these are the dominant sources of degradation, but there are a number of other ones that should be appreciated. Perhaps the next most important source of degradation is interference to the desired signal from other signals within the same system. If the interference is cause by another signal that is located in the same frequency channel as the desired signal, the interference is referred to as co-channel interference. For narrowband signals, co-channel interference is generally caused by interferers in other antenna beams for which the out-of-beam attenuation provided by the satellite antenna is not sufficiently great to render the interfering signal negligible. For spread spectrum signals, some of the co-channel interference may be due to other signals within the same beam. Interference to the desired signal can occur from signals in the adjacent frequency channels as a result of the fact that some of their transmitted energy falls outside of their allotted frequency channel. This type of interference is known as adjacent-channel interference.

For some mobile satellite systems, the ratio of the carrier frequency to the bit rate is many orders of magnitude. When this is the case, a nonnegligible amount of degradation can occur because the phase of the radio frequency carrier differs significantly from its ideal value in a time-varying nature, which is the result of the electronic components in the system. This phenomenon is called phase noise. Common sources of phase noise include imperfect oscillators and frequency syn-

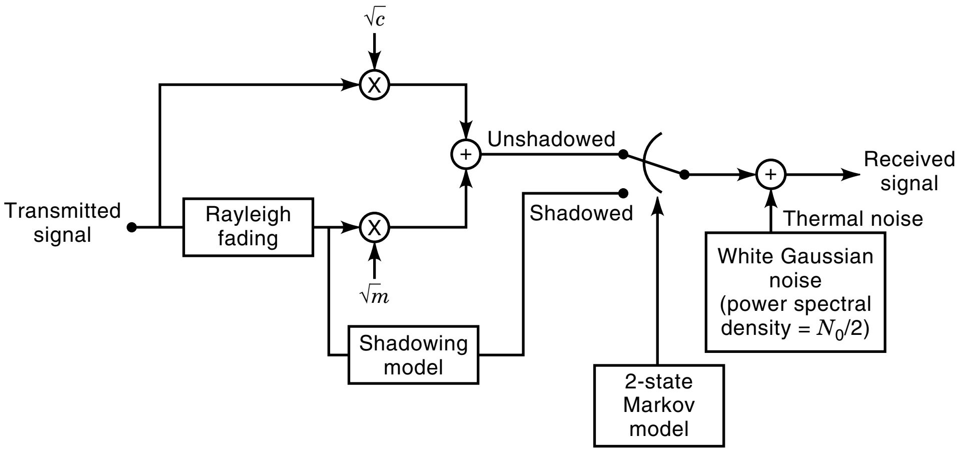

Figure 8. A useful model of fading and shadowing for the evaluation of mobile satellite transmission schemes. Here, c is the average power for the line-of-sight path and m is the average power for the reflected paths.

thesizers, vibration of the mobile terminal's electronic circuitry (known as microphonics), and electronic steering of the mobile terminal's phased array antenna.

Nonlinear power amplification, at several locations in a mobile satellite system, can cause degradation. In the case of the transmitting power amplifier in the mobile terminal, typically only a single carrier (signal) is present, and the distortion of that signal by the amplifier's nonlinear behavior has two effects. First, there will be a small reduction in the power efficiency of the desired transmission. For example, if the signal is digital, a little more transmit power will be necessary to achieve the required bit-error rate. Second, the distortion will often broaden the power spectrum of the transmitted signal resulting in increased interference in the adjacent channels.

Nonlinear distortion will also occur in the transmit power amplifiers of the earth stations and the satellites. Usually, there will be many carriers being amplified simultaneously. In this case, the result is a broadband noiselike signal caused by the intermodulation of the many carriers present in the amplifier.

Depending on the frequency band used by the given mobile satellite system, it may be necessary to account for effects such as ionospheric scintillation, tropospheric scintillation, gaseous absorption, and rain attenuation. In general, these effects become more severe for lower angles of elevation.

THE SIGNAL-PROCESSING PATH

In this section we discuss some of the signal-processing techniques that can be used to increase the efficiency with which the scarce resources of radio frequency spectrum and electrical power are used. Figure 9 shows a high-level block diagram of the processing stages for the transmitting side of the communications chain. The inverse operations are performed on the receiving side to recover the transmitted information. Here, we will discuss the blocks in this processing chain only to the level necessary for understanding their role in a mobile satellite context. More detailed treatment of many of these processing stages can be found elsewhere in this encyclopedia.

The first block in the chain is the Information Source. Examples include telephone-quality speech, data representing text, and multimedia signals representing a composite of audio, video, and data components. Regardless of the type of information that is to be transmitted, it is important to minimize the number of bits required to represent the information subject to constraints such as delay, processing complexity, and quality of the representation. This is the objective of the second block in the chain, entitled "Source coding." Using telephone-quality speech as an example, the analog waveform can be accurately represented using a 64 kbps stream of data, by sampling the waveform at 8 ksamples/s and giving each sample 8 bits of precision. However, using recently developed speech-coding standardized techniques, the bit rate can be reduced a full order of magnitude to 6.4 kbps without a significant reduction in speech quality (6). Very efficient standardized low-rate video-coding techniques also exist (7). Of course, the same techniques as are used for computer storage can be used to reduce the size of data and text files for mobile satellite transmission.

Error Control Coding

Error control coding introduces redundancy into the bit stream by increasing the total number of bits in such a way that each original bit influences several bits in the error-control-coded bit stream. This redundancy can then be used to correct (forward error correction coding) or detect (error detection coding) transmission errors at the receiver. We will consider forward error correction first. Even though the additional bits do result in an increase in the required number of

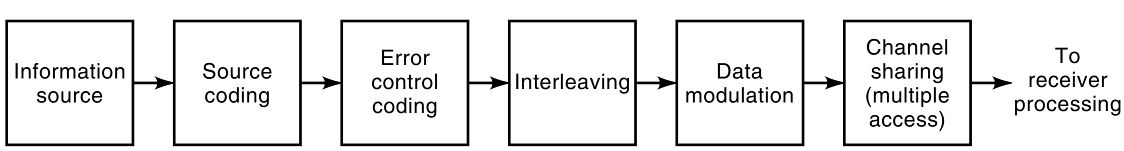

Figure 9. A high-level block diagram of the processing stages for the transmitting side of the communications chain.

bits to be transmitted, appropriate coding and decoding schemes will generally result in a net reduction in the transmitted power required to meet a given bit-error rate. For first-generation mobile satellite systems, rate-1/2 constraint-length-7 convolutional coding has been a fairly standard choice. Note that the rate is the ratio of the number of bits into the coder to those out of the coder. In some cases, punctured versions of this code have been used to achieve a higher coding rate, thereby improving bandwidth efficiency at the expense of power efficiency. A predominant reason for the popularity of this code is that it was one of the first fairly powerful error correction codes for which decoder integrated circuits, capable of processing soft decisions, were commercially available. For decoding in fast fading and shadowing conditions, the soft decision should incorporate channel state information so that the decoder assigns relatively less importance to bits that were received when the signal was faded or blocked.

The achievable coding gain is a strong function of the block length over which coding is performed, with larger blocks allowing for greater gains. For applications for which the packet or frame length is quite short (e.g., most packet data and low-rate speech applications) convolutional coding is still a good choice although constraint lengths greater than 7 can be implemented now. Tail biting (i.e., encoding the input data in a circular buffer) can be performed to eliminate the overhead of transmitting extra bits to terminate the code's decoding trellis (8).

For applications for which the frame length is longer than a couple of hundred bits, turbo coding will be a strong candidate for future systems (9). The performance of turbo coding improves as the block length increases. However, the end-to-end delay of the transmission system increases with increasing block lengths. Consequently, only services that are tolerant of fairly large delays can benefit from the most power efficient error control coding. For rate-1/2 coding, Fig. 10 shows the performance for constraint-length-9 convolutional coding (80-bit block) and turbo coding (512-bit block and

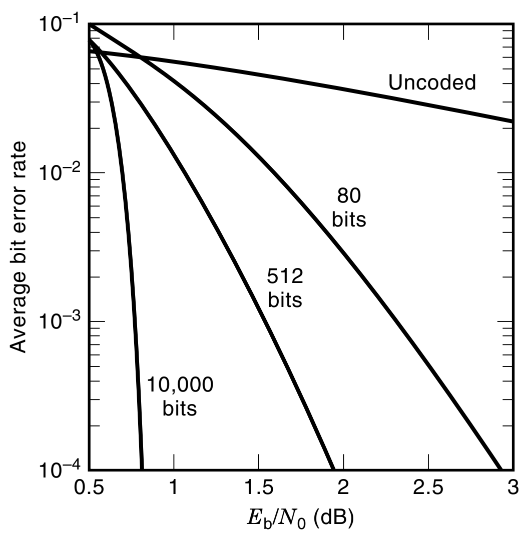

Figure 10. The performance of various rate-1/2 codes in an additive white Gaussian noise environment. Shown are simulation results for a constraint-length-9 convolutional code with tail biting and a block size of 80 bits, and turbo coding with block sizes of 512 bits and 10,000 bits.

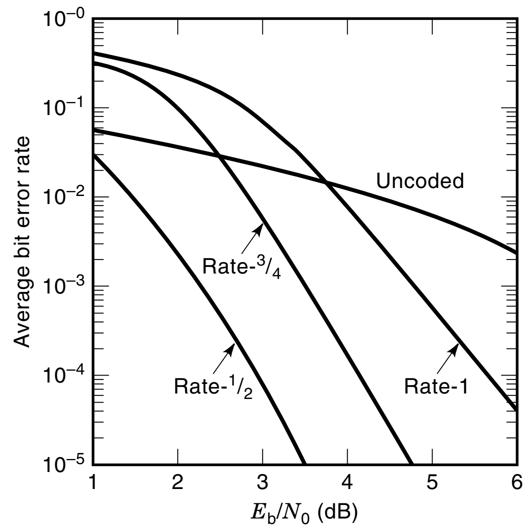

Figure 11. The performance of codes of differing rates in an additive white Gaussian noise environment. Shown are simulation results for the constraint-length 9 rate-1/2 code; the rate-3/4 code, which is a punctured version of the rate-1/2 code; and the rate-1 code, which is a pragmatic trellis-coded modulation with the rate-1/2 code being mapped into a 4-level constellation.

10,000-bit block). This turbo code uses 16-state recursive systematic convolutional codes as its component codes. These performance results assume antipodal signaling (e.g., ideal coherent binary phase-shift keying) with the only channel impairment being Gaussian noise.

In general, the benefit that can be achieved by error correction coding increases with increasing decoding complexity and block (i.e., code word) size, and with decreasing code rate. One way to achieve a higher code rate, for a fixed decoding complexity, is to use puncturing (10). Puncturing increases the code rate by selectively deleting some of the coded bits prior to transmission. In order to increase the code rate beyond 1 bit per symbol, it is necessary for the coder to map the input sequence of bits into a sequence of symbols for which the size of the symbol alphabet is greater than 2. A well-known technique for doing this is trellis-coded modulation (11,12). Some forms of trellis-coded modulations are designed in such a way that standard convolutional decoder integrated circuits can be used to perform the decoding. These forms are referred to as pragmatic trellis-coded modulations (13). An example of the trade-off between power and bandwidth efficiency can be seen in Fig. 11. Here, all three codes are based upon the same convolutional code, with the rate-1/2 code being a constraint-length-9 code, the rate-3/4 code being a punctured version of the rate-1/2 code, and the rate-1 code being a pragmatic trellis-coded modulation with the rate-1/2 code being mapped into a 4-level constellation.

Error detection coding is useful for services that are message or frame based, and it is important to know whether a given message or frame has been received correctly. In these cases a small field of parity bits (e.g., 16 parity bits) is appended to the message, with the parity bits being generated using a cyclic redundancy code. At the receiver, if the parity bits computed from the received data bits do not agree with the received parity bits, the message is known to be in error.

In some systems, a request will then be sent to the transmitter to retransmit the message.

Returning to error correction coding, many forward error correction codes are much better suited to correcting randomly distributed single errors than long bursts of errors, assuming that the average bit error rate is fixed. However, some impairments such as multipath fading cause error patterns that are bursty in nature. To the extent allowed by constraints such as message length and delay restrictions for the service, interleaving can be used between the coder and the modulator in an attempt to eliminate error bursts prior to decoding. Interleaving permutes the order of the coded symbols according to a rule that is known at both the transmitter and the receiver. After demodulation at the receiver, the deinterleaver performs the inverse permutation prior to passing the soft decisions to the decoder. By so doing, sequences of soft decisions corresponding to poor bursts of signal are broken up and mixed with soft decisions that were received under more favorable conditions.

Modulation

After interleaving, the sequence of coded symbols is modulated. Here, we restrict our consideration to linear modulation schemes. For a linear modulation scheme, the transmitted signal is given by

$$\begin{aligned} s(t) &= \mathrm{Re}\left\{\left[\sum_{i=0}^{N-1} a_i g(t-iT)\right] e^{j\omega_0 t}\right\} \\ &= \left[\sum_{i=0}^{N-1} g(t-iT)\mathrm{Re}(a_i)\right]\cos(\omega_0 t) \\ &\quad - \left[\sum_{i=0}^{N-1} g(t-iT)\mathrm{Im}(a_i)\right]\sin(\omega_0 t) \end{aligned} \tag{6}$$

where a_i; $i = 0, \ldots, N-1$ is the sequence of complex modulation symbols, T is the symbol period, $g(t)$ is the unit pulse response of the pulse-shaping filter and is assumed to be real, and ω_0 is the radian carrier frequency. In the second line of Eq. (6), the term inside the square brackets prior to "cos" is referred to as the in-phase component of the signal and the term inside the square brackets prior to "sin" is referred to as the quadrature component of the signal. For M-ary signaling, each a_i is selected from an alphabet of M complex numbers, with the modulus of each complex number representing the amplitude of the given symbol and the phase of each complex number representing the phase of the given symbol. The majority of mobile satellite communications systems uses one or more forms of phase modulation. In the case of phase modulation, each a_i is selected from a symbol alphabet for which all elements have a modulus of one. Therefore, only the phase of the symbol varies. Binary phase shift keying (BPSK) is popular for low rate systems because of its robustness. For BPSK, each a_i is select from the alphabet $\{1, -1\}$ which is purely real, and consequently a BPSK waveform has no quadrature component. A variation of BPSK, that is used in aeronautical satellite communications, is $\pi/2$-BPSK for which subsequent symbols experience a relative phase shift of $\pi/2$ radians. For example, each a_i is select from the alphabet $\{1, -1\}$ when i is even and from $\{j, -j\}$ when i is odd. When used with an appropriate choice of pulse-shaping filter, such as a 40% square-root raised-cosine filter, the result is a waveform that suffers less spectral spreading when passed through a nonlinear amplifier, but enjoys all the robustness of standard BPSK. For systems requiring some additional spectral efficiency, some form of quadrature phase shift keying (QPSK) is usually selected. Standard QPSK can be thought of as two BPSK signals being transmitted in parallel; one as the in-phase component and the other as the quadrature component. A variation of QPSK that is of some interest is $\pi/4$-QPSK, for which subsequent symbols experience a relative phase shift of $\pi/4$ radians. The advantages of selecting $\pi/4$-QPSK are similar to those described previously for $\pi/2$-BPSK. Another variation of QPSK that is even more robust to nonlinear amplification is offset-QPSK for which the symbol timing for the in-phase component is offset by half a symbol period relative to that of the quadrature component.

Multiple Access

Next we consider how the satellite resources of bandwidth and power can be efficiently shared between many users. The sharing of the transmission medium between several users is referred to as multiple access (see MULTIPLE ACCESS MOBILE COMMUNICATIONS). We start from a highly idealized point of view, considering the case where there is only a single beam, perfect synchronization in both time and frequency have been achieved, and no interference is permitted between users.

First, let power be the only constraint. Each user can have as much bandwidth as he wishes but cannot exceed some fixed maximum value of transmit power. Under this constraint, each user attempts to maximize his throughput (i.e., bit rate) subject to the requirement that the average bit error rate is better than some specified value. In general, lowering the coding rate allows for greater power efficiency and consequently a higher throughput for a given amount of power. The achievable region is illustrated by the area under the curve labeled "Power Constraint" in Fig. 12. Note that the coding rate is expressed in bits per dimension, which takes into account the modulation and error control coding. This is the ratio of the number of bits into the error correction coder to the number of dimensions out of the modulator, over a fixed period of time. In Eq. (6), Re$\{a_i\}$ and Im$\{a_i\}$ can be considered as examples of dimensions in the signal space. It is well

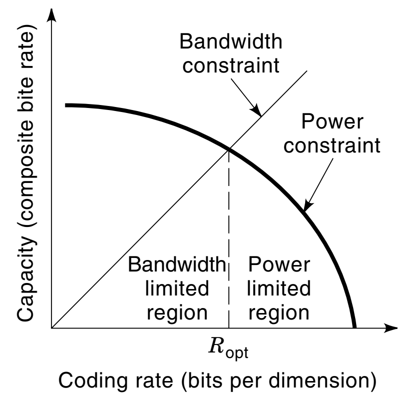

Figure 12. The tradeoff-between capacity and coding rate subject to a power constraint and a bandwidth constraint. R_{opt} is the coding rate that maximizes the capacity.

known that for a bandwidth of B and a time duration of T_s the number of available dimensions is $2BT_s$ (14).

Now let bandwidth be the only constraint being considered. Clearly, the composite bit rate will increase linearly as the users increase their coding rate. The achievable region is illustrated by the area under the curve labeled "Bandwidth constraint" in Fig. 12. If both the power and bandwidth constraints are taken into account, there is an optimal code rate (assuming block size and decoding complexity are fixed) R_{opt} that maximizes the throughput of the system. If the system is operating at a lower rate, it is said to be bandwidth limited, and if it is operating at a higher rate, it is said to be power limited. With most of the early mobile satellite systems, the satellites were comparably weak, the demand for spectrum was low, and few devices were available to support coding rates below rate-1/2. Consequently, most early systems were operating in the power-limited region. With newer systems, much more emphasis is being placed on achieving nearly optimum capacity in the system design.

One example of a set of dimensions (i.e., a basis) for the signal space is the time sample representation of the composite signal, with sampling being performed at the Nyquist rate. If sequential groups of these time samples are apportioned between the users, the sharing arrangement is called time division multiple access (TDMA). Here, the mobile terminals must be fairly accurately synchronized in time so that bursts arriving at the satellite from different terminals can be tightly packed without interfering with each other. Typically, the required timing accuracy is achieved when the terminal requests to initiate communication by sending a short burst on a random access channel, for which accurate timing is not necessary. Then along with an assignment of a set of time slots, the system sends the terminal an accurate clock correction that was calculated by the Earth station based upon the measured time-of-arrival of the burst. Of course many other potentially useful bases exist. If nonoverlapping portions of the total bandwidth are apportioned between the users the arrangement is called frequency division multiple access (FDMA). In this case, timing accuracy is no longer important but narrower band filtering is necessary and the lower data rates present on each carrier tend to make the system more susceptible to phase noise. If orthogonal codes are used to form the basis of the signal space, the sharing is called code division multiple access (CDMA). In a synchronous CDMA system, the carriers must be synchronized in time to within a small fraction of a chip period so that orthogonality is maintained. In the forward direction, this is fairly straightforward to achieve if all the signals are originating from a single Earth station. In an asynchronous CDMA system, time synchronization is not required with the result that the signals are no longer truly orthogonal, resulting in some interference. In the return direction, achieving sufficiently accurate time synchronization amongst all of the mobile terminals is quite challenging so asynchronous CDMA could be preferred over synchronous CDMA. Of course, combinations of these approaches are possible. Most of the mobile satellite systems to date have used FDMA. However, systems based upon narrowband TDMA, which is a combination of FDMA and TDMA, are beginning to appear, even though CDMA is a strong candidate for systems with many beams or where there are severe power spectral density limitations.

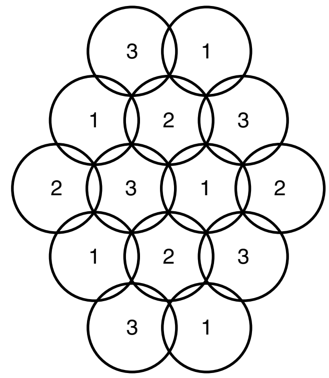

Figure 13. Total coverage area being covered by multiple beams. In this case, there are 14 beams and the total frequency band is subdivided into three subbands. Two of the subbands are used in five beams, whereas the remaining subband is used in four. The resulting frequency reuse factor is 4.667.

More efficient use of both bandwidth and power can be achieved if the satellite's antenna system covers the desired area of the Earth's surface with several smaller beams instead of one large one. The power efficiency of the link is improved as a result of the higher antenna gain associated with the smaller beams. With respect to frequency, the total allocated system bandwidth is divided into a number of distinct subbands, which need not be of equal bandwidth. As illustrated in Fig. 13, each beam is assigned a subband in such a manner that some desired minimum distance between beams with the same subband is maintained. The frequency reuse factor is the ratio of the number of beams to the number of distinct frequency subbands.

For most of the preceding discussion, it was assumed that no interference between users is permitted. In reality, some interference is unavoidable and may even be desirable to decrease system complexity and possibly to improve system capacity. For example, some CDMA systems allow each transmission to be completely asynchronous in chip timing and carrier phase relative to that of other users occupying the same frequency band and period in time. In this case, each transmission appears to be low-level broad band noise to the other users. Unlike the FDMA and TDMA systems for which the interference tends to be dominated by a small number of dominant interferers, the interference experienced by a user is the result of a very large number of other users resulting in a level of interference that is much less variable. Full statistical advantage can be take of voice activation without the need for sophisticated dynamic channel assignment strategies. Powerful error correction coding allows for high levels of both intra- and interbeam interference. This results in the ability to reuse the same frequency bands in every beam and a corresponding high level of capacity in a multibeam satellite system (15).

Because interference is unavoidable, interference mitigation techniques are of interest. One example of such a technique is power control, for which the power of each user terminal is dynamically adjusted with the goal of providing it with just enough power to meet the required grade of service. Allowing terminals additional power would only serve to exacerbate the interference levels experienced in the system. A second example is the use of multiuser detection schemes (16).

PRESENT AND PLANNED SYSTEMS

Here the intent is to provide some examples of systems that are presently offering mobile satellite communications services and of those that are planned for the future. The systems discussed represent only a sampling and not an exhaustive summary.

Global mobile satellite communications got its start in 1976 when 3 Marisat satellites where launched and positioned at approximately equal intervals in geostationary orbits. In 1979, Inmarsat was formed to offer global maritime satellite communications services. Inmarsat is a multinational organization that was created by the United Nations affiliated International Maritime Organization. Even though its original charter restricted its operation to maritime services, its charter was later extended to include aeronautical as well as land mobile and portable services. The nature of the Inmarsat organization continued to evolve with the goal of allowing it to offer an increasing array of mobile satellite services in a commercially competitive environment.

Inmarsat-A was the first system to offer commercial service on a global basis. Its terminals are relatively large and expensive, with the typical antenna being a 1-m diameter parabolic dish and a terminal weight of around 35 kg being representative. Consequently, the majority of the customers are large commercial users with most of the marine terminals installed on ocean-going ships and most of the portable terminals belonging to governments or news gathering organizations. Voice transmission was accomplished using analog frequency modulation, which is neither bandwidth nor power efficient by today's standards. Inmarsat has introduced several new voice and data systems that are based on more recent digital technologies. All Inmarsat's systems operate over geostationary satellites. The first of these new systems is the Inmarsat aeronautical system, which is based upon the work of the International Civil Aviation Organization and the Airlines Electronic Engineering Committee.

The purpose of the aeronautical system is to provide comprehensive aeronautical communications services, including basic air traffic services, aeronautical operational control, and cabin telephone. Inmarsat began by providing the cabin telephone service, with other services to be phased in later. This system is unique in that it is the only mobile satellite system that has been designed in a manner consistent with Open System Interconnect (OSI) principles.

The Inmarsat-M and -B systems were developed in parallel and share a common protocol. The M system offers lower-cost and reduced weight (typically about 10 kg) terminals, which provide communications-quality voice (4.2 kbps voice coding rate with the addition of error control coding bringing the rate up to 6.4 kbps), low-speed data (2.4 kbps), and facsimile services. In addition to marine and land mobile terminals, portable terminals the size of a small briefcase (including the antenna) are available. Telephone booths based on Inmarsat-M technology, that are powered using solar panels, are used in underdeveloped parts of the world.

Inmarsat-B is the designated successor to Inmarsat-A for providing high-quality professional communications services. For operation within the global beam of a satellite, the mobile antenna requirements for the A and B systems are identical, with a typical gain of 20 dBi. Inmarsat-M terminals have smaller antennas, with gains of 14 and 12 dBi for marine and land mobile terminals, respectively. Also available are still smaller "mini-M" terminals that operate only in the higher gain beams provided by the Inmarsat-3 series of satellites, launched in 1996 and 1997.

Inmarsat-C was introduced in 1990 to support store-and-forward packet data services such as telex, electronic mail, messaging, and position reporting. Even though only low-bit rates (600 bps) are supported, the terminals are small and inexpensive relative to those for the other Inmarsat sytems. An antenna with a gain as low as 1 dBi will suffice.

A number of regional systems offer terminals and services similar to those of the mini-M system. One example is the North American MSAT system, for which Canada and the United States each launched a geostationary satellite. A number of future regional systems are planned for Asia and the Middle East, using extremely large geostationary satellites, which should be capable of delivering these services to hand-held terminals, or higher data rate services to larger terminals.

These systems are alike in that they all use geostationary satellites, and the mobile terminals receive their signals in a band around 1,550 MHz and transmit their signals in a band around 1,650 MHz. Systems exist that use completely different frequency bands and in some cases orbits. We will begin with brief discussions of two systems that offer two-way messaging and position determination. These systems have targeted truck fleet management and cargo position reporting as primary application areas.

In 1990, the OmniTRACS system began full operation, providing two-way communications and position reporting services. It was licensed to operate on a secondary basis, which implies that it must not interfere with primary users, in the 12/14 GHz bands using existing geostationary satellites. The early start of service has allowed the OnmiTRACS system to build up a large customer base. A number of novel spread spectrum techniques are employed to safeguard against interfering with other systems.

The Orbcomm system plans to operate with a full constellation of 36 LEO satellites. The mobile terminals will receive their signals at about 138 MHz and transmit their signals at about 150 MHz. The system operators hope to achieve a competitive cost advantage by having small inexpensive satellites, low launch costs (as a result of the small satellites and low orbits), and lower terminal costs caused by the lower-frequency electronics.

A number of planned systems expect to offer hand-held telephone services on a global basis. Three systems that deserve particularly close attention are Globalstar, Iridium, and ICO. Globalstar and Iridium are LEO systems with 48 and 66 active satellites in a full constellation, respectively. The ICO system will use 10 active MEO satellites. The multiple access technique selected for ICO and Iridium is narrowband TDMA, whereas Globalstar will use CDMA. Iridium and Globalstar should be offering global services before the turn of the century, whereas ICO is expected to be a couple of years later.

Early in the next century, a number of satellite systems are planned to offer a broad range of services, including higher rate services which should effectively extend the digital network capabilities that will be available terrestrially. The highest profile of these is the Teledesic system. Originally, this system planned to use 840 LEO satellites! This has now been scaled back to a planned initial constellation of 288 LEO satellites.

For a number of reasons, position determination can be very important for a mobile satellite communications user. In fact, position determination is an integral part of many of the services such as vehicle fleet management and cargo tracking. Some terminals may use position information for antenna steering and to aid in the satellite and antenna beam handoff algorithms. Also, accurate position information is required for obtaining a license to offer service in some countries because the national authority insists on knowing if a call is being made within its territory. Some mobile satellite communications systems are capable of providing fairly coarse position estimation using the signals and satellites within the systems itself. However, accurate position determination is usually done by taking advantage of the Navstar Global Positioning System (GPS) (17).

The GPS system employs 24 satellites distributed in 6 orbital planes, each inclined by 55° with respect to the equator. These satellites are in 12 h medium earth orbits. Even though the system is financed by the US Department of Defense, it is used globally for both civilian and military applications. In addition to the signals generated aboard the Navstar satellites, the Inmarsat-3 satellites have transponders that can relay ground-generated GPS-type signals. These additional signals can be used to improve the accuracy and reliability of the position estimates. A GPS receiver estimates the range to several satellites and then uses these estimates to determine its position by triangulation. Range estimates to three satellites are sufficient to provide two-dimensional position (i.e., on the surface of the earth or if the altitude is known) plus accurate time, whereas four satellites are required to provide three-dimensional position plus accurate time. Each Navstar satellite transmits in two frequency bands; the L_1 carrier is centered at 1,575.42 MHz and the L_2 carrier is centered at 1,227.60 MHz. Frequency-dependent range estimates can be used to compensate for the effect of the ionosphere. The L_1 carrier is modulated with a short coarse/acquisition code (C/A code) at a chip rate of about 1 MHz and a longer precision code (P code) at a chip rate of about 10 MHz. The L_2 carrier is modulated with the P code only. The P code is dithered in a pseudorandom fashion so that the precision is limited for users other than those in the US military. In addition to the previously mentioned ranging codes, the carriers are modulated by a low-rate data stream carrying a navigation message that includes satellite position and satellite clock correction information. Typical civilian GPS receiver sets achieve a position accuracy of about 100 m and a time accuracy of about 10 ns. It is expected that the dithering of the P code will be eliminated within several years, allowing the accuracy for civilian sets to improve to better than 30 m.

TRENDS IN MOBILE SATELLITE SYSTEMS

Increasingly a broader range of services is being offered, with many of the new services requiring data rates that are higher than those currently available. Ultimately the services offered to mobile satellite users will be an extension of those that are available from terrestrial systems, with the result that mobile satellite service offerings will be pulled along by the expansion and convergence that is occurring terrestrially. The upward trend in the data rates will necessitate increased use of the higher frequency bands by mobile satellite systems.

In order to achieve the large numbers of users predicted by market studies, the trend toward smaller and less-expensive terminals will need to continue. Small and simple antennas for the mobile terminals will be essential to achieve this goal. New systems must find ways to provide the extra power needed to offer the combination of higher data rates to smaller terminals. For systems based on geostationary satellites, this will require very powerful satellites with extremely large antennas. Because of reduced path loss, for systems using satellites in lower orbits, the size and power of the satellite can be traded off with the altitude of the orbit. Of course, as the altitude of the orbit decreases, the number of satellites needed to provide global coverage increases.

A large number of systems are in the planning stage, and one can expect fierce competition based upon cost to the user, range of services, quality of services, and availability. Because it is usually not feasible to overcome blockage, satellite diversity to offer improved availability may become an important issue. Systems based upon geostationary satellites will have an advantage for services requiring broad area coverage, such as point-to-multipoint communications, broadcasting, and wide-area paging. On the other hand, systems based upon lower Earth orbits will have an advantage for global point-to-point communications services, particularly if large transmission delays are undesirable. An example of such a service is global hand-held telephony.

From the wide range of technologies and service offerings that characterize planned systems, it is clear that the field of mobile satellite communications is far from being mature.

BIBLIOGRAPHY

1. A. Jahn, et al., Narrow- and wide-band channel characterization for land mobile satellite systems: Experimental results at L-band, *Proc. 4th Int. Mobile Satellite Conf.,* 1995, pp. 115–121.
2. W. Jakes, Multipath interference, in W. Jakes (ed.), *Microwave Mobile Communications,* New York: Wiley, 1974.
3. E. Lutz et al., The land mobile satellite channel;—Recording, statistics, and channel model, *IEEE Trans. Veh. Technol.,* **40**: 375–386, 1991.
4. R. Barts and W. Stutzman, Modeling and simulation of mobile satellite propagation, *IEEE Trans. Antennas Propag.,* **40**: 375–381, 1992.
5. C. Loo, A statistical model for a land mobile satellite link, *IEEE Trans. Veh. Technol.,* **VT-34**: 122–127, 1985.
6. R. Cox and P. Kroon, Low bit-rate speech coders for multimedia communication, *IEEE Commun. Mag.,* **34** (12): 34–41, 1996.
7. K. Rijkse, H.263: Video coding for low-bit-rate communication, *IEEE Commun. Mag.,* **34** (12): 42–45, 1996.
8. H. Ma and J. Wolf, On tail biting convolutional codes, *IEEE Trans. Commun.,* **COM-34**: 104–111, 1986.
9. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.,* **44**: 1261–1271, 1996.

10. Y. Yasuda, K. Kashiki, and Y. Hirata, High rate punctured convolutional codes for soft Viterbi decoding, *IEEE Trans. Commun.,* **COM-32**: 315–319, 1984.

11. G. Ungerboeck, Trellis-coded modulation with redundant signal sets: Part I. Introduction, *IEEE Commun. Mag.,* **25** (2): 5–11, 1987.

12. G. Ungerboeck, Trellis-coded modulation with redundant signal sets: Part II. State of the art, *IEEE Commun. Mag.,* **25** (2): 12–21, 1987.

13. A. Viterbi et al., A pragmatic approach to trellis-coded modulation, *IEEE Commun. Mag.,* **27** (7): 11–19, 1989.

14. C. Shannon, Communications in the presence of noise, *Proc. IRE,* **37**: 10–21, 1949.

15. K. S. Gilhousen et al., Increased capacity using CDMA for mobile satellite communications, *IEEE J. Selec. Areas Commun.,* **8**: 503–514, 1990.

16. A. Duel-Hallen, J. Holtzman, and Z. Zvonar, Multiuser detection for CDMA systems, *IEEE Personal Commun.,* **2** (2): 46–58, 1995.

17. M. Kayton (ed.), *Navigation: Land, Sea, Air and Space,* New York: IEEE Press, 1990.

Reading List

J. Lodge and M. Moher, Mobile satellite sysetms, in J. D. Gibson (ed.), *The Communication Handbook,* Boca Raton, FL: CRC Press, 1997, pp. 1015–1031.

T. Logsdon, *Mobile Communications Satellites,* New York: McGraw-Hill, 1996.

S. Kato, Personal communication systems and low earth orbit satellites, *Proc. Space Radio Sci. Symp.,* U.R.S.I., Brussels, Belgium, 1995, pp. 30–42.

W. Wu et al., Mobile satellite communications, *Proc. IEEE,* **82** (9): 1431–1448, 1994.

J. Lodge, Mobile satellite communications systems: Toward global personal communications, *IEEE Commun. Mag.,* **29** (11): 24–30, 1991.

JOHN LODGE
Communications Research Centre