

receiver with the expected frequency and at the expected time. Thus, to precorrect time, a transmitter would divide the distance between itself and the receiver by the speed of light to get the transmission transit time, and then shift the message transmission timing that much ahead. By transmitting the signal early, it will arrive at the receiver at the appropriate time. The time of arrival at the node is given by

$$T_A = T_t + \frac{d}{c} \quad (10.84)$$

where T_t is the actual transmission start time, d is the transmit distance, and c is the speed of light. Similarly, to precorrect the transmission frequency, the transmitter must allow for the Doppler shift caused by relative motion between the transmitter and the intended receiver. To be received correctly, the required transmission radian frequency is

$$\omega \approx \left(1 - \frac{V}{c}\right) \omega_0 \quad (10.85)$$

where c is the speed of light, V is the relative velocity (positive for decreasing transmission distance), and ω_0 is the nominal transmission radian frequency.

Unfortunately, in practice, neither the time nor the frequency pre correction can be done exactly. Even satellites in nominally geostationary orbits move slightly with respect to a point on the earth, and the behavior of the time and frequency references in the terminal and the central node are never entirely predictable. Thus, there will always be some time and frequency pre correction error. The time error may be expressed as

$$T_e = \frac{r_e}{c} + \Delta t \quad (10.86)$$

where r_e is the error in the range estimate and Δt is the difference between the time reference at the terminal and the reference at the receiver. The frequency error may be expressed as

$$\omega_e = \frac{V_e \omega_0}{c} + \Delta \omega \quad (10.87)$$

where V_e is the error in the measured or predicted relative velocity of the transmitter and receiver (the Doppler error) and $\Delta \omega$ is the frequency difference between the transmitter and the receiver frequency references. There are many other sources of time and frequency error in addition to those mentioned here, but they are typically much less important. Spilker [20] gives a reasonably complete accounting of sources of time and frequency error for satellite systems.

The error terms Δt and $\Delta \omega$ are typically due to random fluctuations in frequency references. The time reference for a transmitter or receiver is generally obtained by counting cycles of the frequency reference, so errors in the accuracy of the time and frequency references are related. The fluctuations in a frequency reference are very difficult to characterize statistically, although the power spectral density of the fluctuations is approximated by a sequence of power-law segments

[15]. Frequency references are often specified in terms of a maximum allowable fractional frequency change per day:

$$\delta = \frac{\Delta\omega}{\omega_0} \text{ hertz/hertz/day} \quad (10.88)$$

Typical values for δ range from 10^{-5} to 10^{-6} , for inexpensive crystal oscillators, to 10^{-9} to 10^{-11} , for high-quality crystal oscillators; to 10^{-12} for rubidium standards, to 10^{-13} for cesium standards. An effect of specifying system-frequency references by the maximum fractional frequency is that if there is no intervention, the offset from the nominal frequency ω_0 can grow linearly with time:

$$\Delta\omega(T) = \omega_0 \int_0^T \delta dt + \Delta\omega(0) = \omega_0 \delta T + \Delta\omega(0) \text{ hertz} \quad (10.89)$$

For a cycle-counting time reference, however, the cumulative time offset is related to the cumulative phase error of the reference:

$$\begin{aligned} \Delta t(T) &= \int_0^T \frac{\Delta\omega(t)}{\omega_0} dt + \Delta t(0) \\ &= \int_0^T \delta t dt + \int_0^T \frac{\Delta\omega(0)}{\omega_0} dt + \Delta t(0) \\ &= \frac{1}{2} \delta T^2 + \frac{\Delta\omega(0)T}{\omega_0} + \Delta t(0) \end{aligned} \quad (10.90)$$

Thus, without intervention, a time-reference error can grow quadratically with time. For open-loop transmitter synchronization systems, this quadratic growth in time error often sets limits on how often the external authority must intervene, either to update the terminal's knowledge of receiver timing, or to reset both the receiver's and the transmitter's time references to nominal. The quadratic error growth usually means that timing errors are more of an operational problem than are frequency errors, although this will depend on the system design.

If the transmitter does not have information from measurements on a return link, the time and frequency offsets as modeled by Equations (10.86) to (10.90) will allow a system designer the ability to determine the maximum interval between interventions on the basis of a probability-of-error criterion. Time- and frequency-reference recalibration is often a burdensome procedure; it should be done rarely as possible.

If a terminal has access to a return link from the central node and the ability to make comparative measurements between the local reference and incoming signal parameters, the interval between recalibrations can be made much longer. Large satellite control stations can measure and model the orbital parameters of nominally geostationary satellites to an accuracy of a few tens of feet in range and a few feet/second in velocity relative to the ground terminal. Thus, for the important special case of a synchronous satellite as the central node, the first terms on the right-hand side of Equations (10.86) and (10.87) are usually negligible. When this is

true, the differences between the incoming signal parameters and those generated by the terminal's time and frequency references will approximate the error terms Δt and $\Delta \omega$. These error terms measured on the downlink can be used to compute appropriate corrections to the uplink transmissions. On the other hand, if the time and frequency references are known to be accurate but the link geometry is somewhat in question—perhaps because the terminal is mobile or the satellite is non-geostationary—the same sort of return link measurement could be used to resolve range or velocity uncertainties. These measures of range or relative velocity can then be used to precorrect uplink timing and frequency.

The case where a terminal is able to utilize measurements made on a return link signal is sometimes called quasi-closed-loop transmitter synchronization. The quasi-closed technique is clearly more adaptable to uncertainties in the communication system than is the purely open-loop system. The purely open-loop system requires complete *a priori* knowledge of all important link parameters in order to operate successfully. Unanticipated changes in the links cannot be tolerated. The quasi-closed-loop system, on the other hand, requires *a priori* knowledge of all but one of the important parameters in each of time and frequency, but the remaining term can be determined from observations of the return link. This adds complexity to the terminal, but it also adds the ability to adapt to certain types of unplanned link changes. This degree of adaptability can greatly reduce the frequency of required system calibration.

10.3.2 Closed-Loop Transmitter Synchronization

Closed-loop transmitter synchronization involves the transmission of special synchronization signals that are used to determine the signal's time or frequency error relative to the desired timing or frequency when the signal arrives at the receiver. The results of this determination are then fed back to the transmitter on a return link. The determination of synchronization errors can be either implicit or explicit. If the central node has sufficient processing capacity, the central node may make an actual error measurement. Such a measurement might be the amount and direction of offset, or perhaps simply the direction alone. This information would be formatted and returned to the transmitter on a return link. If the central node has little processing capability, the special synchronization signal may simply be turned around and returned to the transmitter on the return link. In this case, it becomes part of the transmitter's task to interpret the returned signal for itself. The design of a special synchronization signal that lends itself to easy unambiguous interpretation can be a challenge.

The relative advantages and disadvantages of the two types of closed-loop systems have to do with the location of the signal-processing capability and the efficiency of channel usage. A major advantage of having the processing at the central node is that results of the error measurements that are transmitted on the return link can be a short digital sequence. This efficient use of the return link can be important if a single return link is time-division multiplexed between a large number of terminals. A second potential advantage is that the error-measuring capability in the central node can be shared by all terminals communicating through the node. This can amount to a large saving in system processing capability. The principal

potential advantage in having the processing at the terminal is that the central node may not be easily accessible, and reliability considerations may dictate a simple design. This has typically been the case when the central node is a space satellite. With continuing improvements in satellite technology, simplicity requirements can be expected to be less dominant in the future than in the past. Another potential advantage to having the processing in the terminal is that the response can be quicker because there is little processing delay in the central node. This may be important if link parameters are changing very rapidly. The primary disadvantages are the inefficient use of the return channel and that the return signals may be difficult to interpret. This difficulty would arise when the central node is not just a simple repeater but makes symbol decisions and transmits these decisions on the return link. This symbol decision capability can greatly improve the terminal-to-terminal error performance, but it complicates the synchronization procedure. This is because the effects of a time or frequency offset are resident in the return signal indirectly—that is, only as they have affected the symbol decisions. Consider the example of a BFSK transmission to a central node that makes noncoherent bit decisions. The decisions will be dependent on the detected signal energy in the mark and space detectors. If the transmitted signal is an alternating sequence of marks and spaces, the signal at the central node can be modeled as

$$r(t) = \begin{cases} \sin[(\omega_0 + \omega_s + \Delta\omega)t + \theta] & 0 \leq t \leq \Delta t \\ \sin[(\omega_0 + \Delta\omega)t + \theta] & \Delta t < t \leq T \end{cases} \quad (10.91)$$

where T is the symbol interval, ω_0 is one symbol frequency, $(\omega_0 + \omega_s)$ is the other symbol frequency, $\Delta\omega$ is the frequency error at the central node, Δt is the signal arrival time error at the central node, and θ is an arbitrary phase angle. Now, if

$$x = \frac{1}{T} \int_0^T r(t) \cos \omega_0 t \, dt \quad (10.92)$$

and

$$y = \frac{1}{T} \int_0^T r(t) \sin \omega_0 t \, dt \quad (10.93)$$

represent the detector quadrature components, then the detected signal energy can be expressed as

$$\begin{aligned} z^2 &= x^2 + y^2 \\ &= \left(\frac{\sin[(\omega_s + \Delta\omega)\Delta t/2]}{(\omega_s + \Delta\omega)T} \right)^2 + \left(\frac{\sin[\Delta\omega(T - \Delta t)/2]}{\Delta\omega T} \right)^2 \\ &\quad + \frac{\cos(\Delta\omega\Delta t) + \cos[\Delta\omega T - (\omega_s + \Delta\omega)\Delta t] - \cos(\Delta\omega T) - \cos(\omega_s\Delta t)}{2\Delta\omega(\omega_s + \Delta\omega)T^2} \end{aligned} \quad (10.94)$$

For the special case where the time error, Δt , is zero, Equation (10.94) simplifies to

$$z^2 = \left[\frac{\sin(\Delta\omega T/2)}{\Delta\omega T} \right]^2 \quad (10.95)$$

For the case where the frequency offset is zero,

$$z^2 = \left(\frac{T - \Delta t}{2T} \right)^2 + \left[\frac{\sin(\omega_s \Delta t / 2)}{\omega_s T} \right]^2 \quad (10.96)$$

The important thing to notice in Equations (10.94) to (10.96) is that any time error or frequency offset or combination of both will decrease the detected signal energy in the correct symbol detector and introduce signal energy into the incorrect signal detector. This will reduce the effective distance between signals in signal space and degrade error performance. A measurement of error performance, however, which is all that is available on the return link, gives no insight into whether the problem is a frequency offset, a time error, or a combination of both. Thus the transmission of standard signals is not likely to provide a useful response for synchronization.

A useful technique for determining the correct frequency precorrection for our example of BFSK-signaling is to transmit a constant tone whose frequency is the average of the two symbol frequencies. Such a tone should produce a random binary sequence on the return link with equal numbers of marks and spaces. A frequency offset from the average would produce predominately marks or spaces. Finding the center frequency in this way allows accurate frequency precorrection of the signals. Once the correct frequency is found, the transmitter can transmit an alternating sequence of marks and spaces in order to discover correct timing. By varying the timing of the transmission through a range of half-a-symbol interval, the transmitter can look for the timing that provides the worst error performance. When the transmission arrival at the central node is displaced from correct timing by half-a-symbol interval, the two detectors will detect equal amounts of energy, and the binary sequence on the return link will be random. Determining the time when the transmitted and return signals are decorrelated will allow the transmitter to compute the correct transmission timing. Notice that this procedure works better than attempting to find the point at which error performance is the best. Any well-designed system will have sufficient transmission energy to allow for slight timing offsets, so an error-free return signal could be achieved with less than perfect timing. In fact, the larger the signal-to-noise ratio, the worse a best-finding procedure works. A worst-finding system, however, will work well for any well-designed system, and it will improve in potential accuracy with increasing signal-to-noise ratio. This can be seen intuitively, because increased signal-to-noise ratios will allow the system to tolerate larger timing errors, so the improvement in error performance as the timing error decreases from half-a-symbol time will be more rapid in the large signal-to-noise case than in the smaller signal-to-noise ratio case. This will allow a more precise determination of the half-symbol timing position.

10.4 CONCLUSION

This chapter has outlined the fundamental problems and issues associated with synchronization in digital communications. The trade-offs are generally between expense and complexity on the one hand, and error performance on the other. We have discussed receiver synchronization and phase-locked loops (PLL) in particu-

lar. Typically, it is the receiver that takes the most active role in the synchronization of a communications link. Even in cases where a terminal's transmitter assumes the more active role, as in some satellite links, the process is often aided by a return path that has been acquired by the terminal's receiver. Thus, receiver synchronization is more fundamental. Phase-locked loops and their variations are the primary control circuits used to track variations in phase of an incoming signal. The mathematics needed to describe the response of a PLL to a given input involves the solution to a nonlinear differential equation. It was shown, however, that under steady-state conditions, a linearized model provides a useful approximation to system performance. In circumstances where the linearized model cannot be accurately applied, results by Viterbi [8] for first-order loops were introduced. Although exact for first-order loops only, these results have been shown to be useful approximations to the performance of higher order loops as well [5].

The extremely important special case of suppressed-carrier loops was discussed. Suppressed-carrier loops are required to track the phase of an incoming signal that has no average energy at the carrier frequency. The common example of such a signal is one that has been modulated with standard antipodal BPSK. In this situation, a harmonic of the suppressed carrier is produced through the use of a nonlinearity, and the harmonic is tracked.

The next higher level of synchronization treated here was symbol synchronization. Two primary classes of symbol synchronization were discussed. Open-loop synchronizers operate directly on the modulated signal to produce a symbol transition indication. Closed-loop synchronizers use a closed-cycle control loop to acquire and track the symbol transitions.

The highest level of synchronization considered was frame synchronization. To receive the data in a useful form, the receiver must determine which symbols belong to which frames. This knowledge is equivalent to having frame sync, which is usually accomplished by including with the data symbols some recognizable pattern known to the receiver. The receiver scans the incoming data until it recognizes the pattern. Synchronization can be checked by looking for periodic repetitions of the pattern.

This chapter has necessarily been only an outline of the important problems, issues, and results relating to the synchronization of digital communication systems. The interested reader will find that the references listed are excellent works that will provide much greater depth of coverage than space has allowed here.

REFERENCES

1. Peterson, W. W., and Weldon, E. J., *Error-Correcting Codes*, The MIT Press, Cambridge, Mass., 1972.
2. Lee, E. A. and Messerschmitt, D. G., *Digital Communications*, Kluwer Academic Publications, Boston, 1988.
3. Mengali, U., and D'Andrea, A. N., *Synchronization Techniques for Digital Receivers*, Plenum Press, New York, 1997.

4. Meyr, H. Moeneclaey, M., and Fechtel, S. A., *Digital Communication Receivers*, John Wiley & Sons, Inc., New York, 1998.
5. Gardner, F. M., *Phaselock Techniques*, 2nd ed., John Wiley & Sons, Inc., New York, 1979.
6. Davenport, W. B., and Root, W. L., *Random Signals and Noise*, McGraw-Hill Book Company, New York, 1958.
7. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Book Company, New York, 1965.
8. Viterbi, A. J., *Principles of Coherent Communications*, McGraw-Hill Book Company, New York, 1966.
9. Lindsey, W. C., *Synchronization Systems in Communication and Control*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1972.
10. Lindsey, W. C., and Simon, M. K., "Detection of Digital FSK and PSK Using a First-Order Phase-Locked Loop," *IEEE Trans. Commun.*, vol. COM25, no. 2, Feb. 1977, pp. 200-214.
11. Develet, J. A., Jr., "The Influence of Time Delay on Second-Order Phase Lock Loop Acquisition Range," *Int. Telem. Conf.*, London, 1963.
12. Johnson, W. A., "A General Analysis of the False-Lock Problem Associated with the Phase-Lock Loop," *The Aerospace Corp., Rep. TOR-269(4250-45)-1*, NASA Accession N64-13776, 1963.
13. Tausworthe, R. C., "Acquisition and False-Lock Behavior of Phase-Locked Loops with Noisy Inputs," *Jet Propulsion Laboratory, JPL SPS 37-46*, vol. 4, 1967.
14. Franks, L. E., "Synchronization Subsystems: Analysis and Design," in K. Feher, *Digital Communications, Satellite/Earth Station Engineering*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981, Chap. 7.
15. Simon, M. K., and Yuen, J. H., "Receiver Design and Performance Characteristics," in J. H. Yuen, ed., *Deep Space Telecommunications Systems Engineering*, Plenum Press, New York, 1983.
16. Gardner, F. M., "Hangup in Phase-Lock Loops," *IEEE Trans. Commun.*, COM25, October 1977.
17. Blanchard, A., *Phase-Locked Loops*, John Wiley & Sons, Inc., New York, 1976.
18. Holmes, J. K., *Coherent Spread Spectrum Systems*, John Wiley & Sons, Inc., New York, 1982.
19. Lindsey, W. C., and Simon, M. K., eds., *Phase Locked Loops and Their Applications*, IEEE Press, New York, 1977.
20. Spilker, J. J., Jr., *Digital Communications by Satellite*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1977.
21. Wintz, P. A., and Luecke, E. J., "Performance of Optimum and Suboptimum Synchronizers," *IEEE Trans. Commun. Technol.*, June 1969, pp. 380-389.
22. Anderson, J. B., Aulin, T., and Sundberg, C. E., *Digital Phase Modulation*, Plenum Press, New York, 1986.
23. Laurent, P. A., "Exact and Approximate Construction of Digital Phase Modulations by Superposition of Amplitude Modulated Pulses," *IEEE Trans. Commun.*, COM-34, no. 2, pp. 150-160, Feb. 1986.
24. Lui, G. L., "Threshold Detection Performance of GMSK Signal with $BT=0.5$," *IEEE MILCOM 98 Proceedings*, vol. 2, Boston, October 18-21, 1998, pp. 515-519.

25. Kaleh, G., "Differentially Coherent Detection of Binary Partial Response Continuous Phase Modulation with Index 0.5," *IEEE Trans. Commun.*, COM-39, pp. 1335-40, Sept. 1991.
26. Barker, R. H., "Group Synchronization of Binary Digital Systems," in W. Jackson, ed., *Communicaiton Theory*, Academic Press, Inc., New York, 1953.
27. Willard, M. W., "Optimum Code Patterns for PCM Synchronization," *Proc. Natl. Telem. Conf.*, 1962, paper 5-5.
28. Newman, F., and Hofman, L., "New Pulse Sequences with Desirable Correlation Properties," *Proc. Natl. Telem. Conf.*, 1971, pp. 272-282.
29. Wu, W. W., *Elements of Digital Satellite Communications*, Vol. 1, Computer Science Press, Inc., Rockville, Md., 1984.

PROBLEMS

- 10.1. A transmitter is sending an unmodulated tone of constant energy (a beacon) to a distant receiver. The receiver and transmitter are in motion with respect to each other such that $d(t) = D[1 - \sin(mt)] + D_0$, where $d(t)$ is the distance between the transmitter and receiver (possibly this represents an aircraft doing "figure-eight" maneuvers over a ground station), and D , m , and D_0 are constants. This relative motion will cause a Doppler shift in the received transmitter frequency of

$$\Delta\omega_D(t) = \frac{\omega_0 V(t)}{c}$$

where $\Delta\omega_D$ is the Doppler shift, ω_0 is the nominal carrier frequency, $V(t) = d(t)$ is the relative velocity between the transmitter and receiver, and c is the speed of light. Assuming that the linearized loop equations hold and that the receiver's PLL is in lock (zero phase error) at $t = 0$, show that an appropriately designed first-order loop can maintain frequency lock.

- 10.2. Consider a transmitter and receiver that are in relative motion as in Problem 10.1. Once again assume that the linearized loop equations hold. Under this assumption determine the PLL phase error as a function of time for the all-pass and low-pass loop filters of Equations (10.13) and (10.14). Demonstrate that the validity of the assumption of the linearized loop equations depends on the value of the gain K_0 .
- 10.3. A high-performance aircraft is transmitting an unmodulated carrier signal to a ground terminal. The ground terminal is initially in phase lock with the signal. The aircraft performs a maneuver whose dynamics are described by the equation for acceleration, $a(t) = At^2$, where A is a constant. Assuming that the linearized equations apply, determine the minimum order of the phase-locked loop required to track the signal from this aircraft.
- 10.4. Show that the loop bandwidth of a first-order phase-locked loop is given by $B_L = K_0/4$, where K_0 is the loop gain.
- 10.5. A second-order phase-locked loop has a low-pass loop filter given by

$$F(\omega) = \frac{\omega_1}{j\omega + \omega_1}$$

and a loop gain of K_0 . Under the assumption that $K_0 \geq \omega_1/4$, show that the loop bandwidth of this phase-locked loop is given by $B_L = K_0/8$. [Hint:*

$$\int_{-\infty}^{\infty} \frac{dx}{R} = \frac{\pi \cos(h/2)}{2cq^3 \sin h} \quad \text{for } 4ac > b^2$$

where $R = a + bx^2 + cx^4$, $q = \sqrt[4]{a/c}$, and $\cos h = -b/2\sqrt{ac}$.]

- 10.6.** A first-order phase-locked loop with loop gain K_0 is disturbed by additive white Gaussian noise of normalized (to unit signal energy) two-sided power spectral density of $N_0/2$ watts/hertz. Determine the necessary relationship between noise power spectral density and loop gain if the loop is designed to cycle slip no more often than once per day.
- 10.7.** Viterbi [8] determined that the probability density function of the output phase of a first-order phase-locked loop disturbed by white Gaussian noise is given by

$$p(\phi) = \frac{\exp(\rho \cos \phi)}{2\pi I_0(\rho)}, \quad |\phi| \leq \pi, \quad \rho \geq 0$$

Demonstrate that $p(\phi)$ given above is in fact a probability density function, and compute the mean and variance of ϕ .

- 10.8.** Computer simulations and laboratory measurements have indicated that the time between cycle slips is exponentially distributed; that is, the distribution function of the time between cycle slips, T , is given by

$$p(T) = 1 - \exp\left(-\frac{T}{T_m}\right)$$

Given this distribution function, find the mean time between cycle slips and the variance about this mean as functions of T_m . If the mean time between cycle slips is 1 day, what is the probability of cycle slips less than 1 hour apart? More than 3 days apart?

- 10.9.** Consider a second-order phase-locked loop with a low-pass loop filter

$$F(\omega) = \frac{\omega_1}{j\omega + \omega_1}$$

During aided acquisition it is desired that this loop be scanned throughout a 1000-radian uncertainty region in 1 s. If the relationship between loop gain and filter constant is $K_0 = 2\omega_1$, what is the required relationship between loop gain and the single-sided additive white Gaussian noise power spectral density, N_0 ? Determine the largest value of N_0 that can be accommodated.

- 10.10.** Consider the operation of an open-loop symbol synchronizer whose band pass filter (BPF) has a bandwidth of $0.1/T$ hertz, where T is the symbol period. For a bit energy-to-noise power spectral density ratio (E_b/N_0) of 10 dB, determine the magnitude of the approximate mean and variance of the fractional tracking error, and compute an upper bound on the probability that the tracking error exceeds three times its approximate fractional mean. (Hint: Consider the Chebyshev inequality [7].)

*I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (New York: Academic Press, 1965), 2.161.1.

- 10.11.** A communication system is used to transmit commands to a payload at a data rate of 100 bits/s. Each command is preceded by an N -bit header that identifies it in the data stream. Assuming that except possibly for the header, the bits appear to be random [$P(1) = P(0) = \frac{1}{2}$], what is the minimum-length header that would provide an expected frequency of false alarms of one per year? For a channel bit error probability of 10^{-5} , what is the probability of missing this header? What is the miss probability if the channel error probability is 2×10^{-2} ? If the system is redesigned to accept the header with up to two errors, what is the minimum required length for an expected false alarm rate of one per year? What is the miss probability with this new system and a channel-error probability of 2×10^{-2} ?
- 10.12.** A deep-space probe is moving away from the earth at a nominal velocity of 15,000 m/s with a velocity uncertainty of ± 3 m/s. The probe frequency reference is specified to have a drift rate of no more than 10^{-9} Hz/Hz/day. The nominal downlink transmission frequency is 8 GHz. After a 1-month (30-day) silence the probe begins a scheduled transmission toward an earth terminal. The earth terminal contains a cesium standard. What center frequency and frequency search bandwidth should be used by the ground station? Assuming that the range to the probe was accurately known at the beginning of the month, and that the uncertainty in the probe's time and frequency references were zero [$\Delta t(0) = 0$, $\Delta \omega(0) = 0$], what is the uncertainty in the time of arrival of the downlink transmission?
- 10.13.** A communications link operates at a nominal center frequency of 10 GHz for a single brief period once a day. The receiver operates with a second-order PLL that has a pull-in range of ± 1 kHz. Assuming that the loop acquires using self-acquisition and that both the transmitter and receiver use the same type of frequency reference, what type of frequency reference must this be?
- 10.14.** At some moment in time ($t = 0$), a clock's output signal has an error of -4×10^{-3} seconds with reference to a master clock. At this time, $t = 0$, the clock's driving oscillator is oscillating at the correct frequency f_r , but the oscillator is speeding up at the rate of 2-parts-in- 10^{10} per day.
- How long will it take, in days, before the clock output signal has zero error?
 - If the clock is allowed to operate for 30 days beyond the time of zero error, what will the error be then?
- 10.15.** Under the usual assumptions of zero-mean AWGN and equal energy signals, verify that the right-hand side of Equation (10.67) is in the form of the likelihood function for the estimation of carrier phase and symbol timing.
- 10.16.** Consider the case of full-response MSK signaling with a synchronization "training sequence" of alternating ones and zeros (i.e., $\alpha_k = 1$ for k even, and -1 for k odd).
- Show that for this example there are only four distinct phase states $\{\Phi_k\}$.
 - Derive the form of the filters $h^{(i)}(t)$, given in Equation (10.64).
 - Using the results obtained in part b), derive expressions for Equations (10.68) and (10.69) for this simplified case.
- 10.17.** Develop a rationale for why the iterative procedure suggested in Equations (10.70) and (10.71) will (or will not) work.

QUESTIONS

- 10.1.** What is the definition of *synchronization* in the context of a digital communication system, and why is it important? (See Section 10.1.1.)

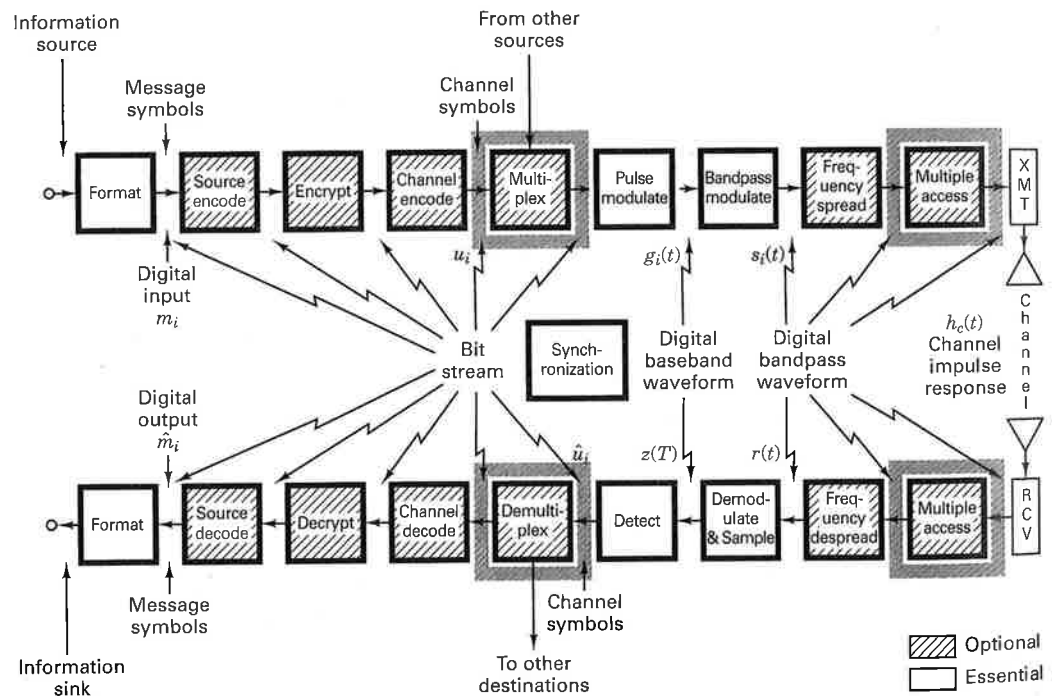
- 10.2. Why may a synchronization system that works well for a home radio receiver possibly be *inadequate* in a high performance aircraft? What modifications are likely to be required for adequate performance? (See Section 10.1.2.)
- 10.3. The *linearized loop equation* depends on an approximation. What is that approximation, why is it appropriate for loops that are in lock or near lock, and why is it not appropriate for acquisition analysis? (See Section 10.2.1.)
- 10.4. Second-order phase locked loops have several advantages in their performance, and as the basis for phase-tracking performance analysis. Name two such advantages. (See Section 10.2.1.1.)
- 10.5. Why are *continuous phase modulation schemes* of increasing importance in modern communications system, and **what synchronization** challenges do they pose? (See Section 10.2.3.1.)
- 10.6. What are the advantages and disadvantages of *data-aided* versus *non-data-aided* synchronizers? (See Section 10.2.3.2.)
- 10.7. Describe a situation in which it may be appropriate to require a transmitter to synchronize itself to the expectations of a receiver. (See Section 10.3.)

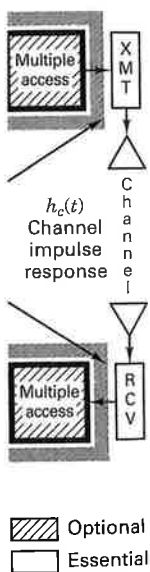
EXERCISES

Using the Companion CD, run the exercises associated with Chapter 10.

CHAPTER 11

Multiplexing and Multiple Access





A communications resource (CR) represents the time and bandwidth that is available for communication signaling associated with a given system. It can be graphically envisioned with a plane, where the abscissa represents time, and the ordinate represents frequency. For the efficient development of a communication system, it is important to plan out the resource allocation among system users, so that no block of time/frequency is wasted, and so that the users can share the resource in an equitable manner.

The terms "multiplexing" and "multiple access" refers to the sharing of a CR. There is a subtle difference between multiplexing and multiple access. With *multiplexing*, users' requirements or plans for CR sharing are fixed, or at most, slowly changing. The resource allocation is assigned *a priori*, and the sharing is usually a process that takes place within the confines of a *local site* (e.g., a circuit board). *Multiple access*, however, usually involves the *remote sharing* of a resource, such as in the case of satellite communications. With a dynamically changing multiple access scheme, a system controller must become aware of each user's CR needs; the amount of time required for this information transfer constitutes an overhead and sets an upper limit on the efficiency of the utilization of the CR.

11.1 ALLOCATION OF THE COMMUNICATIONS RESOURCE

There are three basic ways to increase the throughput (total data rate) of a communications resource (CR). The first way is either to increase the transmitter's effective isotropic radiated power (EIRP) or to reduce system losses so that the

received E_b/N_0 is increased. The second way is to provide more channel bandwidth. The third approach is to make the allocation of the CR more efficient. This third approach is the domain of communications multiple access. The problem, in the context of a satellite transponder, is to efficiently allocate portions of the transponder's fixed CR to a large number of users who seek to communicate digital information to each other at a variety of bit rates and duty cycles. The basic ways of distributing the communications resource, listed under the heading "multiplexing/multiple access" in Figure 11.1, are the following:

1. *Frequency division (FD)*. Specified subbands of frequency are allocated.
2. *Time division (TD)*. Periodically recurring time slots are identified. With some systems, users are provided a fixed assignment in time. With others, users may access the resource at random times.
3. *Code division (CD)*. Specified members of a set of orthogonal or nearly orthogonal spread spectrum codes (each using the full channel bandwidth) are allocated.
4. *Space division (SD) or multiple beam frequency reuse*. Spot beam antennas are used to separate radio signals by pointing in different directions. It allows for reuse of the same frequency band.
5. *Polarization division (PD) or dual polarization frequency reuse*. Orthogonal polarizations are used to separate signals, allowing for reuse of the same frequency band.

The key to *all* multiplexing and multiple access schemes is that various signals share a CR without creating unmanageable interference to each other in the detection process. The allowable limit of such interference is that signals on one CR channel should not significantly increase the probability of error in another channel. Orthogonal signals on separate channels will avoid interference between users. Signal waveforms $x_i(t)$, where $i = 1, 2, \dots$, are defined to be orthogonal if they can be described in the time domain by

$$\int_{-\infty}^{\infty} x_i(t) x_j(t) dt = \begin{cases} K & \text{for } i = j \\ 0 & \text{otherwise} \end{cases} \quad (11.1)$$

where K is a nonzero constant. Similarly, the signals are orthogonal if they can be described in the frequency domain by

$$\int_{-\infty}^{\infty} X_i(f) X_j(f) df = \begin{cases} K & \text{for } i = j \\ 0 & \text{otherwise} \end{cases} \quad (11.2)$$

where the functions $X_i(f)$ are the Fourier transforms of the signal waveforms $x_i(t)$. Channelization, characterized by orthogonal waveforms, as shown in Equation (11.1), is called time-division multiplexing or time-division multiple access (TDM/TDMA), and that characterized by orthogonal spectra, as shown in Equation (11.2), is called frequency-division multiplexing or frequency-division multiple access (FDM/FDMA).

el bandwidth.
ent. This third
oblem, in the
the transpon-
digital infor-
basic ways of
g "multiplex-

llocated.
entified. With
With others,

nal or nearly
el bandwidth)

eam antennas
ions. It allows

e. Orthogonal
of the same

various signals
r in the detec-
ls on one CR
another chan-
between users.
nal if they can

(11.1)

if they can be

(11.2)

iveforms $x_i(t)$.
in Equation
access (TDM/
in Equation
sion multiple

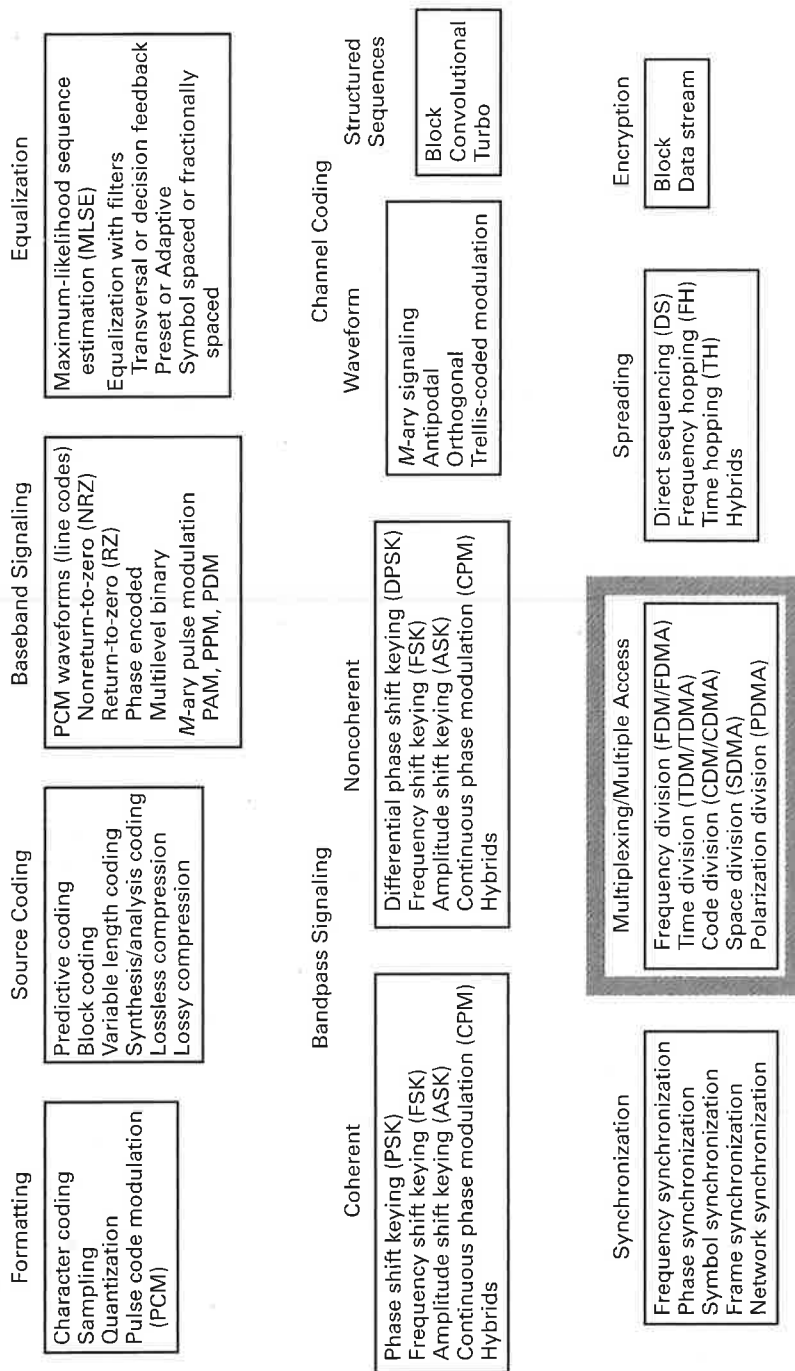


Figure 11.1 Basic digital communication transformations.

11.1.1 Frequency-Division Multiplexing/Multiple Access

11.1.1.1 Frequency-Division Multiplex Telephone

In the early days of telephony, a separate pair of wires was needed for each telephone trunk circuit (trunk circuits interconnect intercity switching centers). As illustrated in Figure 11.2, the skies of all the major cities in the world grew dark with overhead wires as the demand for telephone service grew. A major develop-

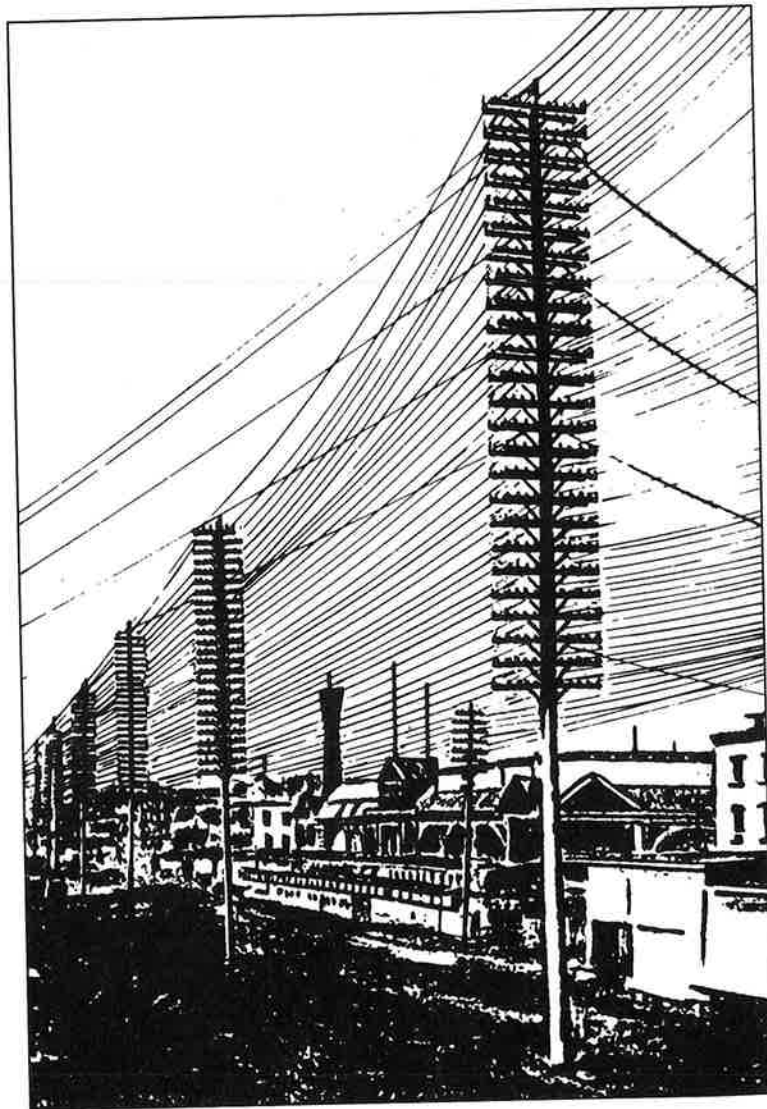


Figure 11.2 In the early days of telephony a pair of wires was needed for each trunk circuit.

ment in the early 1900s, frequency-division multiplex (FDM) telephony, made it possible to transmit several telephone signals simultaneously on a single wire, and thereby transformed the methods of telephone transmission.

The communications resource (CR) is illustrated in Figure 11.3 as the frequency-time plane. The channelized spectrum shown here is an example of FDM or FDMA. The assignment of a signal or user to a frequency band is *long term* or *permanent*; the CR can simultaneously contain several spectrally separate signals. The first frequency band contains signals that operate between frequencies f_0 and f_1 , the second between frequencies f_2 and f_3 , and so on. The spectral regions between assignments, called *guard bands*, act as buffer zones to reduce interference between adjacent frequency channels. We might ask: How does one transform a baseband signal so that it occupies a higher frequency band? The answer: *heterodyning* or *mixing*, also called *modulating*, the signal with a fixed frequency from a sine-wave oscillator.

If two input signals to a mixer are sinusoids with frequencies f_A and f_B , the mixing or multiplication will yield new sum and difference frequencies at f_{A+B} and f_{A-B} . The trigonometric identity

$$\cos A \cos B = \frac{1}{2} [\cos (A + B) + \cos (A - B)] \quad (11.3)$$

describes the effect of the mixer. Figure 11.4a illustrates the mixing of a typical voice-grade telephone signal $x(t)$ (baseband frequency range is 300 to 3400 Hz) with a sinusoid from a 20-kHz oscillator. The baseband two-sided magnitude spectrum, $|X(f)|$, is shown in Figure 11.4a. Can the mixer be a linear device? No. The output signal of a linear device will only consist of the *same* component frequencies as the input signal, differing only in amplitude and/or phase.

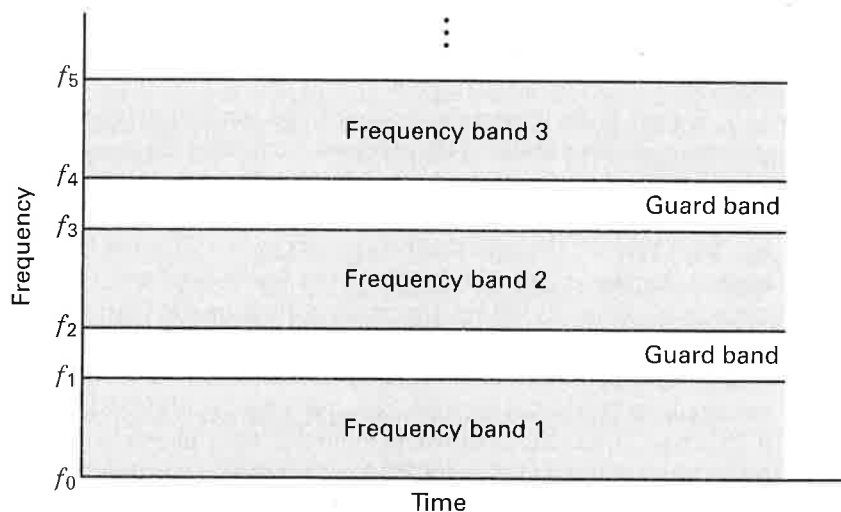


Figure 11.3 Frequency-division multiplexing.

eeded for each
ng centers). As
orld grew dark
major develop-

eeded

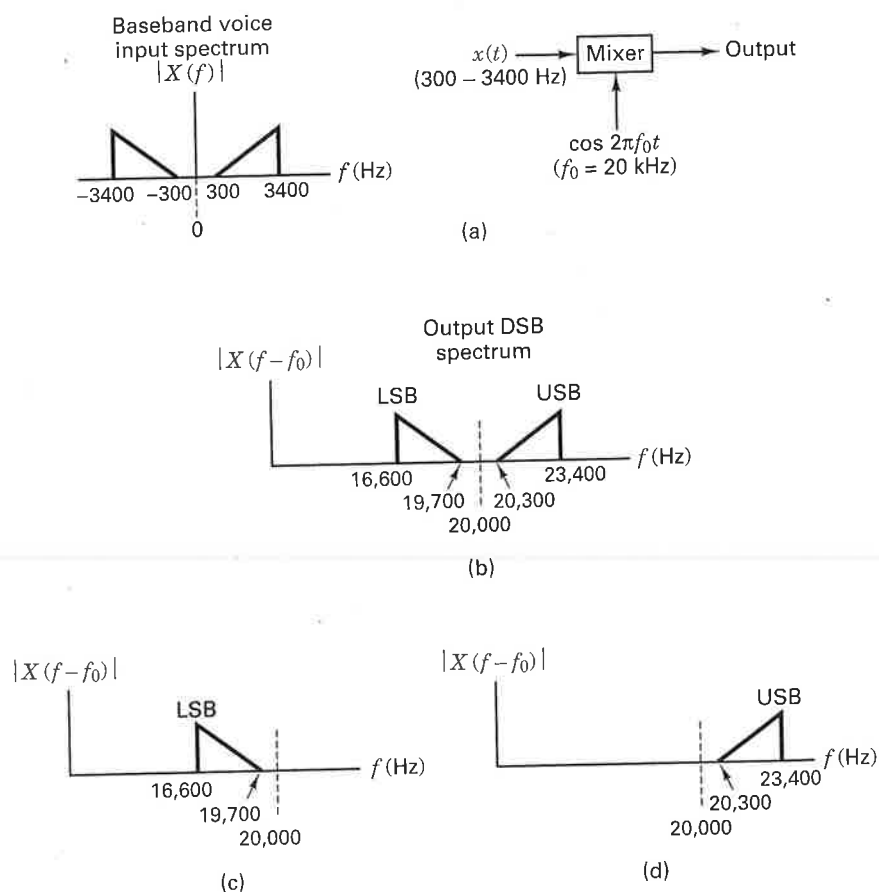


Figure 11.4 Heterodyning (mixing). (a) Mixing operation. (b) Mixer output spectrum. (c) Lower sideband. (d) Upper sideband.

Figure 11.4b illustrates the one-sided magnitude spectrum $|X(f - f_0)|$ at the mixer output. As a result of the mixing described by Equation 11.3, the output spectrum is a frequency-upshifted version of the baseband spectrum, centered at the oscillator frequency of 20 kHz. This spectrum is called a *double-sideband (DSB) spectrum* because the information appears in two different bands of the positive frequency domain. Figure 11.4c shows the lower sideband (LSB), whose frequency range is 16,600 to 19,700 Hz, the result of filtering the DSB spectrum. This sideband is sometimes referred to as the *inverted sideband* because the order of low-to-high frequency components is the reverse of that of the baseband components. Filtering can similarly be used to separate the upper sideband (USB), whose frequency range is 20,300 to 23,400 Hz, as shown in Figure 11.4d. This sideband is sometimes referred to as the *erect sideband* because the order of the low-to-high frequency components corresponds to that of the baseband components. Each sideband of the DSB spectrum contains the same information. Thus, only one sideband, either the USB or the LSB, is needed in order to retrieve the original baseband data.

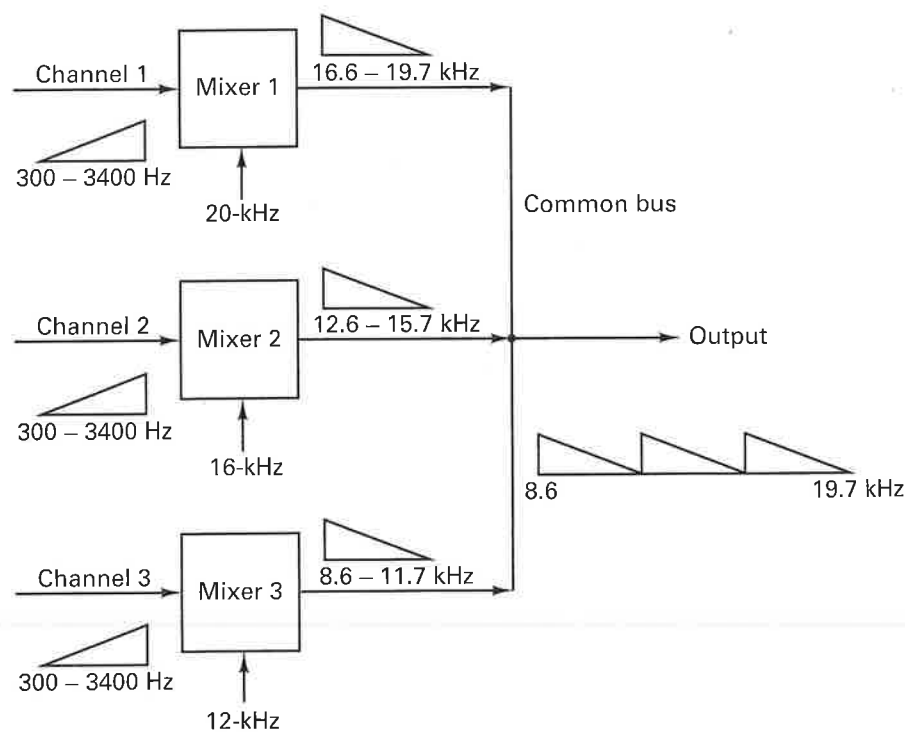


Figure 11.5 Simple FDM example. Three frequency-shifted voice channels.

A simple FDM example with three translated voice channels is seen in Figure 11.5. In channel 1, the 300- to 3400-Hz voice signal is mixed with a 20-kHz oscillator. In channels 2 and 3, a similar type of voice signal is mixed with a 16-kHz and 12-kHz oscillator, respectively. Only the lower sidebands are retained; the result of the mixing and filtering (to remove the upper sidebands) yields the frequency-shifted voice channels shown in Figure 11.5. The composite output waveform is just the sum of the three signals, having a total bandwidth in the range 8.6 to 19.7 kHz.

Figure 11.6 illustrates the two lowest levels of the FDM multiplex hierarchy for telephone channels. The first level consists of a *group* of 12 channels modulated onto subcarriers shown in the range 60 to 108 kHz. The second level is made up of five groups (60 channels) called a *supergroup* modulated onto the subcarriers shown in the range 312 to 552 kHz. The multiplexed channels are now treated as a composite signal that can be transmitted over cables or can be further modulated onto a carrier wave for radio transmission.

11.1.1.2 Frequency-Division Multiple Access of Satellite Systems

Most of the world's communication satellites are positioned in a *geostationary* or *geosynchronous* orbit. This means that the satellite is in a circular orbit, in the same plane as the earth's equatorial plane, and at such an altitude (approximately 19,330 nautical miles) that the orbital period is identical with the earth's rotational period.

SB
1
f (Hz)
3,400

Mixer

$f - f_0$ at the
e output spec-
red at the os-
eband (DSB)
e positive fre-
se frequency
This sideband
of low-to-high
ents. Filtering
quency range
imes referred
y components
he DSB spec-
ie USB or the

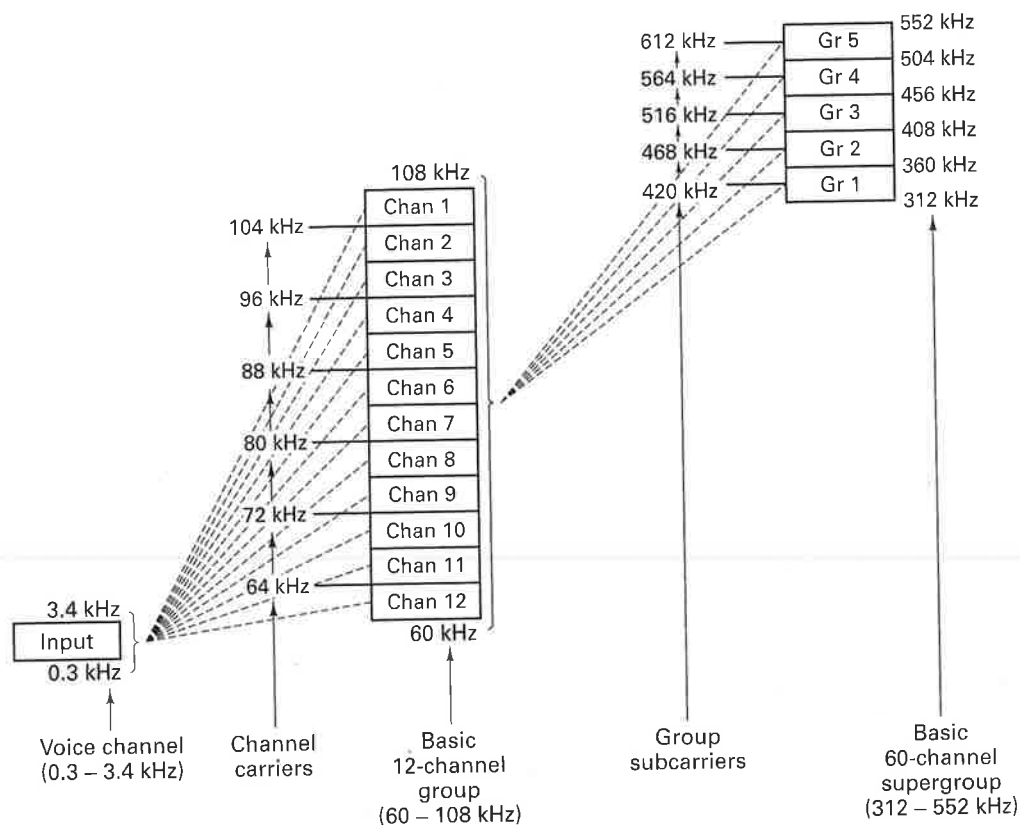


Figure 11.6 Modulation plan of a typical frequency-division multiplex system.

Since such satellites appear stationary when viewed from the earth, three of them spaced 120° apart can provide worldwide coverage (except for the polar regions). Most communication satellite systems are made up of nonregenerative repeaters or transponders. *Nonregenerative* means that the uplink (earth-to-satellite) transmissions are simply amplified, frequency shifted, and retransmitted on the downlink (satellite-to-earth) without any demodulation/remodulation or signal processing. The most popular frequency band for commercial satellite communications, called *C-band*, uses a 6-GHz carrier for the uplink and a 4-GHz carrier for the downlink. For C-band satellite systems, *each satellite* is permitted, by international agreement, to use a 500-MHz-wide spectral assignment. Typically, each satellite has 12 transponders with a bandwidth of 36 MHz each. The most common 36-MHz transponders operate in an FDM/FM/FDMA (frequency-division multiplex, frequency-modulated, frequency-division multiple access) multidestination mode. Let us consider each component of this name:

1. **FDM.** Signals such as telephone signals, each one having a single-sideband 4-kHz spectrum (including guard bands) are FDM'd to form a multichannel composite signal.

552 kHz
504 kHz
456 kHz
408 kHz
360 kHz
312 kHz

Basic
3-channel
per group
2 - 552 kHz

am.

three of them
similar regions).
repeaters or
te) transmis-
the downlink
l processing.
ations, called
ownlink. For
greement, to
12 transpon-
sponders op-
-modulated,
onsider each

gle-sideband
multichannel

2. *FM*. The composite signal is frequency-modulated (FM) onto a carrier and transmitted to the satellite.
3. *FDMA*. Subdivisions of the 36-MHz transponder bandwidth may be assigned to different users. Each user receives a specific bandwidth allocation whereby he or she can access the transponder.

Thus, composite FDM channels are FM modulated and transmitted to the satellite within the bandwidth allocation of an FDMA plan. The major advantage of FDMA (compared with TDMA) is its simplicity. The FDMA channels require no synchronization or central timing; each channel is almost independent of all other channels. Later we discuss some advantages of TDMA compared to FDMA.

11.1.2 Time-Division Multiplexing/Multiple Access

In Figure 11.3, sharing of the communications resource (CR) is accomplished by allocating frequency bands. In Figure 11.7, the same CR is shared by assigning each of M signals or users the full spectral occupancy of the system for a short duration of time called a *time slot*. The unused time regions between slot assignments, called *guard times*, allow for some time uncertainty between signals in adjacent time slots, and thus act as buffer zones to reduce interference. Figure 11.8 is an illustration of a typical TDMA satellite application. Time is segmented into intervals called *frames*. Each frame is further partitioned into assignable user time slots. The frame structure repeats, so that a fixed TDMA assignment constitutes one or more slots that periodically appear during each frame time. Each earth station transmits its data in bursts, timed so as to arrive at the satellite coincident with its designated time slot(s). When the bursts are received by the satellite transponder, they are

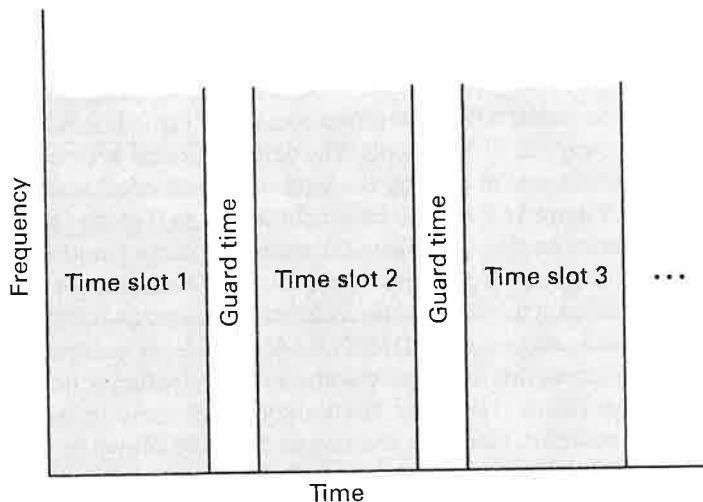


Figure 11.7 Time-division multiplexing.

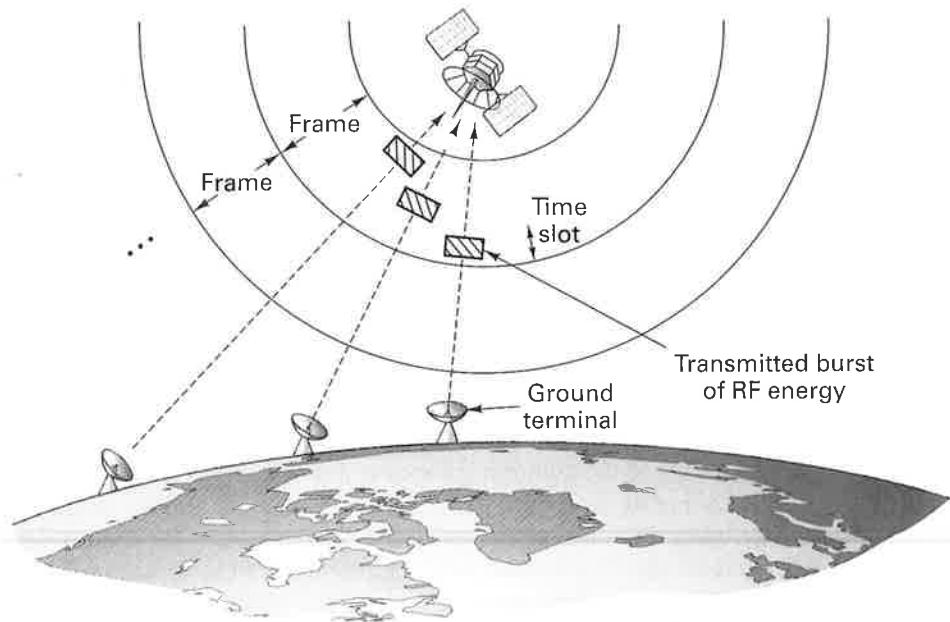


Figure 11.8 Typical TDMA configuration.

retransmitted on the downlink, together with the bursts from other stations. A receiving station detects and demultiplexes the appropriate bursts and feeds the information to the intended user.

11.1.2.1 Fixed-Assignment TDM/TDMA

The simplest TDM/TDMA scheme, called *fixed-assignment TDM/TDMA*, is so named because the M time slots that make up each frame are preassigned to signal sources, long term. Figure 11.9 illustrates, in block diagram form, the operation of such a system. The multiplexing operation consists of providing each source with an opportunity to occupy one or more slots. The demultiplexing operation consists of deslotting the information and delivering the data to the intended sink. The two commutating switches in Figure 11.9 have to be synchronized so that the message corresponding to source 1, for example, appears on the channel 1 output, and so on. The message itself generally comprises a preamble portion and a data portion. The preamble portion usually contains synchronization, addressing, and error-control sequences.

A fixed-assignment TDM/TDMA scheme is extremely efficient when the source requirements are predictable, and the traffic is heavy (the time slots are most always filled). However, for bursty or sporadic traffic, the fixed-assignment scheme is wasteful. Consider the simple example shown in Figure 11.10. In this example there are four time slots per frame; each slot is preassigned to users *A*, *B*, *C*, and *D*, respectively. In Figure 11.10a we see a typical activity profile of the four users. During the first frame time, user *C* has no data to transmit; during the second frame time, user *B* has none, and during the third frame time, user *A* has none. In a

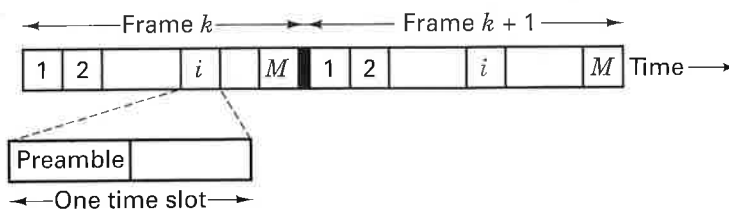
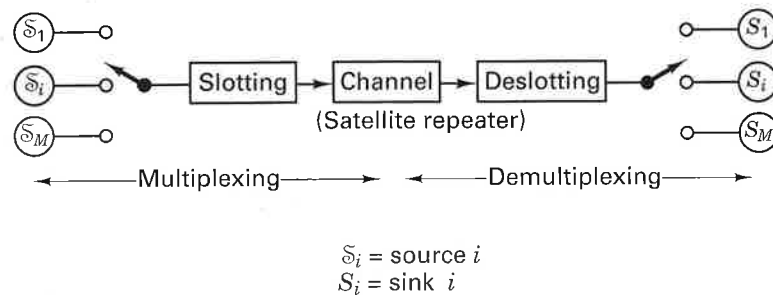


Figure 11.9 Fixed-assignment TDM.

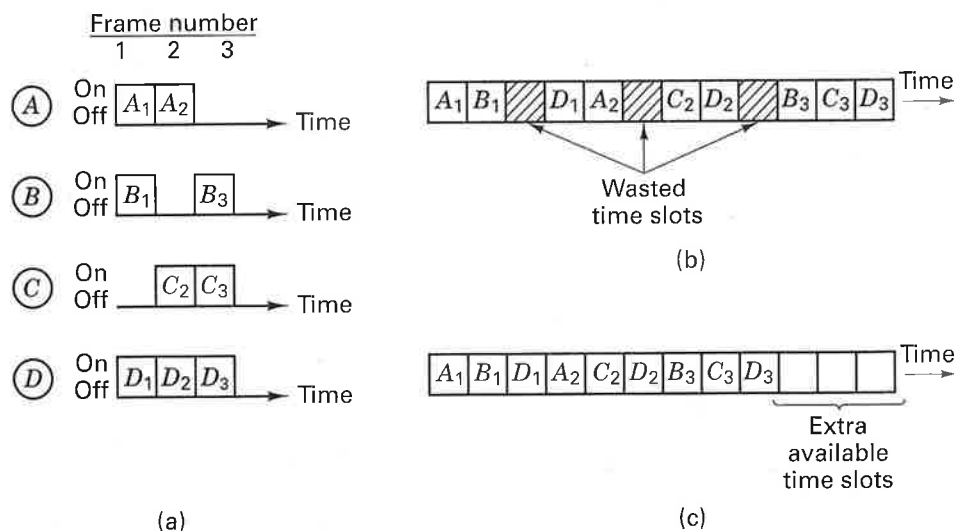


Figure 11.10 Fixed-assignment TDM versus packet switching. (a) Data source activity profiles. (b) Fixed-assignment time-division multiplexing. (c) Time-division packet switching (concentration).

fixed-assignment TDMA scheme, all of the slots within a frame are preassigned. If the “owner” of a slot has *no* data to send during a particular frame, that slot is wasted. The data stream, shown in Figure 11.10b, illustrates the wasted time slots in this example. When source requirements are unpredictable, as in this example, there can be more efficient schemes, involving the dynamic assignment of the slots rather than a fixed assignment. Such schemes are variously known as packet-switched systems, statistical multiplexers, or concentrators; the effect, shown in Figure 11.10c is to use all the slots in a frame in such a way that capacity is conserved.

11.1.3 Communications Resource Channelization

In Figure 11.3 we considered that the CR is partitioned into spectral bands, and in Figure 11.7 we viewed the same CR as being partitioned into time slots. Figure 11.11 represents a more general organization of the CR allowing for the assignment of a frequency band for a prescribed period of time. Such a multiple access scheme is referred to as *combined FDMA/TDMA*. For the assignments of frequency bands, let us assume an equal apportionment of the total bandwidth W , among M user groups or classes, so that M disjoint frequency bands of width W/M hertz are continuously available to their assigned group. Similarly, for the assignment of time slots, the time axis is partitioned into time frames, each of duration T , and the frames are partitioned into N slot times, each of duration T/N . We assume that the users are time synchronized and that the assigned slots are located periodically within the frames. Each user in each frequency band is permitted to transmit during each periodic appearance of the user’s assigned slot, and is permitted to use the assigned channel bandwidth for the slot duration. A slot is uniquely determined as the m th slot within the n th frame. Referring to Figure 11.11, we can describe the time of a particular slot (n, m) with reference to time zero as follows:

$$\text{time of slot } (n, m) = nT + \frac{(m-1)T}{N} \leq t \leq nT + \frac{mT}{N}$$

$$n = 0, 1, \dots; m = 1, 2, \dots, N. \quad (11.4)$$

The n th frame time, T , is denoted by the time interval $[nT, (n+1)T]$. As can be seen in Figure 11.11, the domain of the unit signal is the intersection of the time slot (n, m) and the frequency band (j) . Assume that a modulation/coding system is chosen so that the full bandwidth W of the CR can support R bits/s. In any frequency band having a bandwidth of W/M hertz, the associated bit rate will be R/M bits/s. FDMA alone would provide M bands each with a bandwidth of $1/M$ of the full bandwidth of the CR. TDMA alone would provide the full system bandwidth for each of the N slots, where the duration of each slot is $1/N$ of the frame time.

11.1.4 Performance Comparison of FDMA and TDMA

11.1.4.1 Bit Rate Equivalence of FDMA and TDMA

Figure 11.12 highlights the basic differences between an FDMA and TDMA system in a communications resource capable of supporting a total of R bits/s. In Figure 11.12a the system bandwidth is divided into M orthogonal frequency bands.

preassigned. If
ne, that slot is
ed time slots in
this example,
ent of the slots
wn as packet-
ect, shown in
at capacity is

ids, and in Fig-
re 11.11 repre-
of a frequency
rred to as *com*-
sume an equal
, so that M dis-
their assigned
oned into time
es, each of du-
assigned slots
and is permit-
ot, and is per-
ot is uniquely
11, we can de-
lows:

(11.4)
]. As can be
1 of the time
ing system is
. In any fre-
; will be R/M
of $1/M$ of the
n bandwidth
me time.

and TDMA
f R bits/s. In
ency bands.

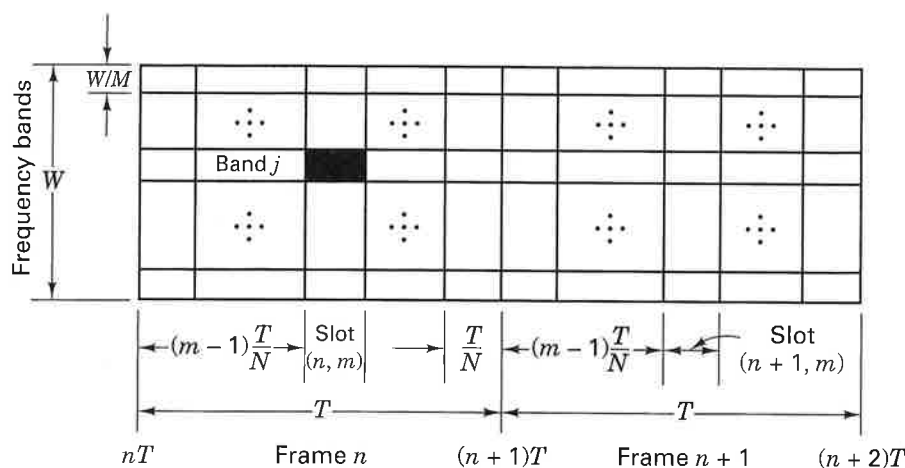


Figure 11.11 Communications resource: time/frequency channelization.

Hence each of the M sources, $\mathcal{S}_m (1 \leq m \leq M)$ can simultaneously transmit at a bit rate of R/M bits/s. In Figure 11.12b the frame is divided into M orthogonal time slots. Hence each of the M sources bursts its transmission at R bits/s, M times faster than the equivalent FDMA user for $(1/M)$ th the time. In both cases, the source $\mathcal{S}_m$ transmits information at an average rate of R/M bits/s.

Let the information generated by each of the sources in Figure 11.12 be organized into b -bit groups, or *packets*. In the case of FDMA, the b -bit packets are transmitted in T seconds over each of the M disjoint channels. Therefore, the total bit rate required is

$$R_{\text{FD}} = M \frac{b}{T} \quad \text{bits/s} \quad (11.5)$$

In the case of TDMA, the b bits are transmitted in T/M seconds from each source. Therefore, the bit rate required is

$$R_{\text{TD}} = \frac{b}{T/M} \quad \text{bits/s} \quad (11.6)$$

Since Equations (11.5) and (11.6) yield identical results, we can conclude that

$$R_{\text{FD}} = R_{\text{TD}} = R = \frac{Mb}{T} \quad \text{bits/s} \quad (11.7)$$

Thus, both systems require the same full CR data rate, R bits/s.

11.1.4.2 Message Delays in FDMA and TDMA

From the previous sections it might appear that the duality between FDMA and TDMA will result in equivalent performance. This is not the case when the metric of performance is the average packet *delay*. It can be shown [1, 2] that

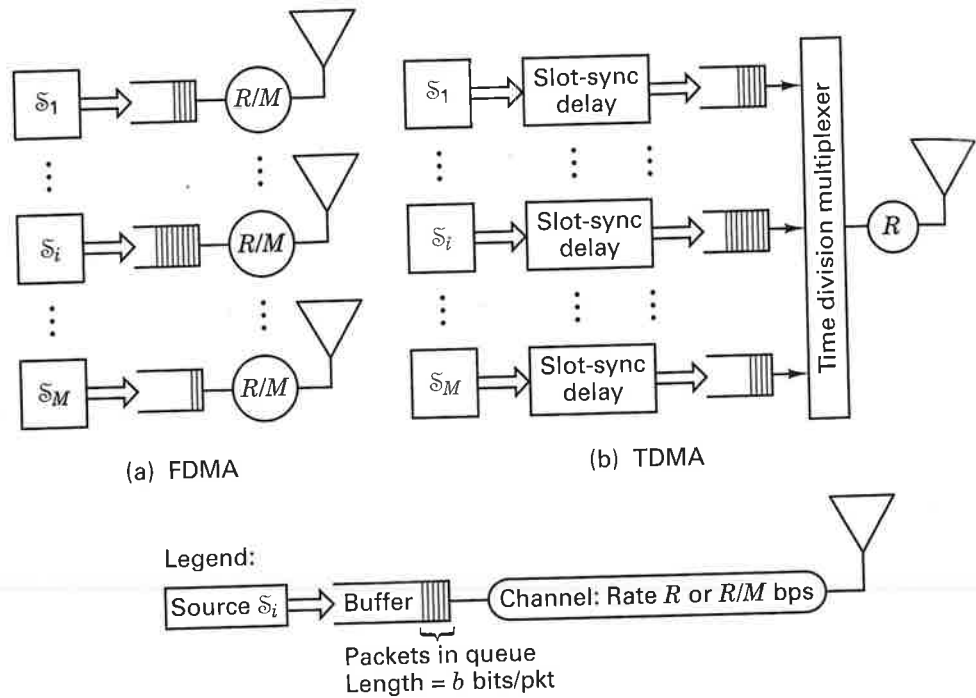


Figure 11.12 (a) FDMA: frequency divided into M orthogonal frequency bands. (b) TDMA: time divided into M orthogonal time slots (one packet per time slot).

TDMA is inherently superior to FDMA in the sense that the average packet delay using TDMA is less than the delay using FDMA.

As before, we assume that in the case of FDMA the system bandwidth is divided into M orthogonal frequency bands, and in the case of TDMA the frame is divided into M orthogonal time slots. For the analysis of message delay, the simplest case is that of deterministic data sources. It is assumed that the CR is 100% utilized, so that all frequency bands in the case of FDMA, and all time slots in the case of TDMA, are filled with data packets. For simplicity, it is also assumed that there are *no* overhead costs such as guard bands or guards times. The message delay can be defined as

$$D = w + \tau \quad (11.8)$$

where w is the average packet waiting time (prior to transmission) and τ is the packet transmission time. In the FDMA case, each packet is sent over a T -second interval, so the packet transmission time for FDMA is simply

$$\tau_{FD} = T \quad (11.9)$$

In the TDMA case, each packet is sent in slots of T/M seconds. We can thus write the TDMA packet transmission time with the use of Equation (11.7) as

$$\tau_{TD} = \frac{T}{M} = \frac{b}{R} \quad (11.10)$$

Since the FDMA channel is continuously available and packets are sent as soon as they are generated, the waiting time, w_{FD} , for FDMA is

$$w_{FD} = 0 \quad (11.11)$$

FDMA and TDMA bit streams are compared in Figure 11.13. For TDMA, Figure 11.13a illustrates that each user's slot begins at a different point in the T -second frame; that is, packet S_{mk} will start at $(m-1)T/M$ seconds ($1 \leq m \leq M$) after the packet generation instant. Therefore, the average waiting time that a TDMA packet sustains before transmission begins is

$$w_{TD} = \frac{1}{M} \sum_{m=1}^M (m-1) \frac{T}{M} = \frac{T}{M^2} \sum_{n=0}^{M-1} n = \frac{T}{M^2} \frac{(M-1)(M)}{2} \quad (11.12)$$

$$= \frac{T}{2} \left(1 - \frac{1}{M} \right)$$

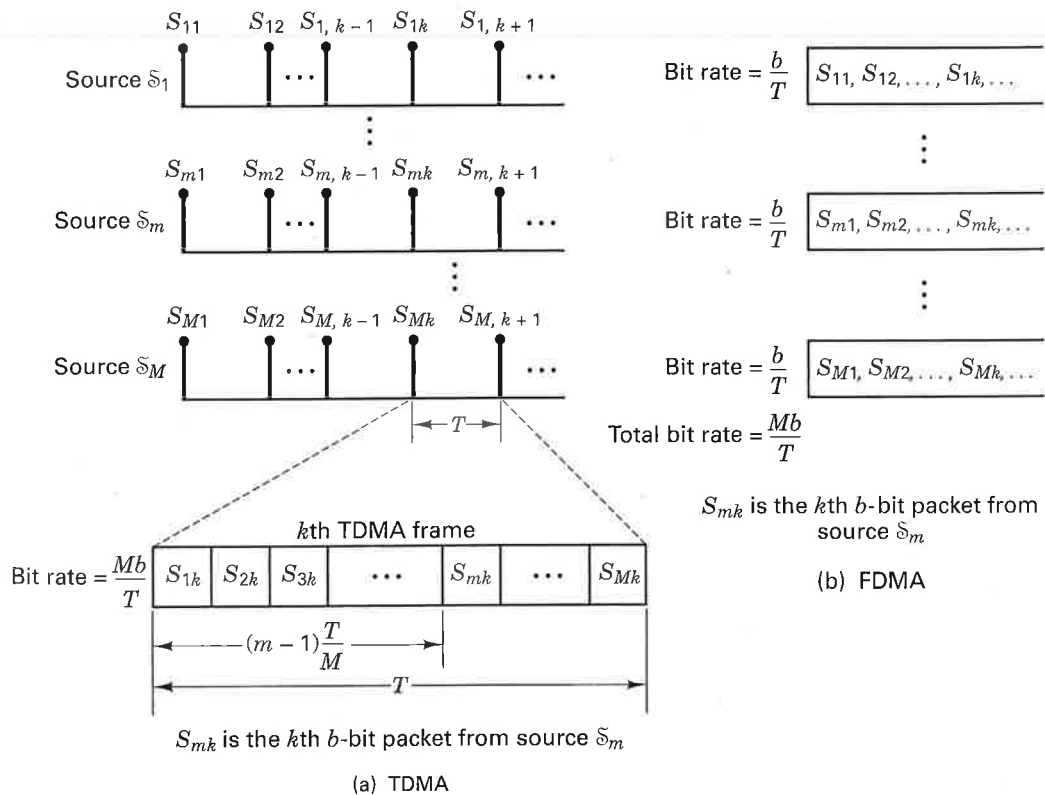


Figure 11.13 (a) TDMA and (b) FDMA channelization.

The maximum waiting time before transmission of a packet is $(M-1)T/M$ seconds, and on the average a packet will wait $\frac{1}{2}(M-1)(T/M) = (T/2)(1 - 1/M)$ seconds, as given by Equation (11.12).

To compare the average delay times, D_{FD} and D_{TD} , for FDMA and TDMA, respectively, we combine Equations (11.9) and (11.11) into Equation (11.8), and similarly combine Equations (11.10) and (11.12) into Equation (11.8), yielding

$$D_{FD} = T \quad (11.13)$$

$$D_{TD} = \frac{T}{2} \left(1 - \frac{1}{M} \right) + \frac{T}{M} = D_{FD} - \frac{T}{2} \left(1 - \frac{1}{M} \right) \quad (11.14)$$

Using Equation (11.7), Equation (11.14) can be written as

$$D_{TD} = D_{FD} - \frac{b}{2R} (M-1) \quad (11.15)$$

The result indicates that TDMA is inherently superior to FDMA, from a message-delay point of view. Although Equation (11.15) assumed that the data source is deterministic, the smaller average message delays for TDMA schemes hold up for any independent message arrival process [1, 2].

11.1.5 Code-Division Multiple Access

In Figure 11.3 the CR plane was illustrated as being shared by slicing it horizontally to form FDMA frequency bands, and in Figure 11.7 the same CR plane was illustrated as being shared by slicing it vertically to form TDMA time slots. These two techniques are the most common choices for multiple access applications. Figure 11.14 illustrates the CR being partitioned by the use of a hybrid combination of FDMA and TDMA known as *code-division multiple access* (CDMA). CDMA is an application of spread-spectrum (SS) techniques. Spread-spectrum techniques can be classified into two major categories: *direct-sequence* SS and *frequency hopping* SS. We introduce frequency hopping CDMA (FH-CDMA) in this chapter, and we treat direct-sequence CDMA together with the overall subject of spread-spectrum techniques in Chapter 12.

It is easiest to visualize *frequency hopping* CDMA, illustrated in Figure 11.14, as the short-term assignment of a frequency band to various signal sources. At each successive time slot, whose duration is usually brief, the frequency band assignments are reordered. In Figure 11.14, during time slot 1, signal 1 occupies band 1, signal 2 occupies band 2, and the signal 3 occupies band 3. During time slot 2, signal 1 hops to band 3, signal 2 hops to band 1, and signal 3 hops to band 2, and so on. The CR can thus be fully utilized, but the participants, having their frequency bands reassigned at each time slot, appear to be playing "musical chairs." Each user employs a pseudonoise (PN) code, orthogonal (or nearly orthogonal) to all the other user codes, that dictates the frequency hopping band assignments. Details of PN code sequences are treated in Section 12.2. Figure 11.14 is an oversimplified view of the way the CR is shared in frequency hopping CDMA, since the symmetry implies that each frequency hopping signal is in time synchronism with each of the

$7/M$ seconds,
seconds, as

and TDMA,
(11.8), and
yielding

$$(11.13)$$

$$(11.14)$$

$$(11.15)$$

a message-
ta source is
nes hold up

horizontally
e was illus-
s. These two
ions. Figure
bination of
CDMA is an
hiques can
icy hopping
ter, and we
ad-spectrum

Figure 11.14,
ces. At each
band assign-
pies band 1,
slot 2, signal
, and so on.
r frequency
airs." Each
al) to all the
s. Details of
ersimplified
e symmetry
each of the

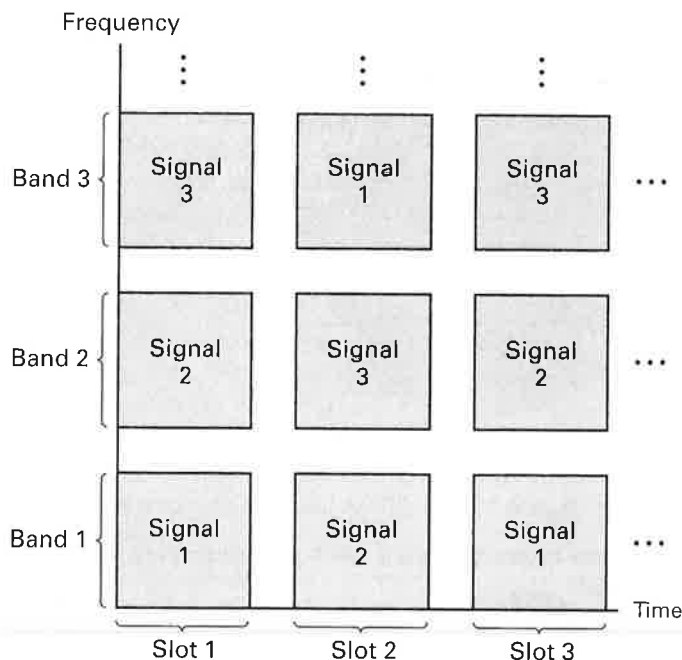


Figure 11.14 Code-division multiplexing.

other signals. This is *not the case*. In fact, one of the attractions of CDMA compared to TDMA is that there is no need for synchronization among user groups (only between a transmitter and a receiver within a group).

The block diagram in Figure 11.15 illustrates the frequency hopping modulation process. At each frequency hop time the PN generator feeds a code sequence to a device called a *frequency hopper*. The frequency hopper synthesizes one of the allowable hop frequencies. Assume that the data modulation has an M -ary frequency shift keying (MFSK) format. The essential difference between a conventional MFSK system and a frequency hopping (FH) MFSK system is that in the conventional system, a data symbol modulates a carrier wave that is *fixed* in frequency, but in the hopping system, the data symbol modulates a carrier wave that *hops* across the total CR bandwidth. The FH modulation in Figure 11.15 can be thought of as a two-step process—data modulation and frequency hopping modulation—even though it can be implemented in a single step, where the modulator produces a transmission tone based on the simultaneous dictates of the PN code and the data. Frequency hopping systems are covered in detail in Section 12.4.

One might ask: Don't the FDMA and TDMA options provide sufficient multiple access flexibility? FDMA and TDMA methods can surely be relied on to apportion the communications resource equitably. Of what use is this hybrid technique? CDMA offers some unique advantages, as follows:

1. *Privacy*. When the code for a particular user group is only distributed among authorized users, the CDMA process provides communications privacy, since

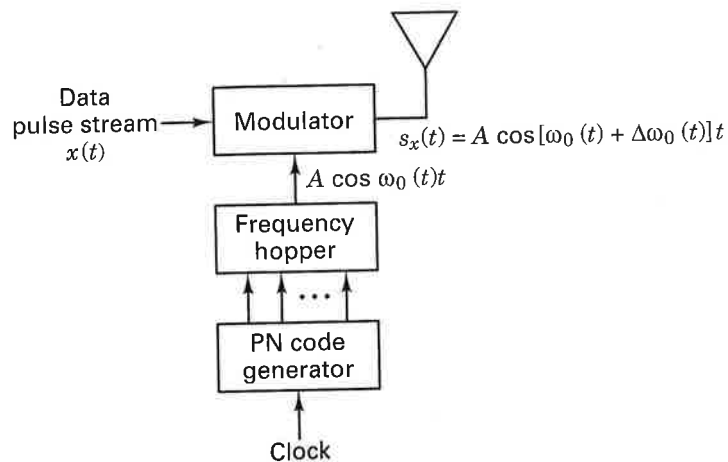


Figure 11.15 CDMA frequency hopping modulation process.

the transmissions cannot easily be intercepted by unauthorized users without the code.

2. *Fading channels.* If a particular portion of the spectrum is characterized by fading, signals in that frequency range are attenuated. In an FDMA scheme, a user who was unfortunate enough to be assigned to the fading position of the spectrum might experience highly degraded communications for as long as the fading persists. However, in a FH-CDMA scheme, only during the time a user hops into the affected portion of the spectrum will the user experience degradation. Therefore, with CDMA, such degradation is shared among all the users.
3. *Jam resistance.* During a given CDMA hop, the signal bandwidth is identical to the bandwidth of conventional MFSK, which is typically equal to the minimum bandwidth necessary to transmit the MFSK symbol. However, over a duration of many time slots, the system will hop over a frequency band which is much wider than the data bandwidth. We refer to this utilization of bandwidth as spread spectrum. In Chapter 12 we develop, in detail, the resistance to jamming that spread spectrum affords a user.
4. *Flexibility.* The most important advantage of CDMA schemes, compared to TDMA, is that there need be no precise time coordination among the various simultaneous transmitters. The orthogonality between user transmissions on different codes is not affected by transmission-time variations. This will become clear upon closer examination of the autocorrelation and cross-correlation properties of the codes, considered in Chapter 12.

11.1.6 Space-Division and Polarization-Division Multiple Access

Figure 11.16a depicts the INTELSAT IVA application of space-division multiple access (SDMA), also called *multiple-beam frequency reuse*. INTELSAT IVA used a dual-beam receive antenna feeding two receivers to allow simultaneous access of

sers without

ized by fading, a user of the spectrum as the fading a user hops degradation. users.

h is identical to the mini- ever, over a / band which ion of band- he resistance

compared to g the various missions on ns. This will n and cross-

sion multiple AT IVA used ous access of

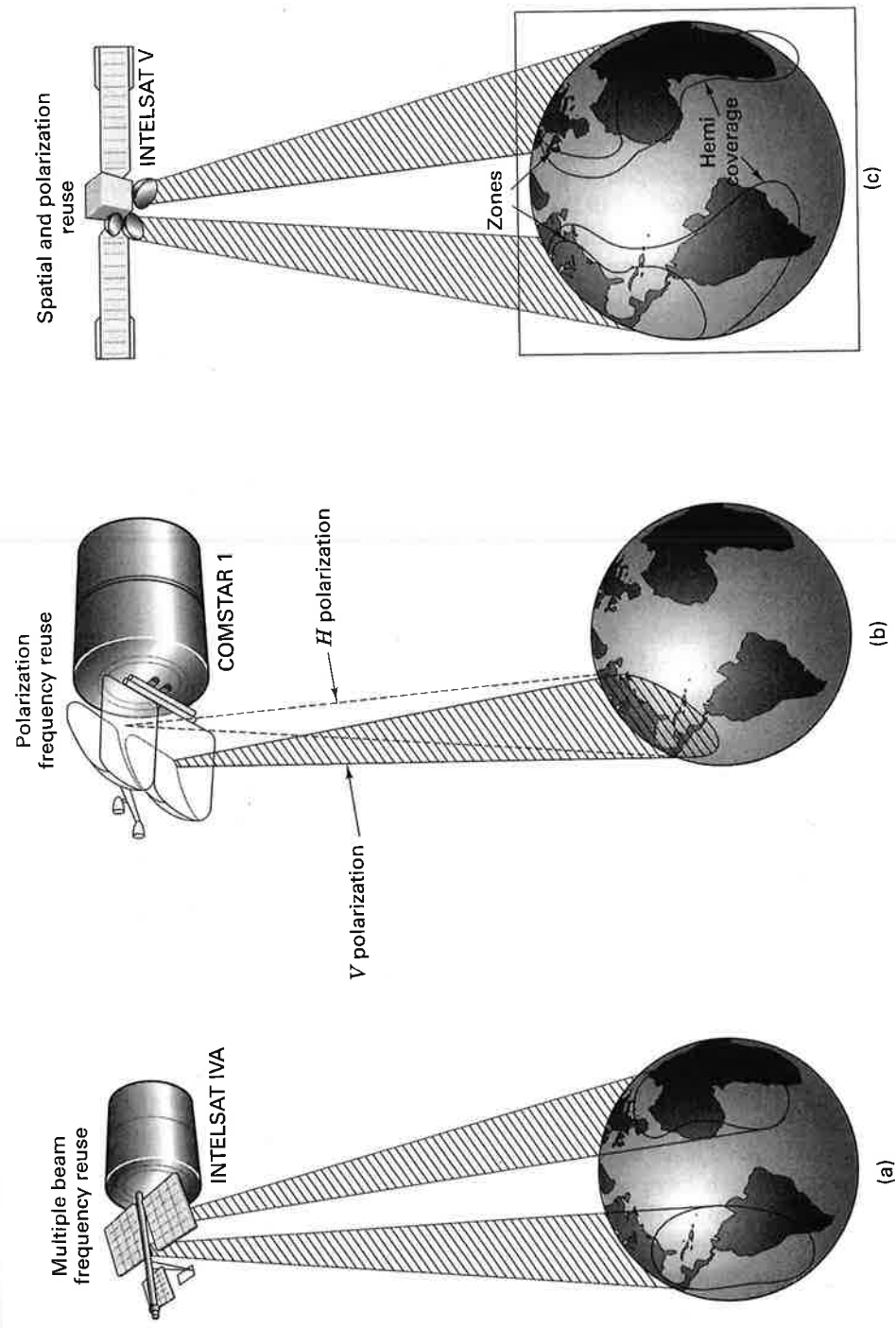


Figure 11.16 SDMA and PDMA. (a) INTELSAT IVA. (b) COMSTAR 1. (c) INTELSAT V (Atlantic coverage).

the satellite from two different regions of the earth. The frequency band allocated to each receive beam was identical because the uplink signals were spatially separated. In such cases, the frequency band is said to be *reused*.

Figure 11.16b depicts an application of polarization-division multiple access (PDMA), also called *dual-polarization frequency reuse*, from COMSTAR 1. Here separate antennas were used, each with different polarization and followed by separate receivers, allowing simultaneous access of the satellite from the same region of the earth. Each corresponding earth station antenna needs to be polarized in the same way as its counterpart in the satellite. (This is generally accomplished by providing each participating earth station with an antenna that has dual polarization.) The frequency band allocated to each antenna beam could be identical because the uplink signals were orthogonal in polarization. As with SDMA, the frequency band in PDMA is said to be reused. Figure 11.16c depicts an application of the simultaneous use of SDMA and PDMA in INTELSAT V. There are two separate hemispheric coverages, west and east. There are also two smaller zone beams; each zone beam overlaps a portion of one of the hemispheric beams and is separated from it by orthogonal polarization. Thus, there is a fourfold reuse of the spectrum.

11.2 MULTIPLE ACCESS COMMUNICATIONS SYSTEM AND ARCHITECTURE

A *multiple access protocol* or *multiple access algorithm* (MAA) is that rule by which a user knows how to use time, frequency, and code functions to communicate through a satellite to other users. A multiple access system is a combination of hardware and software that supports the MAA. The general goal of a multiple access system is to provide communications service in a timely, orderly, and efficient way.

Figure 11.17 illustrates some basic choices for the architecture of a satellite multiple access system. The legend indicates the symbols used for an earth station with and without an MAA controller, and a satellite with and without an MAA controller. Figure 11.17a illustrates the case where one earth station is designated as the master, or the controller. This earth station possesses an MAA computer and responds to the service requests of all other users. Notice that a user's request entails a transmission through the satellite and back down to the controller. The controller's response entails another transmission through the satellite; hence there are two up- and downlink transmissions required for each service assignment. Figure 11.17b illustrates the case where the MAA control is distributed among all the earth stations; there is no single controller. Each earth station uses the same algorithm and they each have identical knowledge regarding access requests and assignments; therefore, only one round trip is required for each service assignment. Figure 11.17c illustrates the case where the MAA controller is in the satellite. A service request goes from user to satellite, and the response from the satellite can follow immediately; therefore, only one round trip is required for each service assignment.

allocated
ally sepa-

ple access
R 1. Here
ed by sep-
ne region
zed in the
ed by pro-
rization.)
ecause the
ency band
e simulta-
ate hemi-
each zone
ed from it
n.

e by which
nmunicate
ination of
a multiple
derly, and

a satellite
rth station
: an MAA
designated
computer
r's request
roller. The
ence there
ment. Fig-
ong all the
same algo-
sts and as-
signment.
satellite. A
atellite can
ach service

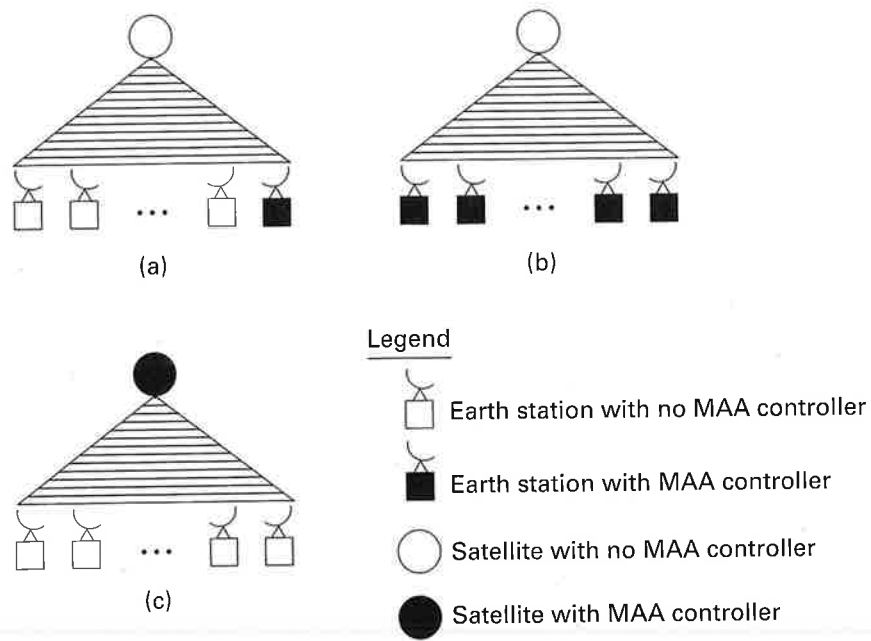


Figure 11.17 Satellite multiple access architecture. (a) Single earth station control. (b) Distributed earth station control. (c) Satellite control.

11.2.1 Multiple Access Information Flow

Figure 11.18 is a flow diagram describing the basic flow of information between the multiple access algorithm (MAA) or controller and an earth station; the numbers below correspond to those on the figure. Recall from the preceding section that the control may be lodged in the satellite, in a master station, or distributed among all the earth stations. The flow proceeds as follows:

1. *Channelization.* This term refers to the most general allocation information [e.g., channels 1 to N may be allocated for the Army and channels $(N+1)$ to M for the Navy]. This information seldom changes, and may be distributed to the earth stations by the use of a newsletter rather than via the communication system.
2. *Network state (NS).* This term refers to the state of the CR. A station is advised regarding the availability of the communications resource and where in the resource (e.g., time, frequency, code position) to transmit its service request(s).
3. *Service request.* Then the station makes its request(s) for service (e.g., allocation for m message slots).
4. Upon receipt of the service request(s), the controller sends the station a schedule regarding where and when to position its data in the CR.
5. The station transmits its data according to its assigned schedule.

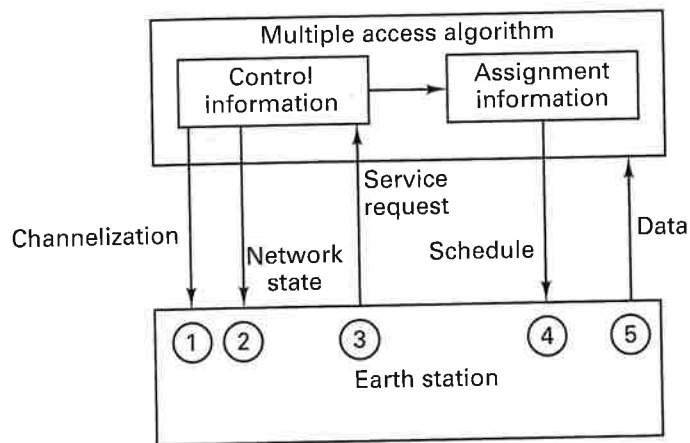


Figure 11.18 Multiple access information flow.

11.2.2 Demand-Assignment Multiple Access

Multiple access schemes are termed *fixed assignment* when a station has periodic access to the channel independent of its actual need. By comparison, dynamic assignment schemes, sometimes called *demand-assignment multiple access (DAMA)*, give the station access to the channel only when it requests access. If the traffic from a station tends to be burst-like or intermittent, DAMA procedures can be much more efficient than fixed-assignment procedures. A DAMA scheme capitalizes on the fact that actual demand *rarely* equals the peak demand. If a system's capacity is equal to the total peak demand and if the traffic is bursty, the system will be underutilized most of the time. However, by using buffers and DAMA, a system with reduced average capacity can handle bursty traffic, at the cost of some queueing delay. Figure 11.19 summarizes the difference between a fixed system, whose capacity is equal to the sum of the user requirements, and a dynamic system, whose capacity is equal to the average of the user requirements.

11.3 ACCESS ALGORITHMS

11.3.1 Aloha

In 1971, the University of Hawaii began operation of its ALOHA system. A communication satellite was used to interconnect the several university computers by use of a random access protocol [3–7]. The system concept was extremely simple, consisting of the following modes:

1. *Transmission mode.* Users transmit at any time they desire, encoding their transmissions with an error detection code.

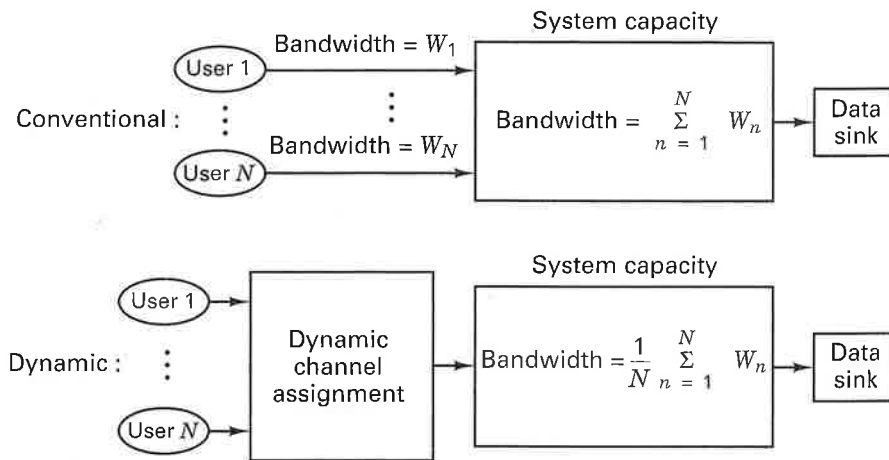


Figure 11.19 Bandwidth reduction for systems using dynamic channel assignment.

2. *Listening mode.* After a message transmission, a user listens for an acknowledgment (ACK) from the receiver. Transmissions from different users will sometimes overlap in time, causing reception errors in the data in each of the contending messages. We say that the messages have *collided*. In such cases, the errors are detected, and the users receive a negative acknowledgment (NAK).
3. *Retransmission mode.* When a NAK is received, the messages are simply retransmitted. Of course, if the colliding users were to retransmit immediately, they would collide again. Therefore, the users retransmit after a *random* delay.
4. *Timeout mode.* If, after a transmission, the user does not receive either an ACK or NAK within a specified time, the user retransmits the message.

11.3.1.1 Message Arrival Statistics

Assume that the total system demand requires an average message or packet arrival rate of λ successful or accepted messages per second. Because of the presence of collisions, some of the messages will be unsuccessful or rejected. Therefore, we define the total traffic arrival rate λ_t , as the acceptance rate λ , plus the rejection rate λ_r , as follows:

$$\lambda_t = \lambda + \lambda_r \quad (11.16)$$

Let us denote the length of each message or packet as b bits. Then we can define the average amount of successful traffic or *throughput*, ρ' , on the channel in units of bits per second, as

$$\rho' = b\lambda \quad (11.17)$$

We can also define the *total traffic*, G' , on the channel, in units of bits per second, as

$$G' = b\lambda_t \quad (11.18)$$

With the channel capacity (maximum bit rate) designated as R bits per second, let us further define a *normalized throughput*

$$\rho = \frac{b\lambda}{R} \quad (11.19)$$

and a *normalized total traffic*

$$G = \frac{b\lambda_t}{R} \quad (11.20)$$

Normalized throughput, ρ , expresses throughput as a fraction ($0 \leq \rho \leq 1$) of channel capacity. Normalized total traffic, G , expresses total traffic as a fraction ($0 \leq G \leq \infty$) of the channel capacity. Notice that G can take on values greater than unity.

We can also define the transmission time of each packet as

$$\tau = \frac{b}{R} \text{ seconds/packet} \quad (11.21)$$

By substituting Equation (11.21) into Equations (11.19) and (11.20), we can write

$$\rho = \lambda\tau \quad (11.22)$$

and

$$G = \lambda_t \tau \quad (11.23)$$

A user can successfully transmit a message as long as no other user began one within the previous τ seconds or starts one within the next τ seconds. If another user began a message within the previous τ seconds, its tail end will collide with the current message. If another user begins a message within the next τ seconds, it will collide with the tail end of the current message. Thus a space of 2τ seconds is needed for each message.

The message arrival statistics for unrelated users of a communication system is often modeled as a Poisson process. The probability of having K new messages arrive during a time interval of τ seconds is given by the Poisson distribution [8] as

$$P(K) = \frac{(\lambda\tau)^K e^{-\lambda\tau}}{K!} \quad K \geq 0 \quad (11.24)$$

where λ is the average message arrival rate. Because the users transmit without regard for each other in the ALOHA system, this expression is useful for calculating the probability that exactly $K = 0$ other messages are transmitted during a time interval 2τ . This is the probability, P_s , that a user's message transmission was successful (experienced no collisions). To compute P_s , assuming that all traffic is Poisson, we use λ_t and 2τ in Equation 11.24. Thus,

$$P_s = P(K=0) = \frac{(2\tau\lambda_t)^0 e^{-2\tau\lambda_t}}{0!} = e^{-2\tau\lambda_t} \quad (11.25)$$

In Equation (11.16) we defined the total traffic arrival rate λ_t in terms of the successful portion λ and the repetition or unsuccessful portion λ_r ; then, by definition, the probability of a successful packet can be expressed as

$$P_s = \frac{\lambda}{\lambda_t} \quad (11.26)$$

By combining Equations (11.25) and (11.26), we have

$$\lambda = \lambda_t e^{-2\tau\lambda_t} \quad (11.27)$$

By combining Equation (11.27) with Equations (11.22) and (11.23), we can write

$$\rho = G e^{-2G} \quad (11.28)$$

Equation (11.28) relates the normalized throughput, ρ , to the normalized total traffic, G , on the channel for the ALOHA system. A plot of this relationship labeled "pure ALOHA" is shown in Figure 11.20; as G increases, ρ increases until a point is reached where further traffic increases create a large enough collision rate to cause

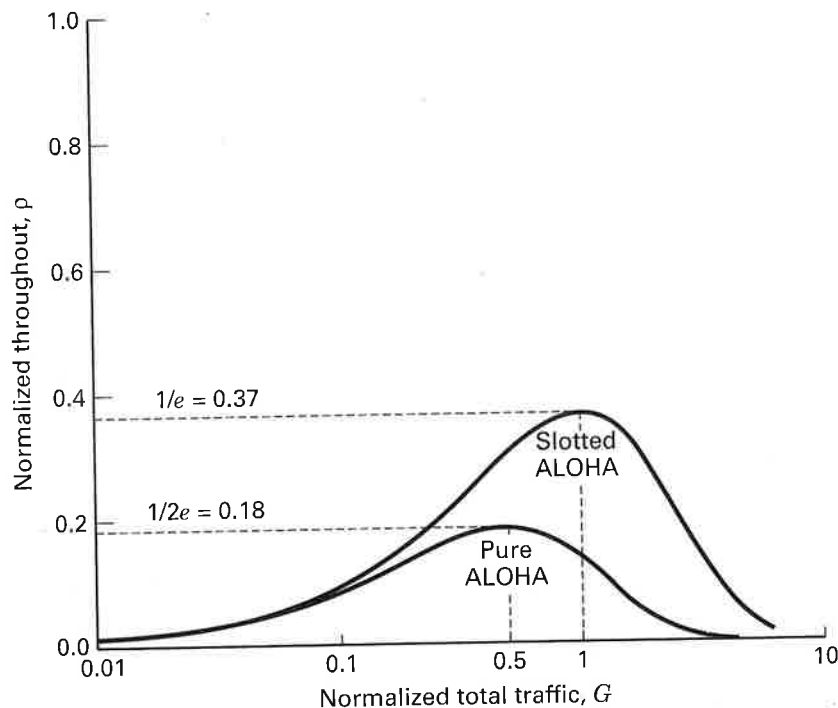


Figure 11.20 Throughput in ALOHA channels (successful transmissions versus total transmissions).

a reduction in the throughput. The maximum ρ , equal to $1/2e = 0.18$, occurs at a value of $G = 0.5$. Therefore, for a pure ALOHA channel, only 18% of the CR can be utilized. Simplicity of control is achieved at the expense of channel capacity [7, 9].

11.3.2 Slotted ALOHA

The pure ALOHA scheme can be improved by requiring a small amount of coordination among the stations. The slotted ALOHA (S-ALOHA) is such a system. A sequence of synchronization pulses is broadcast to all stations. As with pure ALOHA, packet lengths are constant. Messages are required to be sent in the *beginning* of a time slot. This simple change reduces the rate of collisions by half, since only messages transmitted in the same slot can interfere with one another. It can be shown [9, 10] that for S-ALOHA, the reduction in the *collision window* from 2τ to τ results in the following relationship between normalized throughput ρ and normalized total traffic G :

$$\rho = Ge^{-G} \quad (11.29)$$

The plot of Equation (11.29) is shown in Figure 11.20 labeled "slotted ALOHA." Here the maximum value of ρ is $1/e = 0.37$, or an improvement of two times the pure ALOHA protocol.

The retransmission mode described for the pure ALOHA system was modified for S-ALOHA so that if a negative acknowledgment (NAK) occurs, the user retransmits after a *random* delay of an integer number of slot times. Figure 11.21 illustrates the S-ALOHA operation. A packet of data bits is shown transmitted by user k followed by the satellite acknowledgment (ACK). Also shown are users m and n simultaneously transmitting packets, which results in a collision; a NAK is returned. Each using station employs a random-number generator to select its retransmission time. The figure illustrates an example of the m and n retrans-

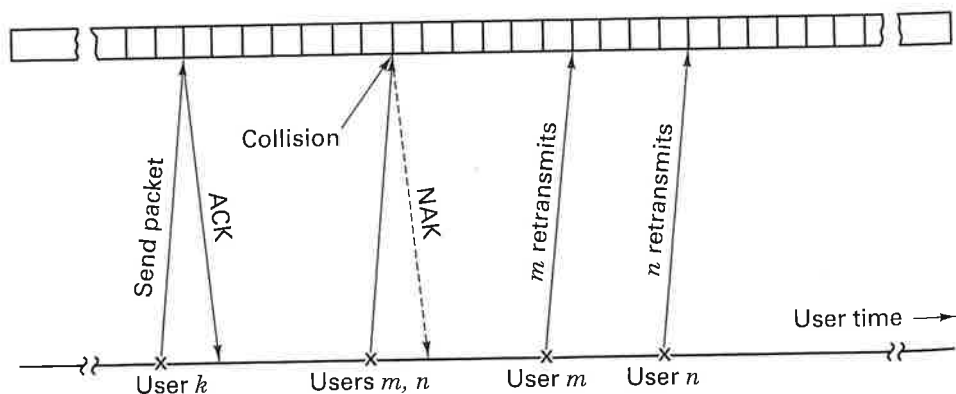


Figure 11.21 Random access scheme: slotted ALOHA operation.

mission at their respective randomly selected times. Of course, there is some probability that users m and n will recollide. However, in that case, they simply repeat the retransmission, using another random delay.

Example 11.1 Poisson Process

Assuming that packet transmissions and retransmission can both be described as a Poisson process, calculate the *probability* that a data packet transmission in an S-ALOHA system will experience a collision with *one other user*. Assume that the total traffic rate $\lambda_t = 10$ packets/s and the packet duration $\tau = 10$ ms.

Solution

$$\begin{aligned} P(K=1) &= \frac{(\tau\lambda_t)^K e^{-\tau\lambda_t}}{K!} \bigg|_{K=1} \\ &= (10 \times 0.01)^1 e^{-0.1} = 0.1e^{-0.1} \\ &= 0.09 \end{aligned}$$

11.3.3 Reservation-ALOHA

A significant improvement was made to the ALOHA system with the introduction of the reservation-ALOHA (R-ALOHA) [11] scheme. The R-ALOHA system has two basic modes: an unreserved mode and a reserved mode; each is described as follows:

Unreserved Mode (Quiescent State)

1. A time frame is established and divided into a number of small reservation subslots.
2. Users use these small subslots to reserve message slots.
3. After requesting a reservation, the user listens for an acknowledgment and a slot assignment.

Reserved Mode

1. The time frame is divided into $M+1$ slots whenever a reservation is made.
2. The first M slots are used for message transmissions.
3. The last slot is subdivided into subslots to be used for reservation/requests.
4. Users send message packets only in their assigned portions of the M slots.

Consider the R-ALOHA example shown in Figure 11.22. In the quiescent state, with no reservations, time is partitioned into short subslots for making reservations. Once a reservation is made, the system is configured so that $M=5$ message slots followed by $V=6$ reservations subslots becomes the timing format. The figure illustrates a request and an acknowledgment in progress. In this example the station seeks to reserve three message slots. The reservation acknowledgment advises the using station where to locate its first data packet. Since the control is distributed so that all participants receive the downlink transmissions and are thus aware

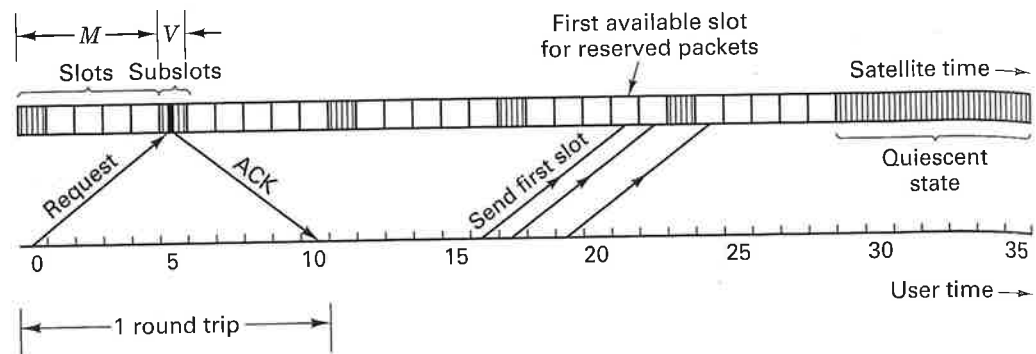


Figure 11.22 Example of reservation ALOHA. Station seeks to reserve three slots ($M = 5$ slots, $V = 6$ subslots).

of the reservations and time format, the acknowledgment need not disclose any more than the location of the first slot. As shown in Figure 11.22, the station sends its second packet in the slot following the first packet. The user further knows that the next slot is comprised of six subslots for reservations, so *no* packets are transmitted during this time. The third and final packet is sent in the following slot. When there are no reservations taking place, the system reverts back to its quiescent format of subslots only. Since the control is distributed, all the participants are made aware of the quiescent format by receiving appropriate synchronizing pulses on the downlink. Other interesting reservation schemes are discussed in Reference [12, 13].

11.3.4 Performance Comparison of S-ALOHA and R-ALOHA

From Chapters 3 and 4 the basic quality measure of a digital modulation scheme is its P_B versus E_b/N_0 curve. This measure is particularly useful because E_b/N_0 is a *normalized signal-to-noise ratio*; being normalized, the curves allow us to compare the performance of various modulation schemes. There is a similar performance measure for multiple access schemes. Here we are interested in the average delay versus normalized throughput. What would an *ideal delay-throughput curve* look like? Figure 11.23 illustrates such a curve. For normalized throughput values of, $0 \leq \rho < 1$, the delay equals zero until $\rho = 1$; then the delay increases without bound. Figure 11.23 also shows a *typical* delay-throughput curve and the direction in which the curve will move as delay performance improves.

Figure 11.24 compares the delay-throughput performance of S-ALOHA with that of R-ALOHA (formatted with two message slots and six reservation subslots). Knowing the location of the *ideal* curve it is easy to compare the delay performance of these two systems. For a throughput of less than approximately 0.20, the S-ALOHA manifests less average delay than does R-ALOHA. But for values of ρ between 0.20 and 0.67, it is apparent that R-ALOHA is superior, since the average delay is less. Why does the S-ALOHA perform better at low traffic intensity? The

te time →

 escent
 rate
 35
 er time →

iree

disclose any
 tation sends
 knows that
 are transmit-
 slot. When
 iescent for-
 ts are made
 ulses on the
 nce [12, 13].

on scheme is
 e E_b/N_0 is a
 to compare
 performance
 verage delay
 t curve look
 at values of,
 hout bound.
 ion in which

LOHA with
 on subslots).
 performance
 0.20, the S-
 r values of ρ
 the average
 tensity? The

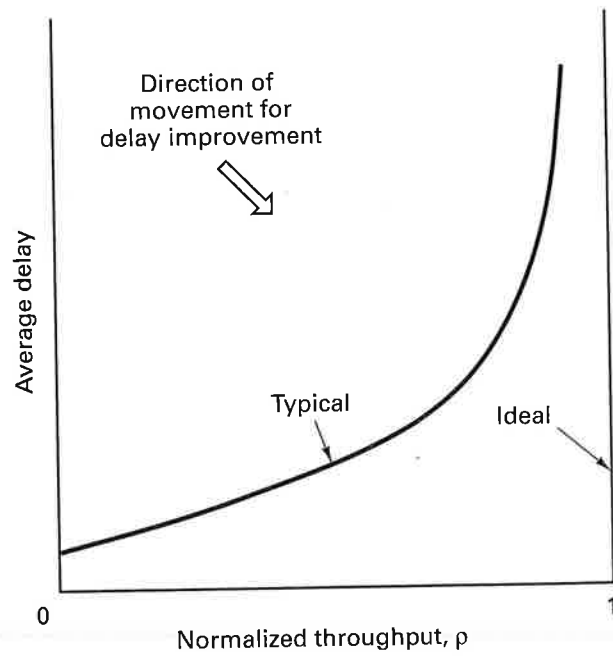


Figure 11.23 Delay-throughput characteristic.

S-ALOHA algorithm does not require the overhead of the reservation subslots as does R-ALOHA. Therefore, at low values of ρ , R-ALOHA pays the price of greater delay due to the greater overhead. For $\rho > 0.2$, the collisions and retransmissions inherent in the S-ALOHA system cause it to incur greater delay (unbounded at $\rho = 0.37$), more quickly than the R-ALOHA system. At higher throughput ($0.2 < \rho < 0.67$), the overhead structure of R-ALOHA ensures that its delay degradation grows in a more orderly manner than S-ALOHA. For R-ALOHA, an unbounded delay is not reached until $\rho = 0.67$.

Example 11.2 Channel Utilization

- (a) Normalized throughput, ρ , is a measure of channel utilization. It can be found by forming the ratio of the successfully transmitted message traffic, in bits per second to the total message traffic, including rejected messages, in bits per second. Calculate the normalized throughput of a channel that has a maximum data rate $R = 50$ kbits/s and operates with $M = 10$ ground stations, each station transmitting at the average rate of $\lambda = 2$ packets/second. The system format provides for $b = 1350$ bits/packet.
- (b) Which of the three ALOHA schemes discussed—pure, slotted, and reservation—could be successfully used with this channel?

Solution

- (a) Generalizing Equation (11.19) to allow for traffic from multiple stations, we have

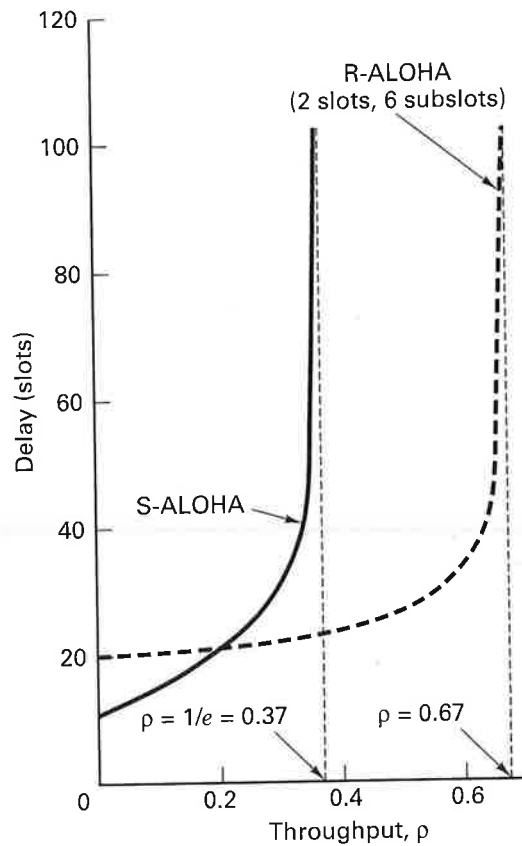


Figure 11.24 Delay-throughput comparison: S-ALOHA versus R-ALOHA on a satellite channel.

$$\begin{aligned}\rho &= \frac{Mb\lambda}{R} \\ &= \frac{10(1350)(2)}{50,000} \\ &= 0.54\end{aligned}$$

- (b) Only the R-ALOHA scheme could be used for this system, since with each of the other schemes, 54% of the resource cannot be utilized.

11.3.5 Polling Techniques

One way to impose order on a system with multiple users having random access requirements is to institute a controller that periodically polls the user population to determine their service requests. If the user population is large (e.g., thousands of terminals) and the traffic is bursty, the time required to poll the population can be an excessive overhead burden. One technique for rapidly polling a user population [4, 14] is called a *binary tree search*. Figure 11.25 illustrates a satellite example of such a tree search to resolve contention among users. In this example, assume that

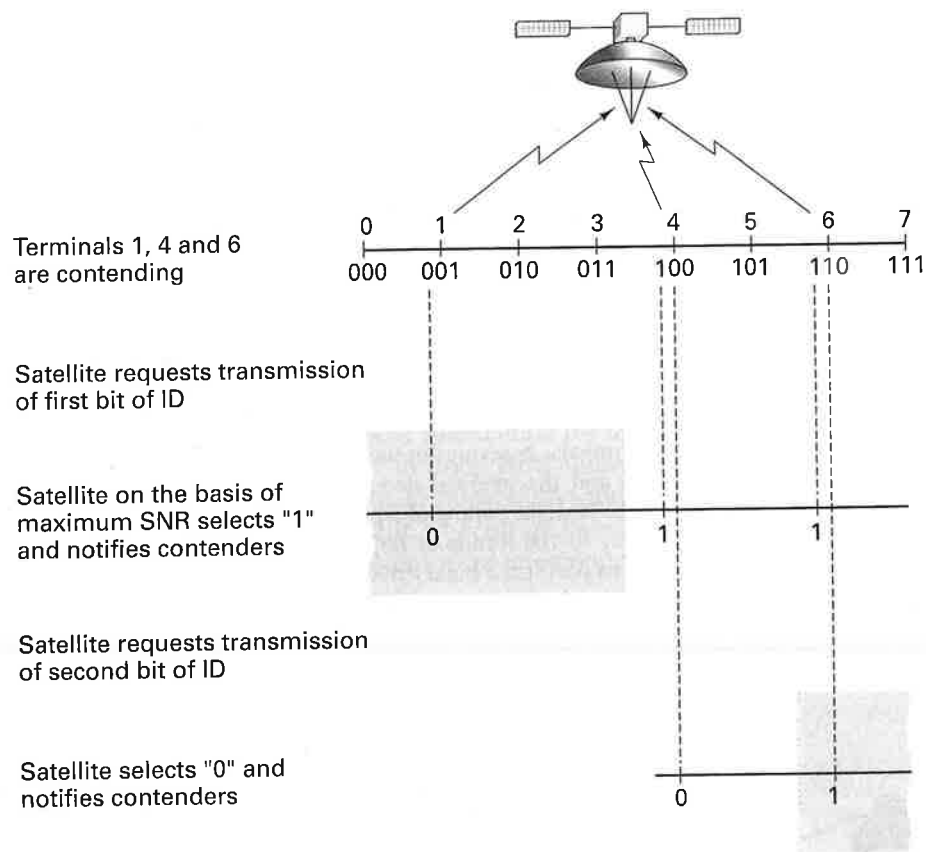


Figure 11.25 Tree search to resolve contention (eight terminal example).

the total user population is eight terminals; let them be identified by the binary numbers 000 to 111 as shown in Figure 11.25. Assume that terminals 001, 100, and 110 are contending for the service of a single channel. The tree search operates by continually partitioning the population until there is just a single branch remaining. The terminal corresponding to that branch is the "winner" and hence the first terminal to access the channel. The operation is repeated and again yields a single terminal that may next use the channel. The algorithm proceeds according to the following steps (see Figure 11.25):

1. The satellite requests the transmission of the contending terminals' first (left-most) bit of their identification (ID) numbers.
2. Terminal 001 transmits a zero, and terminals 100 and 110 each transmit a one. The satellite, on the basis of received signal strength, selects one or zero as the bit it "heard." In this example the satellite chooses binary one and in-

forms the users accordingly. Half the user population now knows that it has not been selected. The terminals in the "losing" half "bow out" of contention during this pass through the tree. In this example terminal 001 bows out.

3. The satellite requests the transmission of the second identifying bit from the remaining contending terminals.
4. Terminal 100 transmits a zero, and terminal 110 transmits a one.
5. Assume that the satellite selects the zero and notifies the contenders accordingly. Terminal 110 bows out. The process continues until it is clear that terminal 100 is free to access the satellite.
6. When the channel becomes available, steps 1–5 are repeated.

Example 11.3 Comparison between Binary Tree Search and Straight Polling

- (a) A binary tree search requires $n = \log_2 Q$ decisions for each pass through a population of Q terminals. A savings in time is possible with a tree search if the population is large and the average demand for service is small. Calculate the time needed for the straight polling of a population of 4096 terminals, to provide channel availability to 100 terminals requesting service. Compare the result with the time needed to perform a binary tree search 100 times, over the same population. Assume that the time required to poll one terminal and the time required for one decision of a binary tree search are each equal to 1 s.
- (b) Develop an expression for Q' , the largest number of terminals that results in the same (or less) time expended for binary tree searching as compared to straight polling.
- (c) Compute Q' for part (a).

Solution

- (a) Straight polling of 4096 terminals:

$$T = 4096 \times 1 \text{ s} = 4096 \text{ s}$$

Binary tree search for 100 terminals requires 100 passes through the binary tree:

$$T' = (100 \times \log_2 4096) \times 1 \text{ s} = 1200 \text{ s}$$

- (b) Q' is the maximum number of terminals that will result in $T' \leq T$ in part (a). This will occur when

$$Q'' \log_2 Q \times 1 \text{ s/decision} = Q \times 1 \text{ s/poll}$$

$$Q' = \lfloor Q'' \rfloor = \left\lfloor \frac{Q}{\log_2 Q} \right\rfloor \quad (11.30)$$

where $\lfloor x \rfloor$ is the largest integer no greater than x .

- (c) Q' for part (a)

$$Q' = \left\lfloor \frac{4096}{\log_2 4096} \right\rfloor = 341 \text{ terminals}$$

A binary tree search for 341 terminals entails a search time of 4092 s.

11.4 MULTIPLE ACCESS TECHNIQUES EMPLOYED WITH INTELSAT

The first commercial, geostationary communication satellite (INTELSAT I, or Early Bird), launched in 1965, represented the start of a new telecommunications era. Its 240 voice circuits provided more capacity than the undersea cables laid between the United States and Europe during the previous 10 years [15].

Early Bird featured a hard-limiting nonlinear transponder using FDMA. When several signals having different carrier frequencies simultaneously occupy a nonlinear device, the result is the production of intermodulation products which are signals at all combinations of sum and difference frequencies [16–18]. The energy apportioned to these intermodulation or IM products represents a *loss* in the useful signal energy. In addition, if these IM products appear within the bandwidth of other signals, the effect is that of added *noise* for the other signals.

The nonlinear transponder in Early Bird allowed for only two earth stations (one in the United States and one in Europe) to simultaneously access the satellite. Figure 11.26 illustrates this satellite's operation between the United States and Europe. Three European earth stations were interconnected via a terrestrial network.

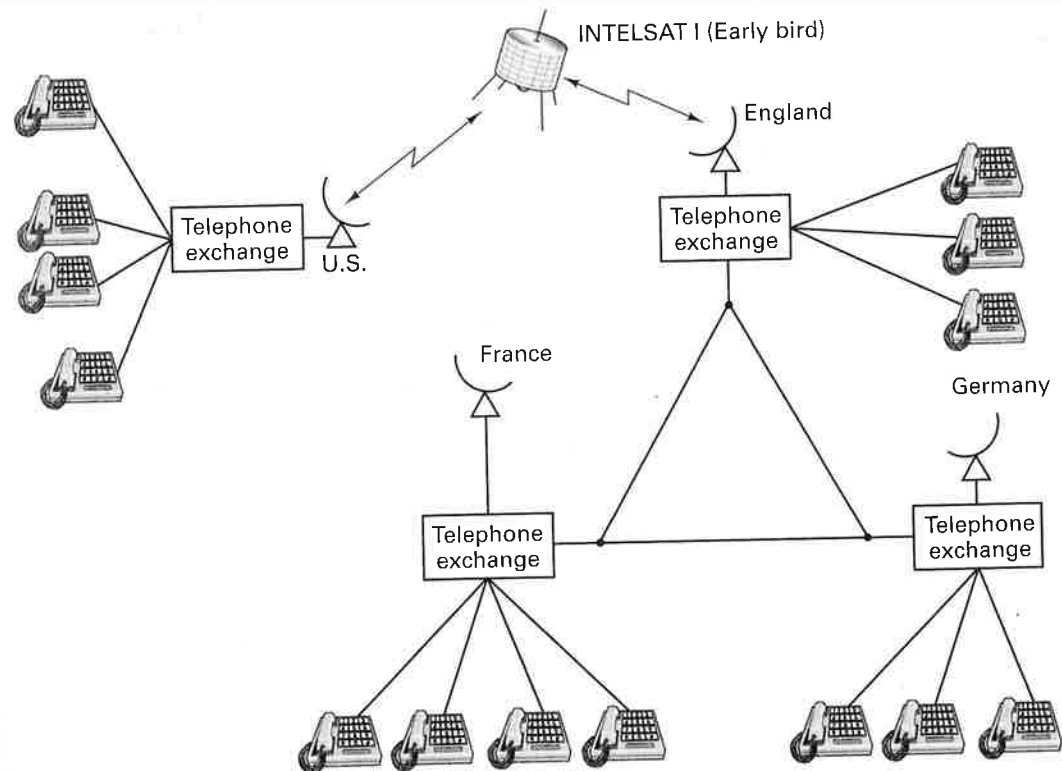


Figure 11.26 Early satellite operation.

Each month a different European station accessed the satellite and distributed the traffic to the other two stations.

11.4.1 Preassigned FDM/FM/FDMA or MCPC Operation

INTELSAT II and III improved multiple access capability by operating their travelling-wave tube amplifiers (TWTA) in the linear region instead of the hard-limiting region. This kept the IM products at an acceptable level, allowing more than two simultaneous accesses. (The price paid was a reduction in power amplifier efficiency.) Thus many FM carriers from various earth stations could simultaneously access these satellites. The operation is designated preassigned multidestination FDM/FM/FDMA or simply FDM/FM, or multichannel per carrier (MCPC), and is illustrated in Figure 11.27. Long-distance calls originating in country *A* enter the telephone exchange and are multiplexed into a supergroup (five groups of 12 voice circuits each). Country *A* transmits the supergroup on a single FM carrier at frequency f_A . Each group within the supergroup has been preassigned to an earth station in country *A* for telephone traffic destined to countries *B* through *F*. These countries each receive the signal on frequency f_A . The received signal is demodulated and demultiplexed at the destination country, selecting only those 12 channels preassigned to it.

11.4.2 MCPC Modes of Accessing an INTELSAT Satellite

INTELSAT has standardized the ways in which each 36-MHz transponder may be shared by specifying the occupied RF bandwidth and the number of 4-kHz channels per user. Some of these standard channels are shown in Table 11.1. Notice that the capacity of the transponder (last column in Table 11.1) drops as the number of carriers increases. The reasons are as follows:

1. Guard bands are needed between carrier bands; the more carriers there are, the more guard bands are needed. Hence capacity is reduced.
2. Multiple carriers in the nonlinear TWTA cause intermodulation (IM) products. If the TWTA is backed off into the linear region to reduce interference, the TWTA can provide less overall power. The channel becomes power limited and can service fewer carriers.

Table 11.1 indicates that a single carrier provides the most efficient use of the transponder. Why doesn't INTELSAT always operate its transponders in this mode? The answer is that not all earth stations have enough traffic to justify the assignment of an entire 36-MHz transponder. The other modes are needed so that various combinations of stations having less traffic will be able to share a transponder.

11.4.2.1 Bandwidth-Limited versus Power-Limited Conditions

In the preceding section it was stated that the backed-off transponder cannot support as many channels as the fully saturated transponder. It is useful to examine the two extreme transponder conditions, bandwidth limited and power limited, in

istributed the

operating their
1 of the hard-
ing more than
amplifier effi-
multaneously
ultidestination
ACPC), and is
enter the tele-
2 voice circuits
uency f_A . Each
country A for
ich receive the
tiplexed at the

ponder may be
of 4-kHz chan-
..1. Notice that
the number of

riers there are,

ion (IM) prod-
e interference,
ecomes power

cient use of the
rs in this mode?
the assignment
so that various
isponder.

sponder cannot
eful to examine
ower limited, in

ess Chap. 11

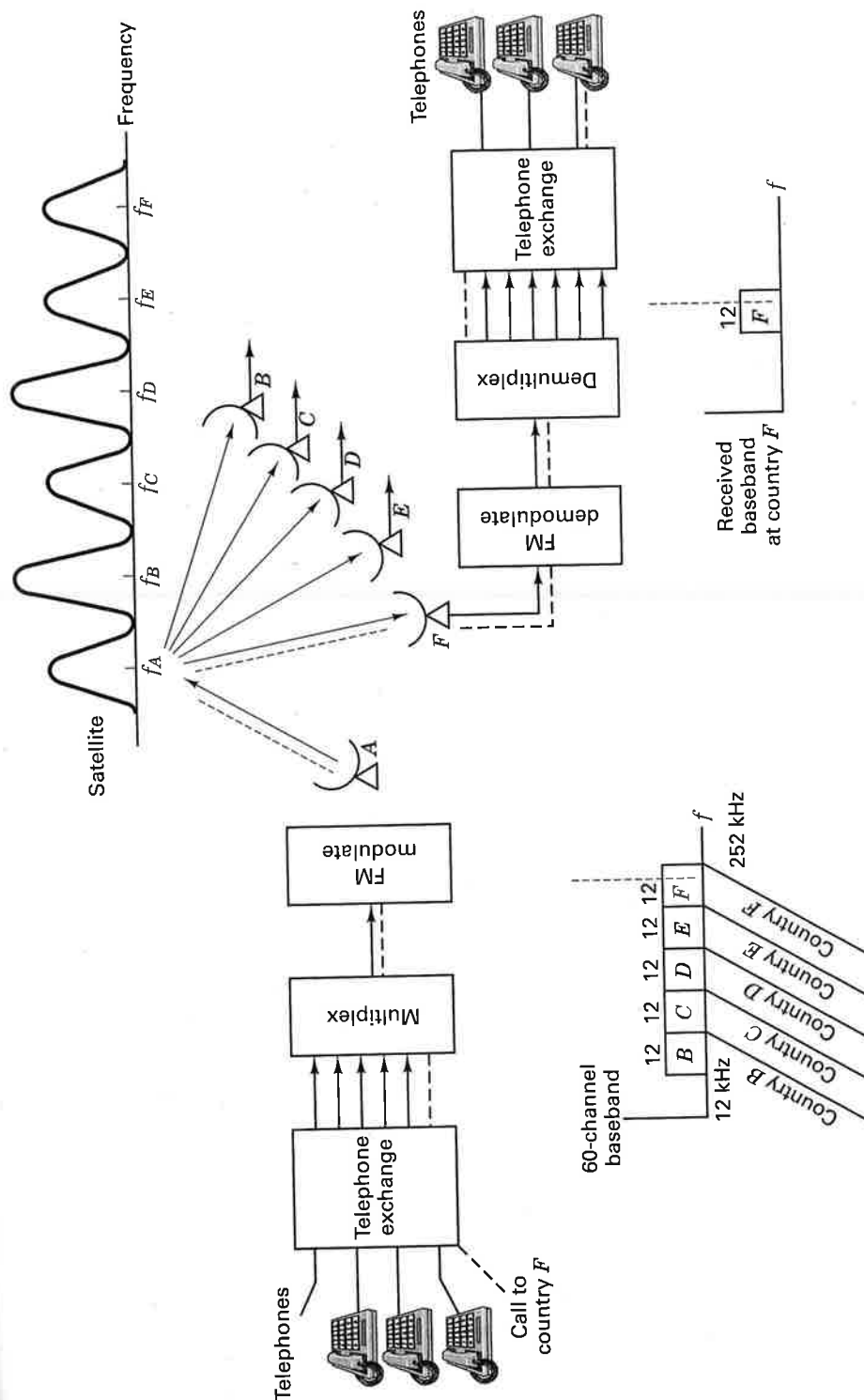


Figure 11.27 Preassigned multidestination FDM/FM carriers. (Reprinted with permission from J. G. Puente and A. M. Werth, "Demand-Assigned Service for the INTELSAT Global Network," *IEEE Spectrum*, Jan. 1971. © 1971 IEEE.)

TABLE 11.1 Standard INTELSAT MCPC Accessing Modes

Number of Carriers per Transponder	Carrier Bandwidth	Number of 4-kHz Channels per Carrier	Number of 4-kHz Channels per Transponder
1	36 MHz	900	900
4	3 at 10 MHz 1 at 5 MHz	132 60	456
7	5 MHz	60	420
14	2.5 MHz	24	336

the context of a satellite transponder. In Figure 11.28 we assume a 36-MHz transponder with a maximum power output of 20 W. Figure 11.28a illustrates an MCPC mode of operation whereby four carriers share the 36-MHz bandwidth. Assume that each carrier requires 4 W. The total output power is 16 W (less than the maximum capability of the amplifier); therefore, there is still power to spare. However, should another user want to access the transponder, the total 36-MHz bandwidth has already been allocated to the existing four carriers; there is no additional bandwidth to spare. Figure 11.28a illustrates this bandwidth-limited case.

Suppose that the previous example results in the production of serious IM products at the transponder. Assume that it is necessary to linearize the transponder by operating it at a reduced maximum power output of 12 W. With only a 12-W capability, the transponder can no longer support four users with 4 W each. One of the users must be “thrown off,” as illustrated in Figure 11.28b. Therefore, we have bandwidth to support another user, but not sufficient power. Figure 11.28b illustrates this power-limited case.

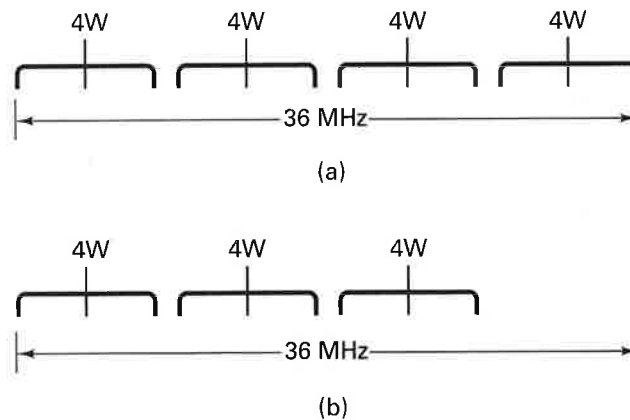


Figure 11.28 Bandwidth-limited versus power-limited configurations. (a) Bandwidth-limited example. (b) Power-limited example.

11.4.3 SPADE Operation

The preassigned MCPC multiple access scheme is very efficient when the traffic is heavy enough so that the channels are most always filled. However, if out of a 12-channel group, only one channel is active, the other 11 cannot be turned off. The FDM/FM transmission is made with or without actual telephone traffic on the channels. Therefore, the long-term preassignment of carriers to stations having light traffic is wasteful. Since there are many light traffic links, a flexible method to service them was needed. Also, an efficient way to handle overflow traffic from medium-capacity preassigned links was needed. Such was the motivation for a novel DAMA scheme known as SPADE, first used with INTELSAT IV. The acronym SPADE stands for "single-channel-per-carrier PCM multiple access demand assignment equipment." The principal features characterizing SPADE operation [15] are the following:

1. A single voice-grade channel is analog-to-digital (A/D) converted at a bit rate of 64 kbits/s.
2. This baseband digital signal modulates a carrier using quadrature phase shift keying (QPSK). Unlike the MCPC case, there is *only one* voice channel per carrier.
3. The channel spacing is 45 kHz. Within a transponder, there is bandwidth available for 800 channel carriers. Six carrier positions are vacant by design; thus there are 794 usable carriers.
4. The carrier is dynamically assigned, *upon demand*.
5. The dynamic assignment is accomplished over a 160-kHz common signaling channel (CSC) used as an "order-wire" or control circuit. The bit rate on the CSC is 128 kbits/s, and the modulation is binary phase shift keying (BPSK).

Figure 11.29 illustrates the frequency allocations for the CSC and the 800 carriers in the SPADE system. The SPADE operation can best be understood with the aid of Figure 11.30. The CSC operates in a fixed-assignment TDMA broadcast mode; that is, all earth stations monitor the CSC and are aware of the current state of channel assignments. Each earth station has a 1-ms time slot on the CSC (once every 50 ms) for requesting or releasing a channel. When an earth station needs a channel, it "seizes" a free one by requesting a frequency pair at random and transmitting its selection on the CSC. Random selection makes it unlikely that two stations will simultaneously request the same channel unless there are very few remaining. As soon as the channel is allocated, each of the other earth station processors deletes it from its list of available channels. The list is continually kept updated via the CSC. Thus, control of the SPADE access scheme is *distributed* among all the participating earth stations.

When the station finishes with the channel, the station indicates the channel's release by transmitting a signal in its time slot on the CSC. Each station receives this signal and designates the released channel as available. If two stations simultaneously seize the same channel, they each get a "busy" indication. They try again, selecting at random from the pool of available channels.

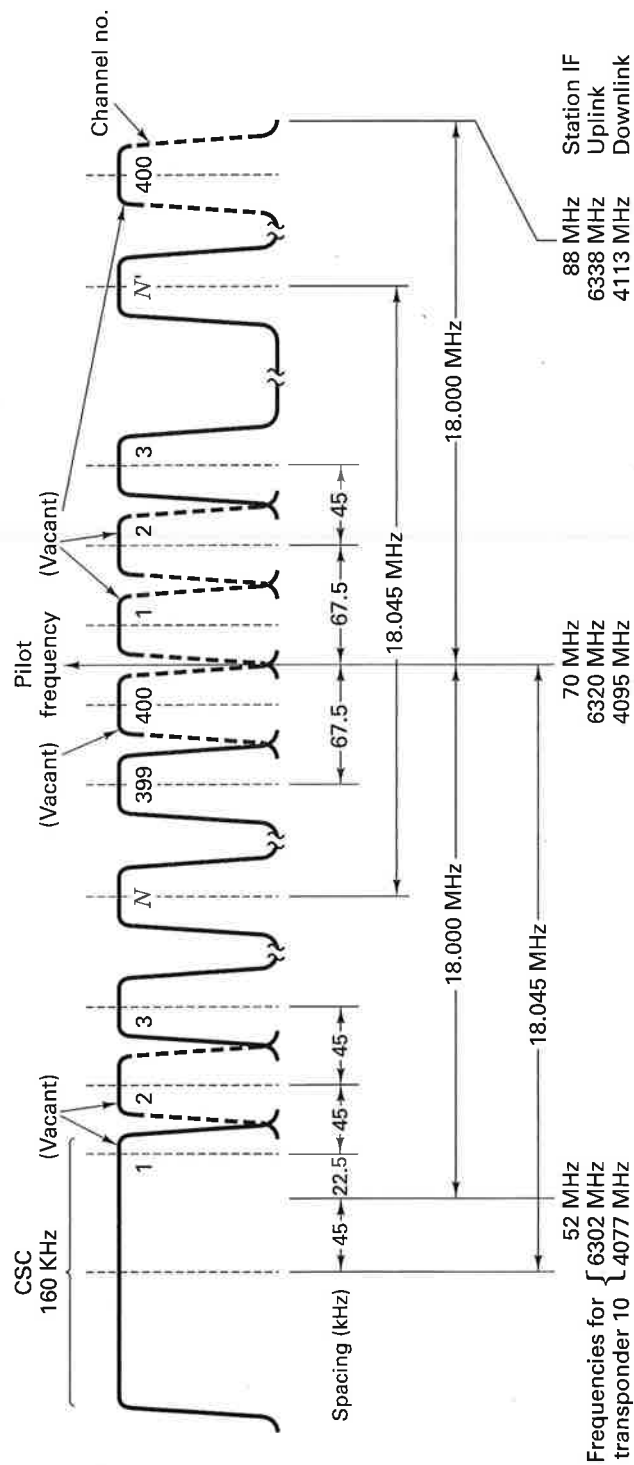


Figure 11.29 SPADE frequency allocations. (Reprinted with permission from J. G. Puente and A. M. Werth, "Demand-Assigned Services for the INTELSAT Global Network," *IEEE Spectrum*, Jan. 1971. © IEEE.)

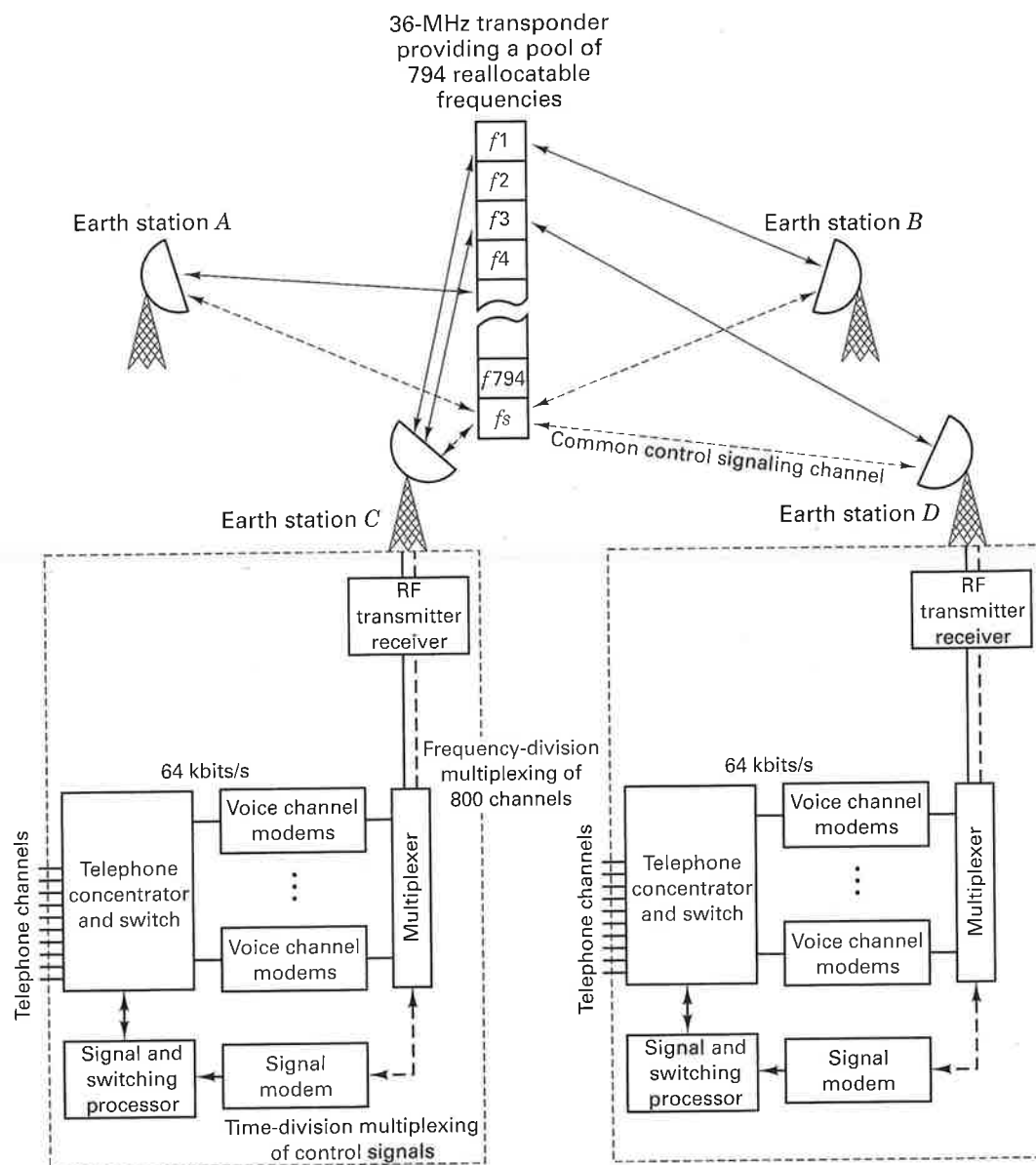


Figure 11.30 SPADE operation. (From James Martin, *Communications Satellite Systems*, © 1978, Fig. 15.2, p. 236. Reprinted by permission of Prentice-Hall, Englewood Cliffs, N.J.)

11.4.3.1 Transponder Capacity Utilization with SPADE

Table 11.2 is a continuation of Table 11.1. We see that the transponder bandwidth utilization with SPADE results in a total capacity of 800 voice channels per transponder. Compare Table 11.2 with Table 11.1. In Table 11.1, as the number of carriers increases from 1 to 14, the total number of channels decreases from 900 to 336. Why doesn't the SPADE system in Table 11.2 exhibit less capacity than the 336 channels associated with 14 carriers? The improved utilization comes about as follows. When there is only one voice channel per carrier, the carrier can be switched off when no speech is detected. Even with all channels operating, they can be switched off approximately 60% of the time. The transponders are power limited; power savings means that more channels can be transmitted. Also, SPADE uses digital voice transmission (QPSK); the bandwidth efficiency of the system is commensurate with the single-carrier FDM/FM case.

11.4.3.2 SPADE Efficiency

With MCPC, capacity is preassigned; a station's unused channels cannot be reallocated to other stations. SPADE is a DAMA system where all channels are shared. The channels are allocated to users as needed. An important telephone system quality measure, called the probability of blocking, is the probability that a requested circuit is not available. To achieve 1% probability of blocking requires four times as many MCPC channels as SPADE channels. A SPADE transponder with 800 channels is equivalent to 3200 MCPC channels [15].

11.4.3.3 Mixed-Size Earth Station Network Using SPADE

A standard-size INTELSAT earth station has a receiver sensitivity $G/T^\circ = 40.7$ dB/K, whereas the smaller size stations have a $G/T^\circ = 35$ dB/K. If 125 SPADE channels are destined for small stations, the total transponder capacity of 800 standard channels is reduced to 525 channels. This is the point at which half the available power is used to service the standard stations. The relationship between transponder capacity and channels allocated to small stations is shown in Figure 11.31. An explanation of this relationship can best be seen in Figure 11.32. When the total TWTA power provides service to large stations, Figure 11.32a illustrates that the 36-MHz bandwidth transponder is occupied by approximately 800 carriers each at a power level of x dBW (the bandwidth-limited case). When half the power is required to service small stations, Figure 11.32b illustrates that 400 carriers (half of the original 800) each at a power level of x dBW are reserved for the standard stations. Consider what happens to the remaining 400 carrier positions. From Chapter 5 we know that the error performance of a link is directly related to the

TABLE 11.2 SPADE Accessing

Number of Carriers per Transponder	Carrier Bandwidth	Number of 4-kHz Channels per Carrier	Numbers of 4-kHz Channels per Transponder
800	45 kHz	1	800

ansponder band-
ice channels per
is the number of
ases from 900 to
apacity than the
1 comes about as
e carrier can be
erating, they can
s are power lim-
1. Also, SPADE
of the system is

nnels cannot be
all channels are
ortant telephone
robability that a
locking requires
DE transponder

nsitivity $G/T^\circ =$
2. If 125 SPADE
acity of 800 stan-
ch half the avail-
onship between
shown in Figure
ure 11.32. When
11.32a illustrates
tely 800 carriers
n half the power
400 carriers (half
for the standard
positions. From
ly related to the

s of 4-kHz
r Transponder
300

ess Chap. 11

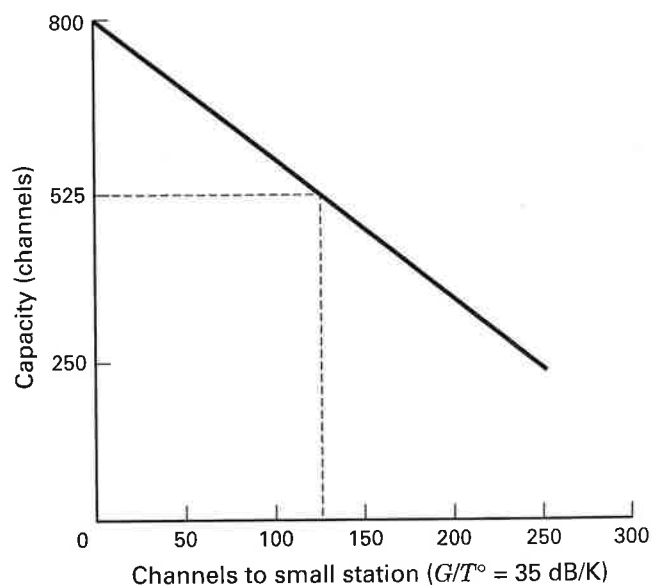


Figure 11.31 SPADE transponder capacity in a mixed-size earth station network.

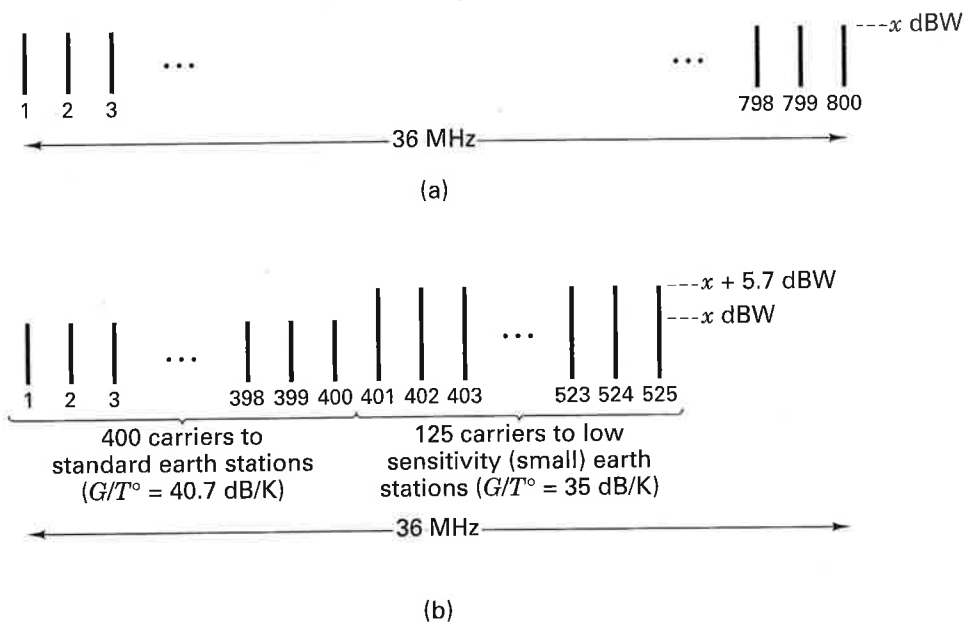


Figure 11.32 Mixed-size earth station network. (a) When the total TWTA power services large stations: bandwidth limited (800 carriers). (b) When half the TWTA power services small stations: power limited (525 carriers).

product of EIRP and G/T° . For any link, one can trade off these two parameters, thereby maintaining a fixed level of performance. Since the small station has a G/T° of 5.7 dB less than that of the standard station, it is necessary to supply the small station with 5.7 dB more EIRP for equivalent performance. The carrier power is *increased* by approximately 5.7 dB for each small station, thus the quantity of the remaining carriers serving these small stations is *decreased* by a similar amount. Therefore instead of 400 carriers, 125 (a reduction of 5.1 dB) are used to serve the small stations; the transponder is now power limited.

At the time a channel is assigned to a call, the transmitting station is apprised of the size of the destination station. Recall that these satellites are nonregenerative so that the apportionment of downlink EIRP is established by the transmitting station (see Section 5.7.1). The transmitting station sets its power level according to the needs of the receiving station.

11.4.4 TDMA in INTELSAT

The first generation of multiple access communication systems has been dominated by FDMA systems. The trend, however, is now in favor of TDMA systems, made possible by the availability of precise clocks and high-speed switching elements [19–24]. INTELSAT IV used a 128-kbits/s TDMA scheme for the common signaling channel that controls the SPADE network. Intelsat V introduced a 120-Mbits/s TDMA scheme for multiple-beam international digital service. One disadvantage or cost in implementing a TDMA scheme is the need for providing precise *synchronization* among the participating earth stations and the satellite. FDMA systems, not having such requirements, are less complex from a networking point of view. Comparisons of TDMA versus FDMA operation are summarized as follows:

1. FDMA can cause IM products. This can be avoided by operating the TWTA in its linear region, thereby reducing the available power output.
2. With TDMA, there is only one carrier present at a time in the TWTA. Thus, IM distortion cannot occur.
3. TDMA earth station equipment is more sophisticated and hence more costly than FDMA equipment. However, for earth stations providing multiple point-to-point channels, FDMA stations require separate radio-frequency (RF) up-conversion and down-conversion signal processing stages. Thus, with FDMA, the amount of equipment grows with the amount of simultaneous connectivity. With TDMA, such growth does not take place since channel selectivity is accomplished in time rather than frequency. Therefore, for a large multiply connected earth station, TDMA can be more cost-effective than FDMA.
4. In multiple-beam systems, each beam may need to communicate with every other beam. TDMA lends itself to conveniently forming connections sequentially as in satellite-switched TDMA (SS/TDMA). INTELSAT VI uses such satellite-switched TDMA (SS/TDMA), described in Section 11.4.5.

An example of the comparative performance of TDMA, FDM/FM, and SPADE is shown for an INTELSAT IV transponder as a plot of channel capacity versus earth station G/T° in Figure 11.33. Figure 11.33a is for an earth coverage antenna, and Figure 11.33b is for a spot-beam antenna. From synchronous altitudes these antennas have half-power beamwidths of 17° and 4.5° , respectively. From these plots it is seen that single-carrier FDM/FM is as efficient as TDMA when the system is operated with standard earth stations ($G/T^\circ = 40.7$ dB/K). For smaller earth stations ($G/T^\circ \leq 31$ dB/K) working through earth-coverage transponders, SPADE is more efficient than TDMA and multicarrier FDM/FM (MCPC); only the four-carrier case is plotted. For earth stations having G/T° in the range 19 to 40.7 dB/K working through a spot-beam transponder, TDMA is superior to SPADE and MCPC. For smaller earth stations having G/T° in the range 6 to 19 dB/K working through a spot-beam transponder, SPADE is superior to TDMA and MCPC. In general, when working through *standard* earth stations it is seen [19] that TDMA was the most efficient multiple access scheme for INTELSAT IV.

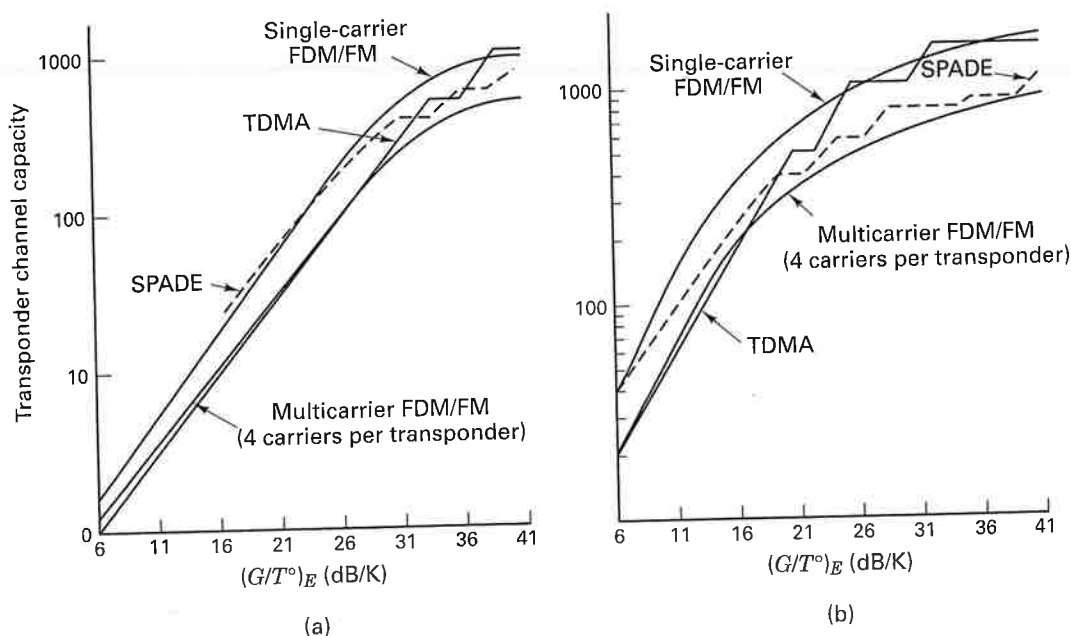


Figure 11.33 Channel capacity versus earth station G/T° for FDMA, TDMA, and SPADE. (a) Global-beam transponder channel capacity as function of $(G/T^\circ)_E$, where $(G/T^\circ)_E$ means earth station G/T° . (b) Spot-beam transponder channel capacity as function of $(G/T^\circ)_E$. [From D. Chakraborty, "INTELSAT IV Satellite System (Voice) Channel Capacity versus Earth Station Performance," *IEEE Trans. Commun. Tech.*, vol. COM19, no. 3, June 1971, pp. 355–362. © 1971 IEEE.]

11.4.4.1 PCM Multiplex Frame Structures

There are two digital telephony standards for PCM frame structures in operation. The North American standard is called *T-Carrier*; it is built around the 193-bit frame shown in Figure 11.34a. There are 24 channels; each channel contains an 8-bit voice sample. Also, there is one bit per frame with alternating value 1 0 1 0 . . . from frame to frame, used for frame alignment. Since a voice-grade telephone channel has a bandwidth of $W = 4$ kHz (including guard bands), the Nyquist sampling rate for recovering the analog information within 4 kHz is $f_s = 2W = 8000$ samples/s. Therefore, the basic PCM frame, called the *Nyquist frame*, which contains 24 voice samples from 24 different message sources, has a frame rate of 8000 frames/s (duration of 125 μ s). Thus the basic T-Carrier bit rate is 193 bits/frame $\times$ 8000 frames/s = 1.544 Mbits/s.

The European standard is built around a 256-bit frame shown in Figure 11.34b. There are 30 message channels, each containing an 8-bit voice sample. Also, one 8-bit time slot is used for frame alignment and another 8-bit time slot is used for signaling (addressing) information. The European frame rate is the same

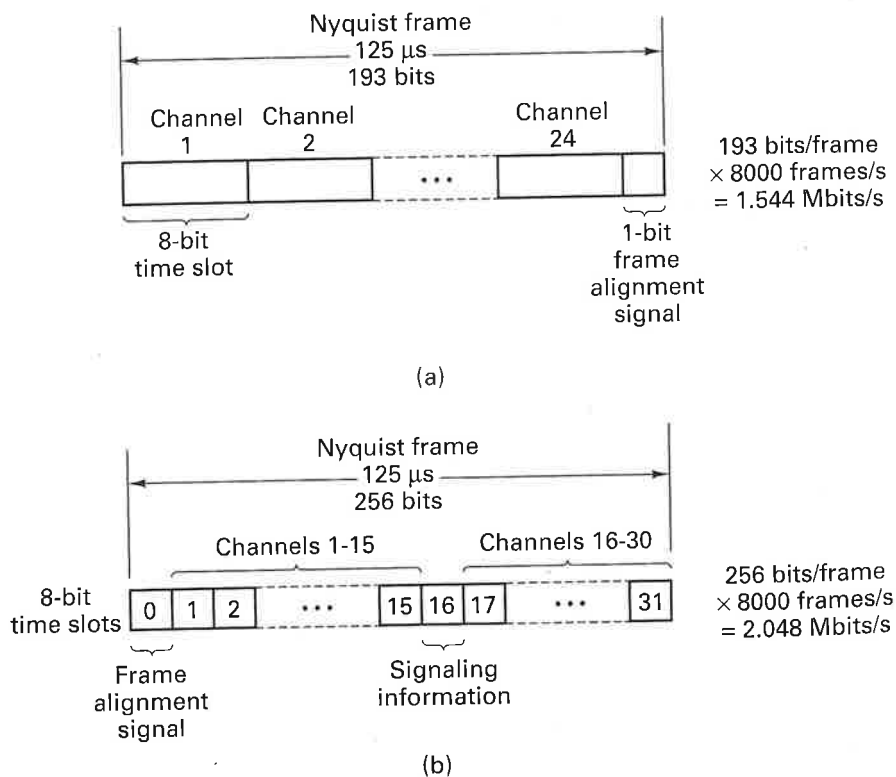


Figure 11.34 PCM multiplex frame structure. (a) Frame structure for T-Carrier (North American) PCM multiplex. (b) Frame structure for the European PCM multiplex.

as that of the T-Carrier. Therefore, the basic European bit rate is $256 \text{ bits/frame} \times 8000 \text{ frames/s} = 2.048 \text{ Mbits/s}$.

11.4.4.2 The High-Rate TDMA Frame for Europe

Sixteen Nyquist frames of the European PCM Multiplex format are shown in Figure 11.35a. Each frame contains an 8-bit sample from each of 30 terrestrial channels, plus 8 bits of framing and 8 bits of signaling information. The TDMA frame duration is

$$16 \text{ Nyquist frames} \times 125 \mu\text{s/Nyquist frame} = 2 \text{ ms}$$

Within this 2-ms frame are maintained

$$16 \text{ Nyquist frames} \times 256 \text{ bits/Nyquist frame} = 4096 \text{ bits}$$

The basic idea behind TDMA is that a user's low-rate data stream can share the CR with similar streams from other users by *bursting* the transmission at a

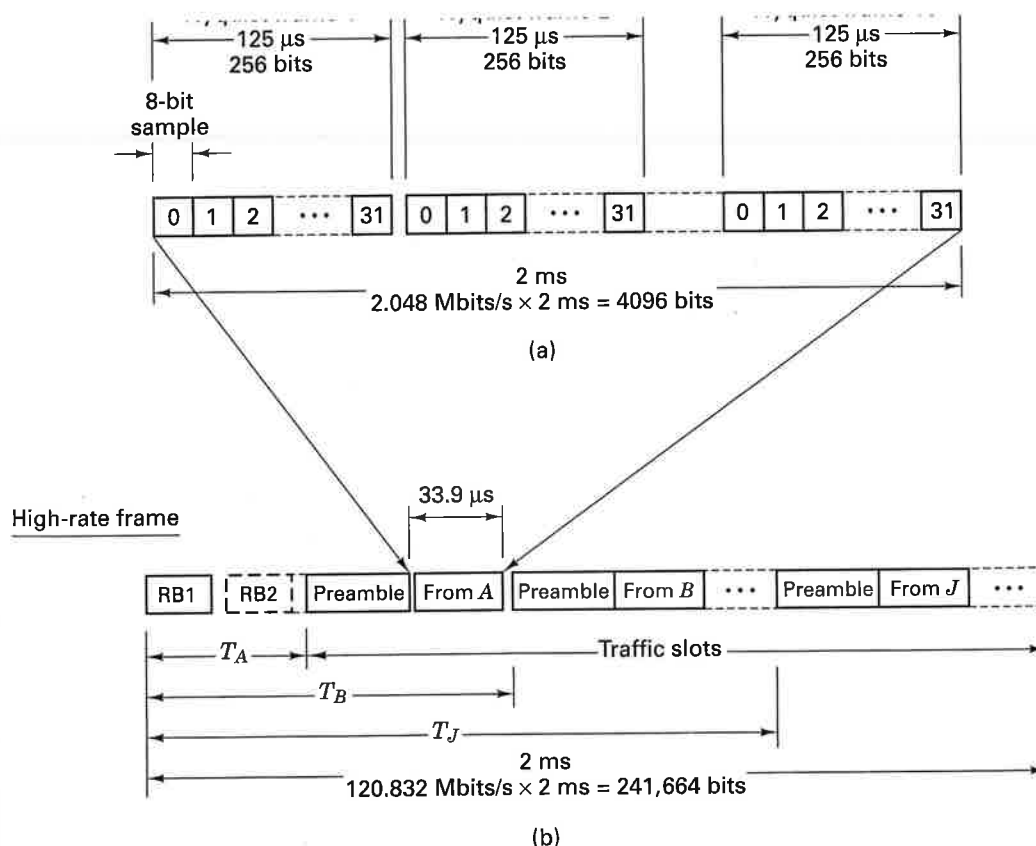


Figure 11.35 INTELSAT digital transmission standards for Europe. (a) Terrestrial PCM multiplex. (b) High-rate frame.

much faster rate than the rate at which it is generated. Figure 11.35b illustrates a 2-ms high-rate TDMA frame. The frame begins with a reference burst, RB1, emitted by a reference station. The burst contains information necessary to enable other stations to precisely position their message traffic bursts in the frame. There may be a second burst, RB2, for reliability, followed by a sequence of traffic slots. The traffic slots may be preassigned, or they may be assigned according to a DAMA protocol [20].

The PCM multiplex signal with a bit rate of $R_0 = 2.048$ Mbits/s and a frame duration of $T = 2$ ms is compressed (by a factor of 59) and transmitted using QPSK modulation at a burst rate of $R_T = 120.832$ Mbits/s (symbol rate of 60.416 megasymbols/s). The duration of the traffic data field T_{tr} in the high-rate TDMA frame is calculated as follows:

$$\begin{aligned} T_{tr} &= \frac{R_0 T}{R_T} \\ &= \frac{2.048 \times 10^6 \times 2 \times 10^{-3}}{120.832 \times 10^6} \\ &= 33.9 \mu\text{s} \end{aligned} \quad (11.31)$$

To obtain the total duration of a traffic burst, the time used for the preamble must be added. If the preamble contains S_P symbols, then, assuming QPSK modulation, the total length of the traffic burst measured in number of symbols is

$$S_T = \frac{R_0 T}{2} + S_P \quad (11.32)$$

and the burst-time duration is

$$T_T = \frac{2S_T}{R_T} \quad (11.33)$$

If the preamble contains 300 symbols, then

$$\begin{aligned} S_T &= \frac{2.048 \times 10^6 \times 2 \times 10^{-3}}{2} + 300 \\ &= 2348 \text{ symbols} \end{aligned}$$

Using this in Equation (11.33), we obtain

$$T_T = \frac{2 \times 2348}{120.832 \times 10^6} = 38.9 \mu\text{s}$$

11.4.4.3 The High-Rate TDMA Frame for North America

The INTELSAT TDMA burst (bit) rate of $R_T = 120.832$ Mbits/s was chosen to be compatible with both the European and North American standards. Figure 11.36 is similar to Figure 11.35 except that the PCM multiplex signal is the 24-channel T-Carrier instead of the 30-channel European standard. The essential

illustrates a burst, RBI, try to enable frame. There traffic slots. ording to a

and a frame using QPSK 16 megasym- MA frame is

(11.31)

reamble must modulation,

(11.32)

(11.33)

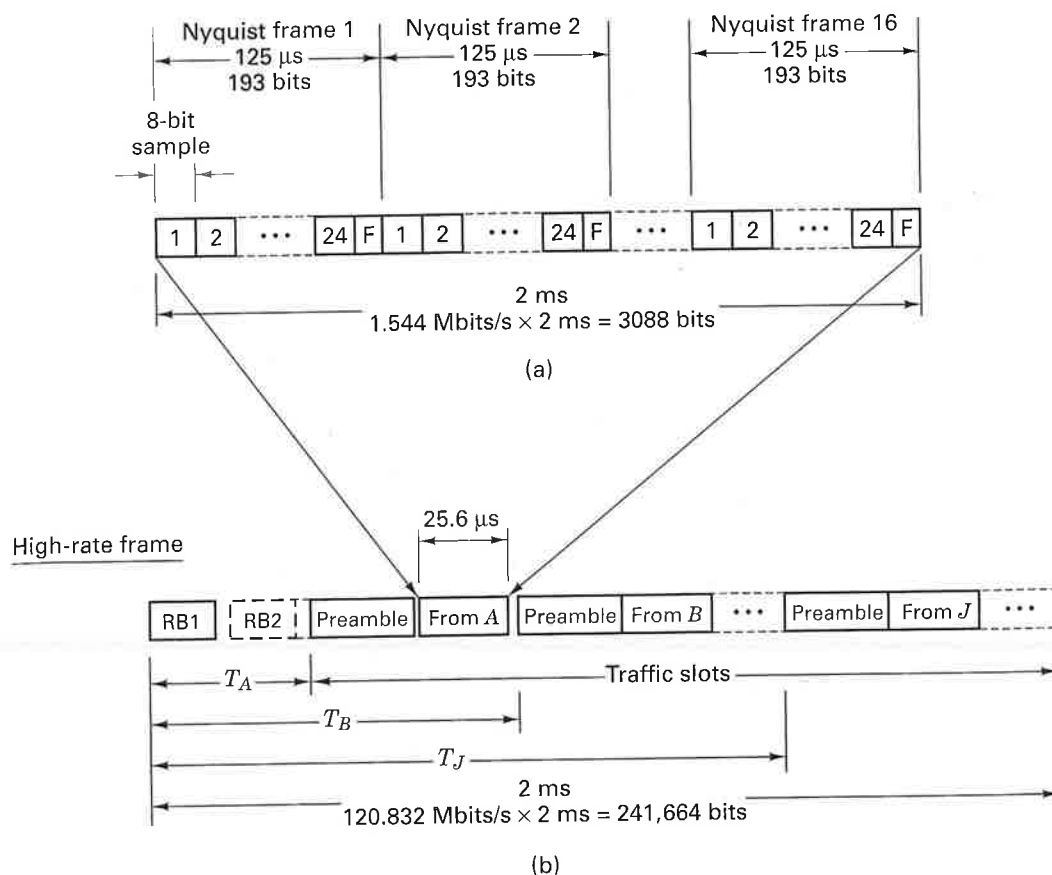


Figure 11.36 INTELSAT digital transmissions standards for T-Carrier. (a) Terrestrial PCM multiplex. (b) High-rate frame.

T-Carrier features that are different from the European standard are listed as follows:

1. Each Nyquist frame comprises 24 channels or samples × 8 bits + 1 frame alignment bit = 193 bits.
2. The 16 Nyquist frames contain 16 × 193 = 3088 bits.
3. The T-Carrier data rate is 1.544 Mbits/s.
4. The duration of the traffic data field in the high-rate TDMA frame is calculated from Equation (11.31).

$$T_{tr} = \frac{1.544 \times 10^6 \times 2 \times 10^{-3}}{120.832 \times 10^6} = 25.6 \mu s$$

/s was chosen dards. Figure signal is the The essential

11.4.4 INTELSAT TDMA Operation

At the transmitting earth station, the continuous low-rate data stream enters one of a pair of buffers illustrated in Figure 11.37a. When one buffer is filling at the low rate (1.544 Mbits/s or 2.048 Mbits/s), the other is emptying at the burst rate (120.832 Mbits/s). The buffers alternate functions at each TDMA frame. The time of application of the high-rate clock is controlled so that the traffic burst is transmitted in the proper interval to arrive at the satellite in its assigned position in the TDMA frame.

At the receiving station, the received traffic burst is routed to one of a pair of expansion buffers, shown in Figure 11.37b, that have the inverse function of the compression buffers in Figure 11.37a. When one buffer is filling at the high rate, the other is emptying at the desired output rate.

The most critical aspect of TDMA operation is the precise synchronization needed to assure orthogonality of the time slots [20]. Figure 11.38 illustrates the general idea behind most commercial satellite synchronization schemes. One station is designated as the master or control station. This station transmits periodic bursts of reference timing pulses. User stations also transmit their timing pulses, designated as slave pulses in Figure 11.38. On the downlink, the using station receives the master or reference pulses in addition to its own slave pulses. The time difference between the master and slave pulses corresponds to the timing error. The station adjusts its clock so as to reduce this timing error.

11.4.5 Satellite-Switched TDMA in INTELSAT

Modern communication satellites often employ several regional antenna beams. For a satellite based over the Atlantic Ocean, separate beams might be aimed at North America, Europe, South America, and Africa. Switches are used to allow

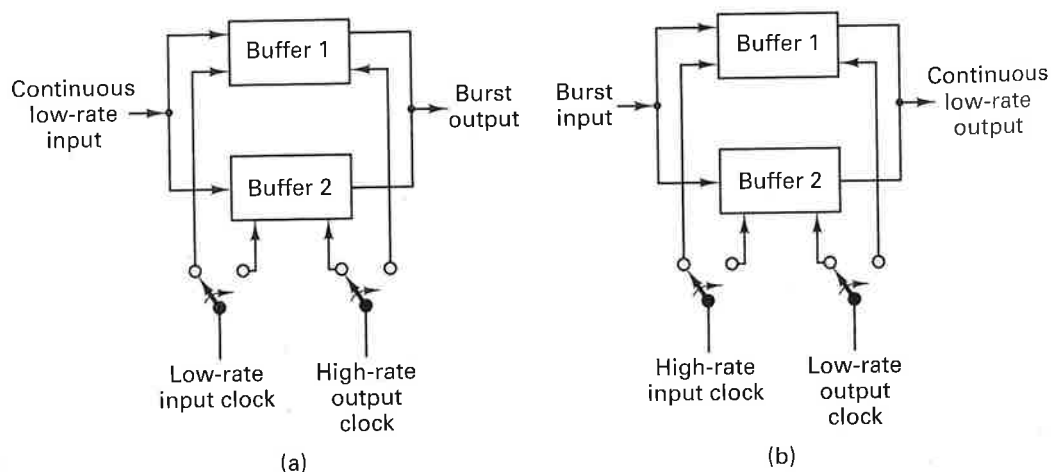


Figure 11.37 Burst compression and expansion buffers. (a) Compression buffers at transmitter. (b) Expansion buffers at receiver.

stream enters
filling at the
burst rate
ne. The time
burst is trans-
position in the

e of a pair of
action of the
he high rate,

ynchronization
illustrates the
nes. One sta-
mits periodic
iming pulses,
ng station re-
ses. The time
timing error.

tenna beams.
t be aimed at
used to allow

Continuous
low-rate
output

buffers

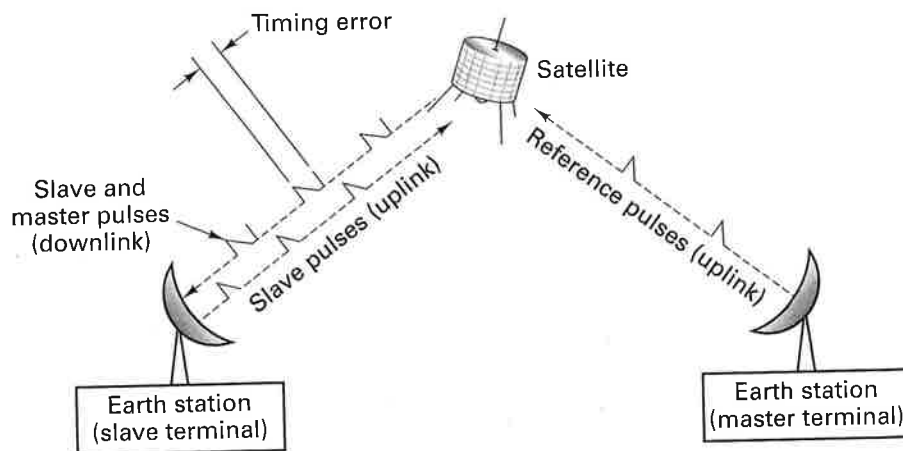


Figure 11.38 TDMA synchronization concept.

the interconnection of stations in one region to communicate with stations in another region. The basic goal of a satellite-switched TDMA (SS/TDMA) scheme is to provide an efficient way of cyclically providing interconnection of TDMA data among various coverage regions.

The heart of the system consists of a microwave switch matrix, located in the satellite, that is programmed via ground control to change states cyclically in rapid sequence, thus interconnecting distinct uplink beams to distinct downlink beams at each switching time. An earth station in the network communicates with those in other beams by transmitting TDMA bursts in the proper timing positions in the sequence. The pattern of switch states is selected so as to maximize the usable system capacity under the constraints of the traffic demands [21]. For complete interconnectivity between N beams, a total of $N!$ different satellite switch states or *modes* are required. Table 11.3 illustrates the six modes required for the full interconnectivity of a three-beam system.

In mode 1, the satellite receivers in beams A , B , and C are connected to the satellite transmitters for beams A , B , and C , respectively. An earth station in one of these beams can then communicate with other earth stations in the same beam. The beam is said to be *looped back* on itself.

TABLE 11.3 Three-Beam Satellite Switch Modes

Input	Output					
	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
A	A	A	B	B	C	C
B	B	C	A	C	A	B
C	C	B	C	A	B	A

Figure 11.39 illustrates a three-beam (beams A, B, and C) example of a SS/TDMA system. The satellite microwave switch matrix is configured in a *crossbar* design. This design can be thought of as being made up of row and column lines; when one row and one column are energized, contact is made at the intersection. A crossbar design only permits a single row to communicate with a single column at a time. If uplink A_U is connected to downlink B_D , neither A_U nor B_D can be simultaneously connected to any other beam.

In Figure 11.39, three different traffic patterns during time slot intervals T_1 , interval T_2 , and T_3 , with three different switch states S_1 , S_2 , and S_3 are shown. During interval T_1 , switch state S_1 interconnects the beams in a loop-back fashion which permits the uplink messages in slot T_1 to be delivered to their correct destinations. During time interval T_2 , switch state S_2 interconnects uplink beam A_U to downlink beam B_D , uplink beam B_U to downlink beam C_D , and uplink beam C_U to downlink beam A_D . This connection pattern assures that the uplink messages in slot T_2 are delivered to their correct destinations. During time interval T_3 , switch state S_3 similarly connects uplink transmissions to downlink beams to assure correct delivery of the data.

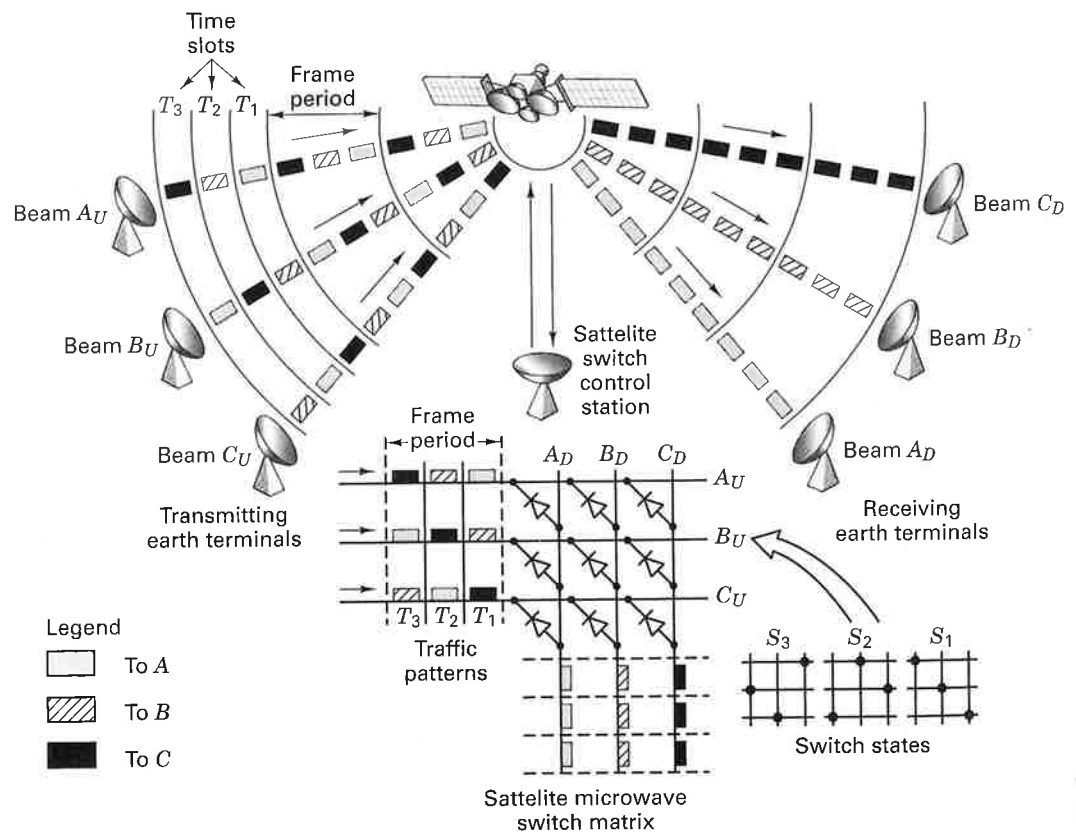


Figure 11.39 Satellite-switched TDMA (SS/TDMA).

The traffic patterns and their durations are programmed to optimize the resource capacity and to serve the users as efficiently as possible. The cyclic pattern can be reprogrammed by ground command to meet changing traffic requirements.

11.4.5.1 Traffic Matrix

Figure 11.40 is a matrix describing the communication traffic among N spot-beam coverages. In this figure, t_{ij} is the traffic volume from the i th beam to the j th beam. The subtotal

$$S_i = \sum_{j=1}^N t_{ij} \quad (11.34)$$

is the total traffic originating from the i th uplink beam, and

$$R_j = \sum_{i=1}^N t_{ij} \quad (11.35)$$

is the total traffic received in the j th downlink beam. When the traffic in a SS/TDMA system is controlled by a nonblocking switch (one that allows for the transmission of *all* messages, without any "busy" signals) a k -second time slot will be assigned to each channel in the TDMA frame. For efficient utilization of the CR, the total traffic in Figure 11.40 should be transmitted within a frame time T which should be made as short as possible. The minimum frame time for providing such nonblocking connectivity can be expressed [22] as

$$T_{\min} = k \max(\{S_i\}, \{R_j\}) \quad (11.36)$$

where $\max(\{S_i\}, \{R_j\})$ is the maximum value taken over the set of all $\{S_i\}$ and $\{R_j\}$. Equation 11.36 describes the minimum time to communicate *all* of the traffic in the traffic matrix, for equal bandwidth per channel.

	Destination						Originating traffic
	1	2	...	j	...	N	subtotals
1	t_{11}	t_{12}		t_{1j}		t_{1N}	S_1
2	t_{21}	t_{22}		t_{2j}		t_{2N}	S_2
...							
i	t_{i1}	t_{i2}		t_{ij}		t_{iN}	S_i
...							
N	t_{N1}	t_{N2}		t_{Nj}		t_{NN}	S_N
Received traffic subtotals	R_1	R_2		R_j		R_N	Total

Figure 11.40 Traffic matrix.

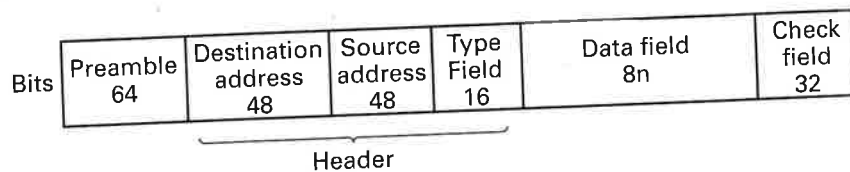
11.5 MULTIPLE ACCESS TECHNIQUES FOR LOCAL AREA NETWORKS

A local area network (LAN) can be used to interconnect computers, terminals, printers, and so on, located within a building or a small set of buildings. While long-haul networks use the public telephone network for economic reasons, LAN designers usually lay their own high-bandwidth cables. Bandwidth is not as scarce as it is in the long-haul cases. Not being forced to optimize bandwidth, a LAN can use simple access algorithms [6, 25–27].

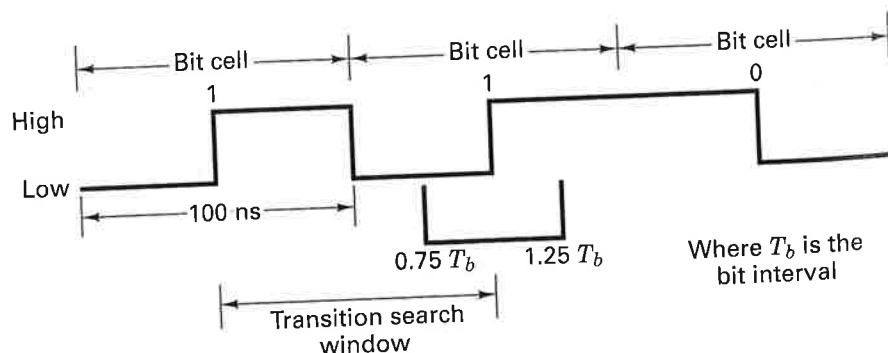
11.5.1 Carrier-Sense Multiple Access Networks

Ethernet is a LAN access scheme developed by the Xerox Corporation. The Ethernet scheme is based on the assumption that each local machine can sense the state of a common broadcast channel before attempting to use it. The technique is known as *carrier-sense multiple access with collision detection* (CSMA/CD). The word “carrier,” here, means any electrical activity on the cable. Figure 11.41a illustrates the bit field format for the Ethernet specification; the details are listed as follows:

1. The maximum packet size is 1526 bytes, where a byte is 8 bits. The packet breakdown is 8-byte preamble + 14-byte header + 1500-byte data + 4-byte parity.



(a)



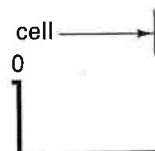
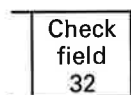
(b)

Figure 11.41 Ethernet bit field and PCM format. (a) Ethernet specification. (b) Manchester PCM format.

rs, terminals,
ldings. While
easons, LAN
not as scarce
width, a LAN

n. The Ether-
se the state of
que is known
)). The word
illustrates the
ollows:

s. The packet
data + 4-byte



T_b is the
interval

ecification.

s Chap. 11

2. The minimum packet size is 72 bytes, consisting of an 8-byte preamble + 14-byte header + 46-byte data + 4-byte parity.
3. The minimum spacing between packets is 9.6 μ s.
4. The preamble contains a 64-bit synchronization pattern of alternating ones and zeros, ending with two consecutive ones: (1 0 1 0 1 0 ... 1 0 1 0 1 1).
5. The receiving station examines a destination address field in the header to see if it should accept a particular packet. The first bit indicates the type of address (0 = single address, 1 = group address); an entire field of ones means an all-station broadcast.
6. The source address is the unique address of the transmitting machine.
7. The type field determines how the data field is to be interpreted. For example, bits in the type field can be used to describe such things as data encoding, encryption, message priority, and so on.
8. The data field is an integer number of bytes from a minimum of 46 to a maximum of 1500.
9. The parity check field houses the parity bits which are generated by the following generating polynomial (see Section 6.7):

$$X^{32} + X^{26} + X^{23} + X^{22} + X^{16} + X^{12} + X^{11} + X^{10} \\ + X^8 + X^7 + X^5 + X^4 + X^2 + X + 1$$

The Ethernet multiple access algorithm defines the following user action or response:

1. *Defer*. The user must not transmit when the carrier is present or within the minimum packet spacing time.
2. *Transmit*. The user may transmit if not deferring until the end of the packet or until a collision is detected.
3. *Abort*. If a collision is detected, the user must terminate packet transmission and transmit a short jamming signal to ensure that all collision participants are aware of the collision.
4. *Retransmit*. The user must wait a random delay time (similar to the ALOHA system) and then attempt retransmission.
5. *Backoff*. The delay before the n th attempt is a uniformly distributed random number from 0 to $2^n - 1$, for $(0 < n \leq 10)$. For $n > 10$, the interval remains 0 to 1023. The unit of time for the retransmission delay is 512 bits (51.2 μ s).

Figure 11.41b illustrates a 10-Mbits/s data stream with Manchester PCM formatting from the Ethernet specification. Notice that with such formatting, each bit cell or bit position contains a transition. A binary one is characterized by transitioning from a low level to a high level, while a binary zero has the opposite transition. Therefore, the presence of data transitions denotes to all "listeners" that the carrier is present. If a transition is not seen between 0.75 and 1.25 bit times since the last transition, the carrier has been lost, indicating the end of a packet.

11.5.2 Token-Ring Networks

A carrier-sense network consists of a cable onto which all stations are passively connected. A *ring network*, by comparison, consists of a series of point-to-point cables between consecutive stations. The interfaces between the ring and the stations are active rather than passive. Figure 11.42a illustrates a typical unidirectional ring with interface connections to several stations. Figure 11.42b illustrates the state of the interface for the listen mode and the transmit mode. In the *listen mode* the input bits are copied to the output with a delay of one bit time. In the *transmit mode*, the connection is broken so that the station can enter its own data onto the ring. The token is defined as a special bit pattern (e.g., 1 1 1 1 1 1 1) which circulates on the ring whenever all stations are idle. How does the system ensure that message data do not contain a tokenlike sequence? *Bit stuffing* is used to prevent this pattern from occurring in the data. For the 8-bit token example shown, a bit-stuffing algorithm would insert a zero into the data stream after each sequence of seven consecutive ones. The data receiver would use a similar algorithm to dispose of the inserted bit following any sequence of seven consecutive ones. The token-ring access scheme works as follows:

1. A station wanting to send a message monitors the token appearing at the interface. When the last bit of the token appears, the station inverts it (e.g., 1 1 1 1 1 1 1 0). The station then breaks the interface connection and enters its own data onto the ring.

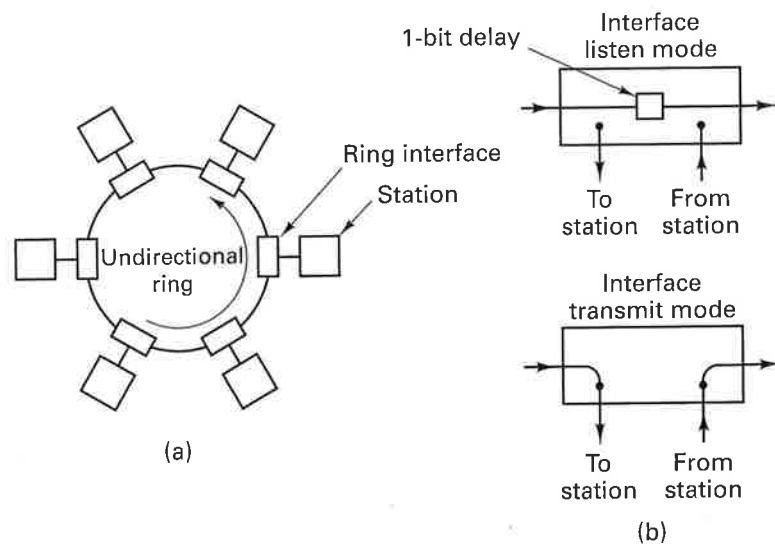


Figure 11.42 Token-ring network. (a) Network. (b) Listen and transmit modes.

2. As bits come back around the ring, they are removed by the sender. There is no limit on the size of the packets, because the entire packet never appears on the ring at one instant.
3. After transmitting the last bit of its message, the station must regenerate the token. After the last data bit has circled the ring and has been removed, the interface is switched back to the listen mode.
4. Contention is not possible with a token-ring system. During heavy traffic, as soon as a token is regenerated, the next downstream station requiring service will see and remove the token. Thereby, permission to transmit rotates smoothly around the ring. Since there is only one token, there is no contention.

The ring itself must have sufficient delay to enable a complete token to circulate when all stations are idle. A major issue in ring network design is the propagation distance or "length" of a bit. If the data rate is R Mbits/s, a bit is emitted every $(1/R)$ microseconds. Since the propagation rate along a typical coaxial cable is $200 \text{ m}/\mu\text{s}$, each bit occupies $200/R$ meters on the ring.

Example 11.4 Minimum Ring Size

If an 8-bit token is to be used on a 5-Mbits/s token-ring network, calculate the minimum *propagation distance*, d_p , needed for the ring circumference. Assume that the propagation velocity v_p is $200 \text{ m}/\mu\text{s}$.

Solution

$$R = 5 \text{ Mbits/s}$$

Time to emit one bit, t_b :

$$t_b = \frac{1}{5 \times 10^6} \text{ s}$$

Time to emit the 8-bit token, t_t :

$$t_t = \frac{8}{5 \times 10^6} \text{ s}$$

Propagation distance for the 8-bit token:

$$\begin{aligned} d_p &= t_t \times v_p \\ &= \frac{8}{5} \mu\text{s} \times 200 \text{ m}/\mu\text{s} \\ &= 320 \text{ m} \end{aligned}$$

11.5.3 Performance Comparison of CSMA/CD and Token-Ring Networks

Figure 11.43 compares the delay-throughput characteristics of a CSMA/CD network with a token-ring network. In each case, the cable length is 2 km, there are 50 stations on the network, the average packet length is 1000 bits, and the header size

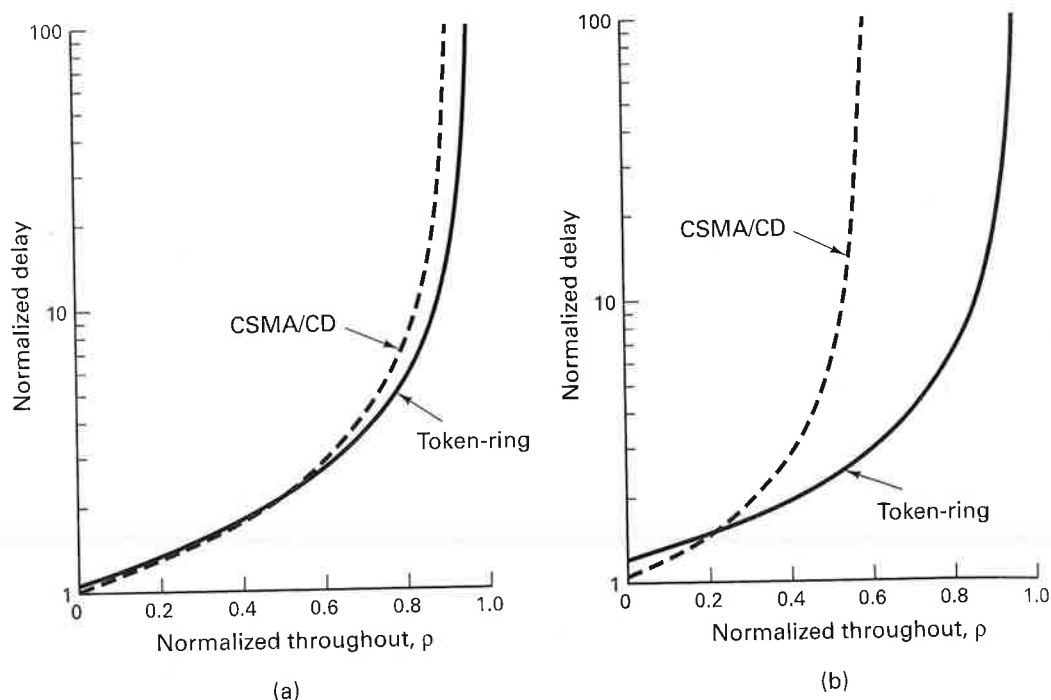


Figure 11.43 Delay versus throughput performance for CSMA/CD and token-ring networks. (a) Transmission rate = 1 Mbits/s. (b) Transmission rate = 10 Mbits/s. (Reprinted with permission from W. Bux, "Local-Area Subnetworks: A Performance Comparison," *IEEE Trans. Commun.*, vol. COM29, no. 10, Oct. 1981, pp. 1465–1473. © 1981 IEEE.)

is 24 bits. Figure 11.43a, the case where the transmission rate is 1 Mbits/s, illustrates that under these assumptions, CSMA/CD and token ring perform almost equally well. In Figure 11.43b, only one parameter has been changed as compared to Figure 11.43a; the transmission rate was increased to 10 Mbits/s. The difference for CSMA/CD is considerable; for normalized throughput, $\rho < 0.22$, CSMA/CD performs better than token ring. However, for $\rho > 0.22$, token ring clearly manifests better delay-throughput characteristics. To understand the reason for the poor CSMA/CD performance in Figure 11.43b, let us review the definition of ρ , described in Equations (11.17) and (11.19) and shown as

$$\rho = \frac{b\lambda}{R} = \frac{\rho'}{R}$$

where $\rho' = b\lambda$ is the channel throughput in bits per second and R is the channel capacity (maximum transmission bit rate). As R increases, channel throughput must increase accordingly for a given value of ρ . At higher channel throughput rates, a significant portion of the CSMA/CD transmission attempts ends in collision [26].

11.6 CONCLUSION

In this chapter we have outlined the concepts of resource sharing. The classical approaches of FDM/FDMA and TDM/TDMA were discussed in some detail. We also described a hybrid multiple access technique called CDMA, and introduced some of the satellite multiple access techniques that became popular in the 1970s and 1980s, known as multiple-beam frequency reuse and dual-polarization frequency reuse.

We described the demand-assignment (DAMA) techniques in the context of several versions of the ALOHA algorithm, and we considered several of the multiple-access techniques employed with INTELSAT, such as FDM/FM, SPADE, TDMA, and SS/TDMA. Finally, we examined two popular algorithms used for local area networks: carrier-sense multiple access with collision detection (CSMA/CD) and a token-ring network. The goals of the chapter were to introduce an assortment of multiple access techniques rather than attempting a rigorous treatment of any of them.

REFERENCES

1. Rubin, I., "Message Delays in FDMA and TDMA Communication Channels," *IEEE Trans. Commun.*, vol. COM27, no. 5, May 1979, pp. 769-777.
2. Nirenberg, L. M., and Rubin, I., "Multiple Access System Engineering—A Tutorial," *IEEE WESCON/78 Professional Program*, Modern Communication Techniques and Applications, session 21, Los Angeles, Sept. 13, 1978.
3. Abramson, N., "The ALOHA System—Another Alternative for Computer Communications," *Proc. Fall Joint Comput. Conf. AFIPS*, vol. 37, 1970, pp. 281-285.
4. Hayes, J. F., "Local Distribution in Computer Communications," *IEEE Commun. Mag.*, Mar. 1981, pp. 6-14.
5. Schwartz, M., *Computer-Communication Network Design and Analysis*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1977.
6. Tanenbaum, A. S., *Computer Networks*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981.
7. Abramson, N., "The ALOHA System," in N. Abramson and F. F. Kuo, eds., *Computer Communication Networks*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973.
8. Kleinrock, L., *Queueing Systems*, Vol. 1, *Theory*, John Wiley & Sons, Inc., New York, 1975.
9. Abramson, N., "Packet Switching with Satellites," *AFIPS Conf. Proc.*, vol. 42, June 1973, pp. 695-702.
10. Rosner, R. D., *Packet Switching*, Lifelong Learning Publications, Wadsworth Publishing Company, Inc., Belmont, Calif., 1982.
11. Crowther, W., Rettberg, R., Walden, D., Ornstein, S., and Heart, F., "A System for Broadcast Communication: Reservation ALOHA," *Proc. Sixth Hawaii Int. Conf. Syst. Sci.*, Jan. 1973, pp. 371-374.
12. Roberts, L., "Dynamic Allocation of Satellite Capacity through Packet Reservation," *AFIPS Conf. Proc.*, vol. 42, June 1973, p. 711.

13. Binder, R., "A Dynamic Packet-Switching System for Satellite Broadcast Channels," *Proc. Int. Conf. Commun.*, June 1975, pp. 41-1-41-5.
14. Capetanakis, J., "Tree Algorithms for Packet Broadcast Channels," *IEEE Trans. Inf. Theory*, vol. IT25, Sept. 1979, pp. 505-515.
15. Puente, J. G., and Werth, A. M., "Demand-Assigned Service for the INTELSAT Global Network," *IEEE Spectrum*, Jan. 1971, pp. 59-69.
16. Jones, J. J., "Hard Limiting of Two Signals in Random Noise," *IEEE Trans. Inf. Theory*, vol. IT9, Jan. 1963, pp. 34-42.
17. Bond, F. E., and Meyer, H. F., "Intermodulation Effects in Limiter Amplifier Repeaters," *IEEE Trans. Commun. Technol.*, vol. COM18, no. 2, Apr. 1970, pp. 127-135.
18. Shimbo, O., "Effects of Intermodulation, AM-PM Conversion, and Additive Noise in Multicarrier TWT Systems," *Proc. IEEE*, vol. 59, Feb. 1971, pp. 230-238.
19. Chakraborty, D., "INTELSAT IV Satellite System (Voice) Channel Capacity versus Earth-Station Performance," *IEEE Trans. Commun. Technol.*, vol. COM19, no. 3, June 1971, 355-362.
20. Campanella, S., and Schaefer, D., "Time Division Multiple Access Systems (TDMA)," in K. Feher, *Digital Communications, Satellite/Earth Station Engineering*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1983.
21. Scarcella, T., and Abbott, R. V., "Orbital Efficiency Through Satellite Digital Switching," *IEEE Commun. Mag.*, May 1983, pp. 38-46.
22. Muratani, T., "Satellite-Switched Time-Domain Multiple Access," *Proc. IEEE Electron. and Aerosp. Conf. (EASCON)*, 1974, pp. 189-196.
23. Dill, G. D., "TDMA, The State-of-the-Art," *Rec. IEEE Electron. Aerosp. Syst. Conv. (EASCON)*, Sept. 26-28, 1977, pp. 31-5A-31-5I.
24. Jarett, K., "Operational Aspects of Intelsat VI Satellite-Switched TDMA Communication System," *AIAA Tenth Commun. Satell. Syst. Conf.* Mar. 1984, pp. 107-111.
25. Stallings, W., "Local Network Performance," *IEEE Commun. Mag.*, vol. 22, No. 2, Feb. 1984, pp. 27-36.
26. Bux, W., "Local-Area Subnetworks: A Performance Comparison," *IEEE Trans. Commun.*, vol. COM29, no. 10, Oct. 1981, pp. 1465-1473.
27. Dixon, R. C., Strole, N. C., and Markov, J. D., "A Token-Ring Network for Local Data Communications," *IBM Syst. J.*, vol. 22, no. 1-2, 1983, pp. 47-62.

PROBLEMS

- 11.1. Design an FDM signal set consisting of five voice channels, each in the frequency range 300 to 3400 Hz. The multiplexed composite is to be made up of inverted sidebands and is to occupy the spectral region from 30 to 50 kHz.
 - (a) Draw the composite spectrum, indicating individual spectrum and guard band frequency locations.
 - (b) Draw a block diagram showing the heterodyning and filtering details and the required local oscillator values.
- 11.2. A receiver is tuned to receive the lower sideband (LSB) of a radio-frequency (RF) carrier wave with frequency, $f_c = 8$ MHz. The bandwidth of the LSB signal is 100 kHz. The receiver employs a local oscillator (LO) with frequency, f_{LO} , for hetero-

dyning the received signal down to a lower intermediate frequency (IF). Assume that $f_{LO} > f_c$, and that the IF amplifier is centered at 2 MHz. Draw a block diagram of the heterodyning conversion, including the RF filter, the LO, and the IF filter. Indicate the center frequency of each filter and typical spectra of the signals at various points in the diagram.

- 11.3.** Equations (11.13) to (11.15) demonstrate that the **average message delay time** for TDMA is less than that for FDMA. Discuss the **practical benefits** of such **reduced delay** in TDMA, as a function of frame time, for a satellite link with a one-way range of 36,000 km. For what values of frame time can there be a significant advantage of TDMA over FDMA?
- 11.4.** A group of stations share a 56-kbits/s pure ALOHA channel. Each station outputs a packet on the average of once every 10 s, even if the previous one has not yet been sent (i.e., the stations buffer the packets). Each packet comprises 3000 bits. What is the maximum number of stations that can share this channel, assuming that the arrival process is Poisson?
- 11.5.** A group of three stations share a 56-kbits/s pure ALOHA channel. The average bit rate transmitted from each of the three stations is $R_1 = 7.5$ kbits/s, $R_2 = 10$ kbits/s, and $R_3 = 20$ kbits/s. The size of each packet is 100 bits/packet. Find the normalized total traffic on the channel, the normalized throughput, the probability of successful transmission, and the arrival rate of successful packets. Assume that the arrival process is Poisson.
- 11.6.** Verify that for a pure ALOHA access scheme, the **normalized throughput** is bounded by $1/2e$ and that this maximum occurs when the **normalized total traffic** is equal to 0.5.
- 11.7.** (a) Verify that Equation (11.24) is a valid probability density function (pdf) for a discrete random variable.
 (b) Calculate the mean of a discrete random variable having a pdf like the one given in Equation (11.24).
 (c) Show that your result in part (b) is consistent with the claim that λ is the average packet arrival rate.
- 11.8.** Consider the pure ALOHA arrival scenario shown in Figure P11.1. The vertical arrows indicate packet arrival times. N_n is the number of arriving packets in the time interval $(T_{n-1}, T_n]$, where $(t_x, t_y]$ indicates the interval $t_x < t \leq t_y$. N_{n+1} is the number of arriving packets in $(T_n, T_{n+1}]$ and τ is the time duration per packet in seconds. The average arrival rate is λ . Assume the arrivals are independent of each other.
 (a) Write an expression for the joint pdf of N_n and N_{n+1} .

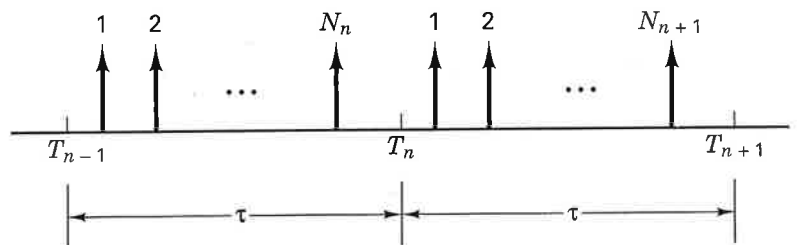


Figure P11.1

- (b) Let T_n define the time at which user A 's packet arrives. Express, in terms of the joint pdf of N_n and N_{n+1} , the probability that user A 's transmission will be successful.
- 11.9. Let $N = N_n + N_{n+1}$, where N_n and N_{n+1} are as defined in Problem 11.8. Write an expression for the pdf of N , and give an interpretation for N .
- 11.10. Six thousand stations are competing for the use of a single slotted ALOHA channel. The average station makes 30 requests per hour, where each request is for one slot of 500- μ s duration. Calculate the normalized total traffic on the channel.
- 11.11. Consider the arrival scenario of Figure P11.1; the location of the packet arrival times are permissible as shown under pure ALOHA, but not under slotted ALOHA, where arrivals are permitted only at the discrete times T_i , where $i = 0, 1, \dots$. Assume that the average arrival rate is λ_i .
- (a) How would Figure P11.1 need to be modified if slotted ALOHA is used? How would the pdfs of N_n and N_{n+1} change?
- (b) If user A 's packet arrives at time T_n , what is the probability of successful transmission?
- 11.12. A group of slotted-ALOHA stations generate a total of 120 requests per second, including both original and retransmissions. Each request is for a 12.5-ms duration slot.
- (a) What is the normalized total traffic on the channel?
- (b) What is the probability of a successful transmission on the first attempt?
- (c) What is the probability of exactly two collisions before a successful transmission?
- 11.13. Measurements of a slotted-ALOHA channel show that 20% of the slots are idle.
- (a) What is the normalized total traffic on the channel?
- (b) What is the normalized throughput?
- (c) Is the channel underloaded or overloaded?
- 11.14. Show that the sum of two Poisson processes, with rates λ_1 and λ_2 , is also a Poisson process, with rate $\lambda_i = \lambda_1 + \lambda_2$. Generalize your result for the sum of n Poisson processes.
- 11.15. A 10-MHz transponder is occupied by 200 identical carriers, half servicing stations with $G/T = 40$ dB/K, the other half servicing stations with $G/T = 37$ dB/K. All stations have a requirement to operate with a bit error probability of 10^{-5} . The transponder is power limited under this configuration.
- (a) What is the maximum possible bandwidth for each carrier?
- (b) Suppose that each carrier has a bandwidth of 40 kHz, and the transponder is required to service a group of larger ($G/T = 40$ dB/K) stations only. How many stations can the transponder handle? Will the transponder be power or bandwidth limited?
- (c) Repeat part (b) for the case where the transponder is to service a group of small ($G/T = 37$ dB/K) stations only.
- 11.16. A TDMA system operates at 100 Mb/s with a 2-ms frame time. Assume that all slots are of equal length and that a guard time of 1 μ s is required between slots.
- (a) Compute the efficiency of the communications resource (CR) for the case of 1, 2, 5, 10, 20, 50, and 100 slots per frame.
- (b) Repeat part (a) assuming that a 100-bit preamble is required at the start of each slot. Compute the efficiency of the CR in terms of the desired information transmission.
- (c) Graph the results of parts (a) and (b).

11.17. With reference to Equation 11.36:

- (a) Discuss the efficiency of the CR use if all S_i and R_j are equal.
- (b) Discuss the effect of a few S_i or R_j being much larger than the majority. How can the efficiency of the CR be improved?
- (c) When are the distributions of S_i and R_j likely to be similar? Dissimilar?

11.18. (a) Consider a token-ring network operating at a transmission rate of 10 Mbits/s over a cable having a propagation velocity of 200 m/ μ s. How many meters of cable is equal to a delay of 1 bit at each ring interface?

- (b) If the token is 10 bits long, and all but three stations are switched off during evening hours, what is the minimum cable length needed for the ring?

QUESTIONS

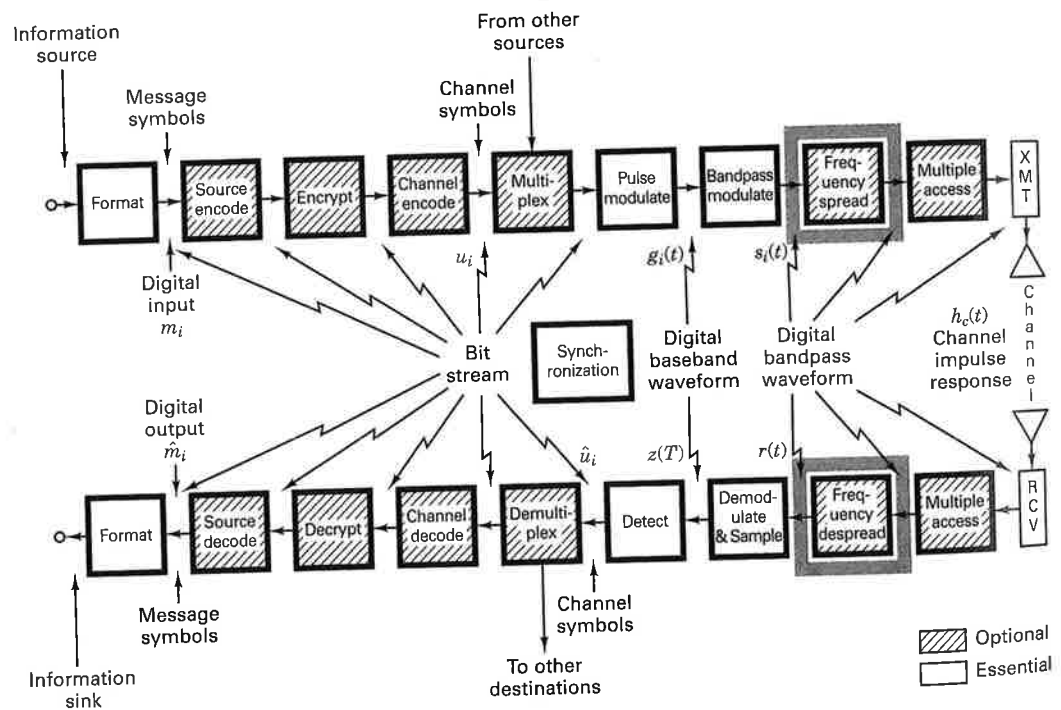
- 11.1. What is typically meant by the terms *communications resource*? (See Chapter 11, introduction.)
- 11.2. What are the similarities and differences between the terms *multiplexing* and *multiple access*? (See Chapter 11, introduction.)
- 11.3. Why is it not possible to use a *linear device* as a mixer (See Section 11.1.1.1 and Appendix A.)
- 11.4. Is there any theoretical capacity advantage in providing users with FDMA service versus TDMA service? (See Section 11.1.4.1.)
- 11.5. What benefits might there be in using CDMA versus either FDMA or TDMA access schemes? (See Section 11.1.5.)

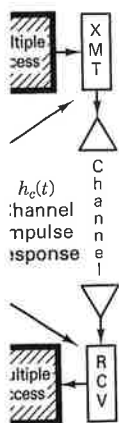
EXERCISES

Using the Companion CD, run the exercises associated with Chapter 11.

CHAPTER 12

Spread-Spectrum Techniques





Optional

Essential

12.1 SPREAD-SPECTRUM OVERVIEW

The initial application of spread-spectrum (SS) techniques was in the development of military guidance and communication systems. By the end of World War II, spectrum spreading for jamming resistance was already a familiar concept to radar engineers [1], and during subsequent years, SS investigation was motivated primarily by the desire to achieve highly jam-resistant communication systems. As a result of this research, there emerged an assortment of other applications in such areas as energy density reduction, high-resolution ranging, and multiple access, which will be discussed in later sections. The techniques considered in this chapter are called *spread spectrum* because the transmission bandwidth employed is much greater than the minimum bandwidth required to transmit the information. A system is defined to be a spread-spectrum system if it fulfills the following requirements:

1. The signal occupies a bandwidth much in excess of the minimum bandwidth necessary to send the information.
2. Spreading is accomplished by means of a *spreading signal*, often called a *code signal*, which is independent of the data. The details of some spreading signals are described in later sections.
3. At the receiver, despreading (recovering the original data) is accomplished by the correlation of the received spread signal with a *synchronized replica* of the spreading signal used to spread the information.

Standard modulation schemes such as frequency modulation and pulse code modulation also spread the spectrum of an information signal, but they do not qualify as spread-spectrum systems since they do not satisfy all the conditions outlined above.

12.1.1 The Beneficial Attributes of Spread-Spectrum Systems

12.1.1.1 Interference Suppression Benefits

White Gaussian noise is a mathematical model that, by definition, has infinite power spread uniformly over all frequencies. Effective communication is possible with this interfering noise of infinite power because only the finite-power noise components that are present within the signal space (in other words, share the *same coordinates* as the signal components) can interfere with the signal. The balance of the noise power may be thought of as noise that is effectively tuned out by the detector (see Section 3.1.3). For a typical narrowband signal, this means that only the noise in the signal bandwidth can degrade performance. Since spread-spectrum (SS) techniques were initially developed as a military application to permit reliable communications in the face of an enemy interferer (jammer), we begin by focusing on the anti-jam (AJ) capabilities of SS. (Commercial applications are treated in Sections 12.7 and 12.8.)

The idea behind a spread-spectrum AJ system is as follows. Consider that many orthogonal signal coordinates or dimensions are available to a communication link and that only a small subset of these signal coordinates are used at any time. We assume that the jammer cannot determine the signal subset that is currently in use. For signals of bandwidth W and duration T , the number of signaling dimensions can be shown [2] to be approximately $2WT$. Given a specific design, the error performance of such a system is only a function of E_b/N_0 . Against white Gaussian noise, with *infinite* power, the use of spreading (large $2WT$) offers no performance improvement. However, when the noise stems from a jammer with a *fixed finite* power and with uncertainty as to where in the signal space the signal coordinates are located, the jammer's choices are limited to the following:

1. Jam *all* the signal coordinates of the system, with an *equal* amount of power in each one, with the result that *little* power is available for each coordinate.
2. Jam a *few* signal coordinates with *increased* power in each of the jammed coordinates (or more generally, jam all the coordinates with various amounts of power in each).

Figure 12.1 compares the effect of spectrum spreading in the presence of white noise with spreading in the presence of an intentional jammer. The power spectral density of the signal is denoted $G(f)$ before spreading, and $G_{ss}(f)$ after spreading. For simplicity, the figure treats the frequency dimension only. In Figure 12.1a it can be seen that the single-sided power spectral density of white noise, N_0 , is unchanged as a result of expanding the signal bandwidth from W to W_{ss} . The av-

nd pulse code
t they do not
he conditions

n, has infinite
on is possible
-power noise
ds, share the
gnal. The bal-
tuned out by
is means that
Since spread-
-ization to per-
-er), we begin
lications are

Consider that
- communica-
- used at any
-t that is cur-
-r of signaling
ic design, the
-against white
-offers no per-
-mmer with a
-ce the signal
-ing:

unt of power
-coordinate.

the jammed
ious amounts

presence of
- The power
- $G_{ss}(f)$ after
-ly. In Figure
-ite noise, N_0 ,
- W_{ss} . The av-

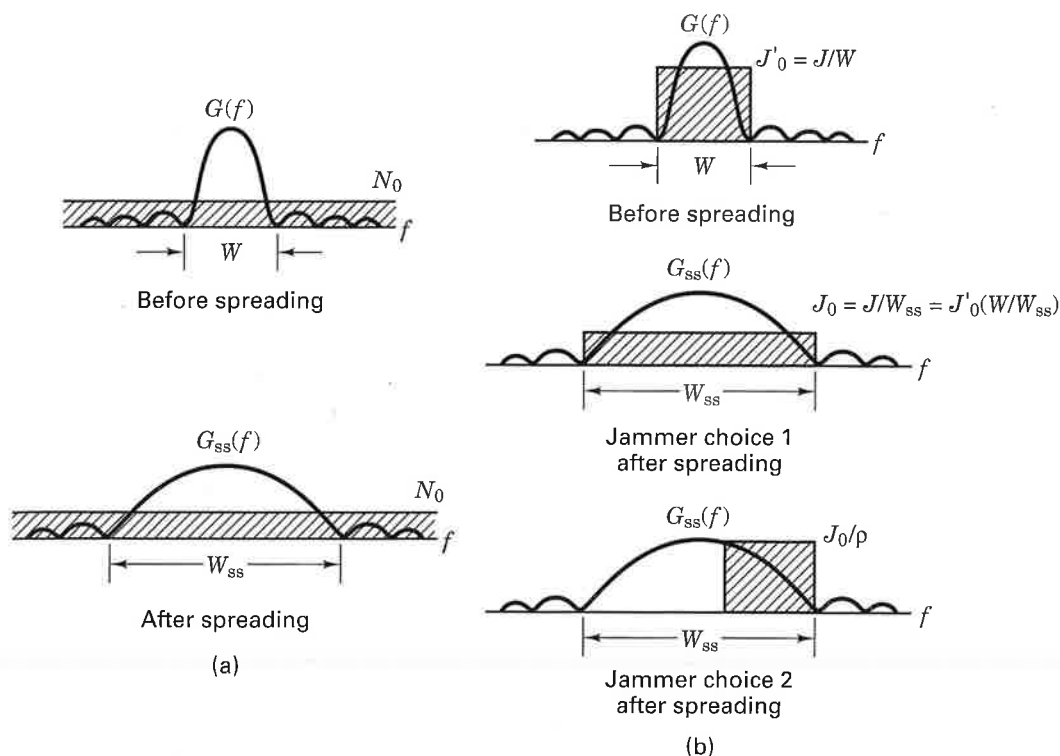


Figure 12.1 Effect of spectrum spreading. (a) Spectrum spreading in the presence of white noise. (b) Spectrum spreading in the presence of an intentional jammer.

erage power of white noise (area under the spectral density curve) is infinite. Hence, the use of spreading offers no performance improvement here. Figure 12.1b (upper diagram) illustrates the case of received (fixed finite) jammer power, J , and power spectral density, $J'_0 = J/W$, where W is the unsread bandwidth being jammed. Once the signal bandwidth is spread, the jammer can make one of the two choices listed earlier—choice 1 results in a reduction in jammer noise spectral density, J'_0 , by a factor (W/W_{ss}) across the spread spectrum. The resulting noise spectral density, $J_0 = J/W_{ss}$, is referred to as the *broadband jammer noise spectral density*. Choice 2 results in a reduction in the number of signal coordinates that the jammer occupies. However, with choice 2 the jammer can increase its noise spectral density from J_0 to J_0/ρ ($0 < \rho \leq 1$), where ρ is the portion of the spread-spectrum band the jammer elects to jam. If the jammer makes a poor choice in the coordinates to be jammed, the average effect of jamming will be less than if it makes a good choice. The larger the dimensionality of the signal set or the more signal coordinates the communicator can choose from, the greater is the jammer's uncertainty regarding the effectiveness of the jamming technique, and the better will be the protection against jamming. The comparison of unsread- versus spread-spectrum signaling

should be done under the assumption that there is the same amount of total average power in both cases. Since the area under the power spectral density (psd) curves represent total average power, there should be equal area under each of the psd curves for the unspread and the spread examples. Hence, it should be clear that in Figure 12.1, the $G_{ss}(f)$ plots are not to scale in both parts a and b.

Jamming is not always the result of an intentional act. Sometimes, the jamming signal is caused by natural phenomena, and sometimes it is the result of self-interference caused by *multipath*, in which delayed versions of the signal, arriving via alternative paths, interfere with the direct path transmission.

12.1.1.2 Energy Density Reduction

One can imagine situations where it is desired that a communications link be operated without being detected by anyone other than the intended receiver. Systems designed for this special task are known as *low probability of detection* (LPD) or *low probability of intercept* (LPI) communication systems. These systems are designed to make the detection of their signals as difficult as possible by anyone but the intended receiver. The goal of such a system is to use the minimum signal power and the optimum signaling scheme that results in the minimum probability of being detected. Since, in spread-spectrum systems, the signal is spread over many more signaling coordinates than in conventional modulation schemes, the resulting signal power is, on average, spread thinly and uniformly in the spread domain. Therefore, not only can the spread-spectrum signal be made difficult to jam, but additionally, the signal's very existence may be rendered difficult to perceive. To anyone who does not possess a synchronized replica of the spreading signal, the spread-spectrum signal will seem "buried in the noise."

A *radiometer* is a simple power measuring instrument that can be used by an adversary to detect the presence of spread-spectrum signals within some bandwidth W . The radiometer, illustrated in Figure 12.2, consists of a bandpass filter (BPF) with bandwidth W , a squaring circuit to ensure a positive output value (since a measure of *signal energy* is being detected), and an integrating circuit. At time $t = T$, the output of the integrator is compared to a preset threshold. If the output of the integrator is larger than the threshold, a signal is declared present; otherwise, the signal is declared absent. References [3, 4] provide details on the detectability of spread-spectrum signals, using radiometers and other more complicated instruments that make use of the features of the SS signal itself.

Spread-spectrum systems that are designed to exhibit LPI may also exhibit a *low probability of position fix* (LPPF), which means that even if the presence of the signal is perceived, the direction of the transmitter is difficult to pinpoint. Some spread-spectrum systems also exhibit a *low probability of signal exploitation* (LPSE), which means that the identification of the source is difficult to ascertain.

Another, unrelated application of spread-spectrum signaling deals with the fact that in some cases energy density reduction may be required to meet national allocation regulations. Downlink transmissions from satellites must meet international regulations on the spectral density that impinges on the earth. By spreading the downlink energy over a wider bandwidth, the total transmitted power can be

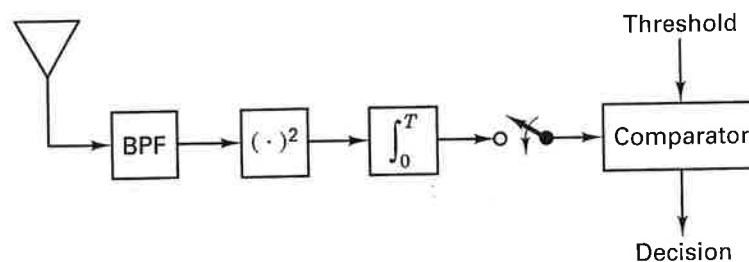


Figure 12.2 Radiometer.

increased and hence performance improved, while the energy density regulations are followed.

12.1.1.3 Fine Time Resolution

Spread-spectrum signals can be used for ranging or determination of position location. Distance can be determined by measuring the time delay of a pulse as it traverses the channel. Uncertainty in the delay measurement is inversely proportional to the bandwidth of the signal pulse. This can be seen by the illustration in Figure 12.3. The uncertainty of the measurement, Δt , is proportional to the rise time of the pulse, which is inversely proportional to the bandwidth of the pulse signal; that is,

$$\Delta t \approx \frac{1}{W} \quad (12.1)$$

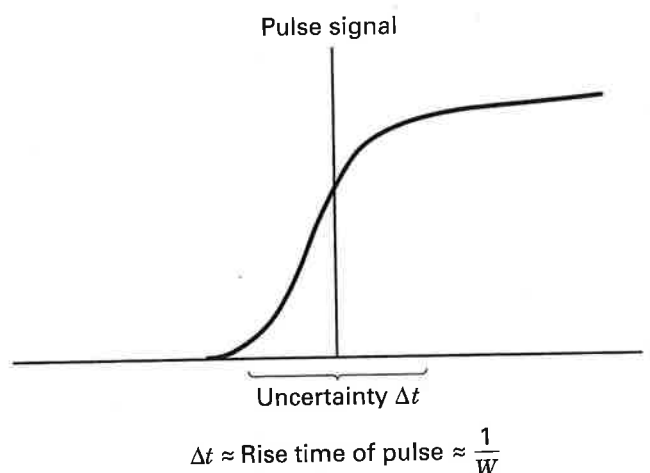


Figure 12.3 Time-delay measurement.

The larger the bandwidth, the more precisely one can measure range. Over a Gaussian channel, a one-shot measurement on a single pulse is not very reliable. The spread-spectrum technique, however, uses a code signal consisting of a long sequence of polarity changes (e.g., a binary PSK-modulated signal) in place of the single pulse. Upon reception, the received sequence is correlated against a local replica and the results of the correlation are used to perform an accurate time-delay or range measurement.

12.1.1.4 Multiple Access

Spread-spectrum methods can be used as a multiple access technique, in order to share a communications resource among numerous users in a coordinated manner. The technique, termed *code-division multiple access* (CDMA), since each simultaneous user employs a unique spread-spectrum signaling code, was discussed briefly in Chapter 11. One of the by-products of this type of multiple access is the ability to provide communication privacy between users with different spreading signals. An unauthorized user (a user not having access to a spreading signal) cannot easily monitor the communications of the authorized users. (A more detailed treatment is presented in a later section.)

12.1.2 A Catalog of Spreading Techniques

Figure 12.4 highlights the popular techniques for spreading the information signal over a large number of signal coordinates or dimensions. For signals of bandwidth W and duration T , the dimensionality of the signaling space is approximately $2WT$. To increase the dimensionality, we can either increase W by spectrum spreading, or increase T by time spreading or time hopping (TH). With spectrum spreading the signal is spread in the frequency domain. With time hopping, a message with data rate R is allocated a longer transmission-time duration than would be used with a conventional modulation scheme. During this longer time the data are sent in bursts according to the dictates of a code. We can say that with time hopping the signal is spread in the time domain. For both cases, frequency spreading and time spreading, a jammer will be uncertain regarding the signaling subset that is currently in use.

In Figure 12.4, the first two items listed under the category of spreading, *direct sequencing* (DS) and *frequency hopping* (FH), are the most commonly used techniques for spectrum spreading. As a jamming-rejection technique, *time hopping* (TH), the third item in the list, is similar to spread spectrum in that the location of the signal coordinates is hidden from potential adversaries. Also, there are hybrid combinations of the spreading techniques, for example, DS/FH, FH/TH, and DS/FH/TH; however, these techniques can be viewed as simple extensions of the material presented here and we will not elaborate on them. In this chapter, we focus only on the two major spread-spectrum techniques: direct sequencing and frequency hopping.

re range. Over a
ot very reliable.
sisting of a long
nal) in place of
elated against a
orm an accurate

ss technique, in
in a coordinated
MA), since each
le, was discussed
iple access is the
ferent spreading
ding signal) can-
A more detailed

information signal
als of bandwidth
roximately $2WT$.
um spreading, or
im spreading the
essage with data
d be used with a
data are sent in
time hopping the
reading and time
g subset that is

ory of spreading,
t commonly used
unique, *time hop-*
i in that the loca-
s. Also, there are
DS/FH, FH/TH,
ple extensions of
1 this chapter, we
t sequencing and

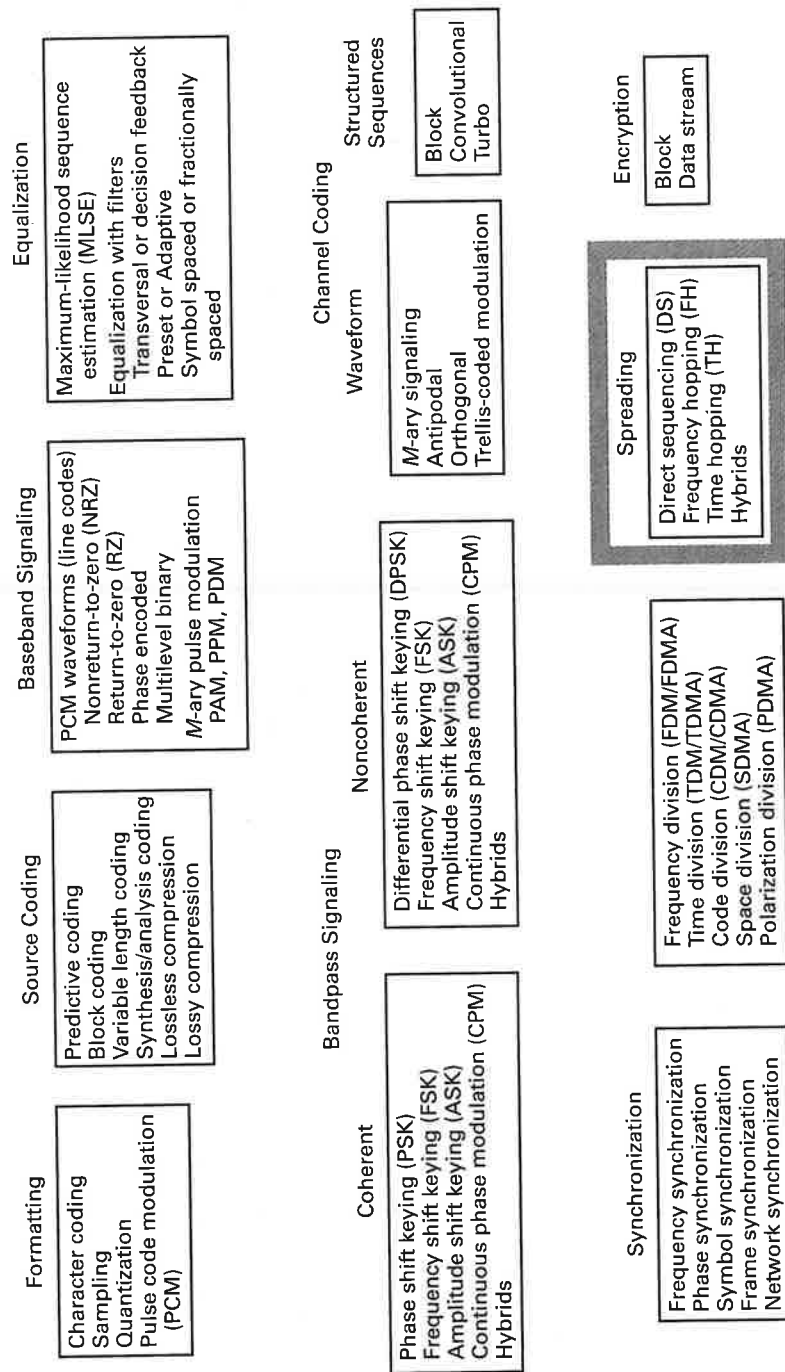


Figure 12.4 Basic digital communication transformations.

12.1.3 Model for Direct-Sequence Spread-Spectrum Interference Rejection

Figure 12.5 illustrates a model for direct-sequence spread-spectrum (DS/SS) interference rejection. At the modulator, the information signal $x(t)$, with a data rate of R bits/s, is multiplied by a spreading code signal $g(t)$, having a code symbol rate, usually called the *chip rate*, R_{ch} chips/second. Assume that the transmission bandwidths for $x(t)$ and $g(t)$ are R hertz and R_{ch} hertz, respectively. Multiplication in the time domain transforms to convolution in the frequency domain:

$$x(t)g(t) \leftrightarrow X(\omega) * G(\omega) \quad (12.2)$$

Therefore, if the data signal is narrowband compared to the spreading signal, the resulting product signal $x(t)g(t)$ will have approximately the bandwidth of the spreading signal. (See Section A.5.)

At the demodulator, the received signal is ideally multiplied by a synchronized replica of the spreading code signal, $g(t)$, which results in the despreading of the signal. A filter with bandwidth R is used to remove any spurious higher-frequency components. If there is any undesired signal at the receiver, the multiplication by $g(t)$ will spread this undesired signal, in the same way that the multiplication by $g(t)$ at the transmitter spread the desired signal originally. Consider the effect on a jammer that attempts to position a narrowband jamming signal within the information bandwidth. The first operation at the receiver input is multiplication by the spreading signal. Hence, the jamming tone is spread to the bandwidth of the spreading signal.

The essence behind the interference rejection capability of a spread-spectrum system can be summarized as follows:

1. Multiplication by the spreading signal *once* spreads the signal bandwidth.
2. Multiplication by the spreading signal *twice*, followed by filtering, recovers the original signal.

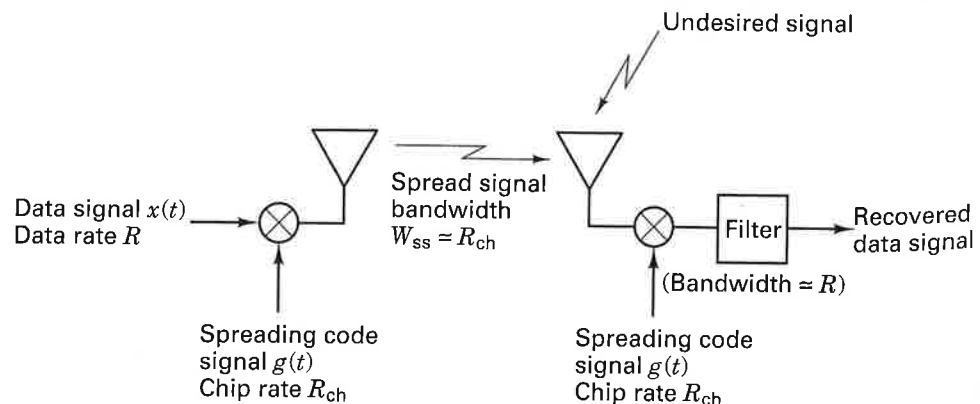


Figure 12.5 Basic spread-spectrum technique.

3. The desired signal gets multiplied *twice*, but the interference signal gets multiplied only *once*.

12.1.4 Historical Background

12.1.4.1 Transmitted Reference versus Stored Reference

During the early years of spread-spectrum investigation, one technique that was considered for operating a transmitter and receiver synchronously with a *truly random* spreading signal, such as wideband noise, was called a *transmitted reference* (TR) system. In a TR system, the transmitter would send two versions of an unpredictable wideband carrier—one modulated by data and the other unmodulated. These two signals were transmitted on separate channels. The receiver used the unmodulated carrier as the reference signal for despreading (correlating) the data-modulated carrier. The principal advantage of a TR system was that there were no significant synchronization problems at the receiver, since the data-modulated signal and the spreading signal used for despreading were transmitted simultaneously. The principal disadvantages of TR systems were that (1) the spreading code was sent in the clear and thus was available to any listener; (2) the system could be easily spoofed by a jammer sending a pair of waveforms acceptable to the receiver; (3) performance degraded at low signal levels since noise was present on both signals; and (4) twice the bandwidth and transmitted power were required because of the need to transmit the reference.

Modern spread-spectrum systems all use a technique called *stored reference* (SR), whereby the spreading code signal is independently generated at both the transmitter and the receiver. The main advantage of an SR system is that a well-designed code signal cannot be predicted by monitoring the transmission. Note that the noiselike code signal in an SR system cannot be truly random as it could in the case of a TR system. Since the same code must be generated independently at two or more sites, the code sequence must be deterministic, even though it should appear random to unauthorized listeners. Such random-appearing deterministic signals are called pseudonoise (PN), or pseudorandom signals; their generation is treated later in greater detail.

12.1.4.2 Noise Wheels

In the late 1940s and early 1950s, Mortimer Rogoff, working at ITT, demonstrated the fundamental operation of spectrum spreading systems with a novel experiment [5]. Using photographic techniques, Rogoff built a “noise wheel” for storing a noiselike signal. He randomly selected 1440 numbers not ending in 00 from the Manhattan telephone directory, and radially plotted the middle two of the last four digits so that the radius at every $\frac{1}{4}^\circ$ represented a new random number. The drawing was transferred to the wheel-shaped film shown in Figure 12.6. When the wheel was rotated past a slit of light, the resulting intensity-modulated light beam provided a stored noiselike spreading signal to be sensed by a photocell.

Rogoff mounted two such identical wheels on a single axis driven by a 900-rpm synchronous motor. One wheel’s noiselike spreading signal was modulated

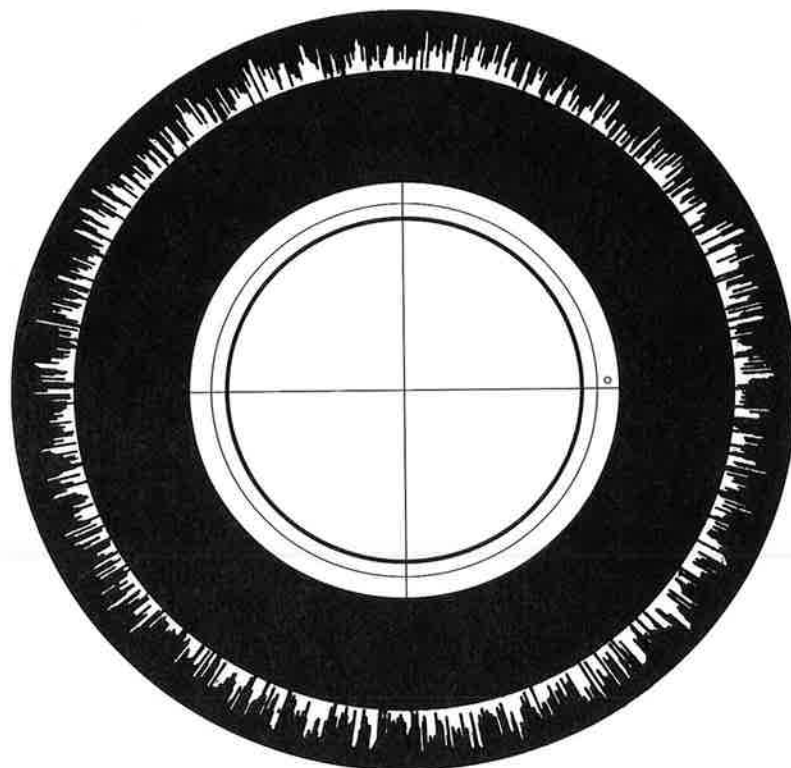


Figure 12.6 Rogoff's noise wheel. [Reprinted from Section I (Communications) of "Application of Statistical Methods to Secrecy Communication Systems," Proposal 946, Fed. Telecomm. Lab., August 28, 1950, Fig. 6, courtesy of ITT.]

with data (and interference) to provide one input to the receiving correlator, while the other wheel's unmodulated spreading signal provided the other input to the correlator. These baseband experiments, performed with the data rates of 1 bit/s, demonstrated the feasibility of conveying information hidden in noiselike signals [4].

12.2 PSEUDONOISE SEQUENCES

The spread-spectrum approach called *transmitted reference* (TR) can utilize a *truly* random code signal for spreading and despreading, since the code signal and the data-modulated code signal are simultaneously transmitted over different regions of the spectrum. The *stored reference* (SR) approach *cannot* use a truly random code signal since the code needs to be stored or generated at the receiver. For the SR system a *pseudonoise* or *pseudorandom* code signal must be used.

How does a pseudorandom signal differ from a random one? A random signal *cannot* be predicted; its future variations can only be described in a statistical sense. However, a pseudorandom signal is not random at all; it is a deterministic, periodic signal that is known to both the transmitter and receiver. Why the name “pseudonoise” or “pseudorandom”? Even though the signal is deterministic, it appears to have the statistical properties of sampled white noise. It appears, to an unauthorized listener, to be a truly random signal.

12.2.1 Randomness Properties

What are these randomness properties that make a pseudorandom signal appear truly random? There are three basic properties that can be applied to any periodic binary sequence as a test for the appearance of randomness. The properties, called *balance*, *run*, and *correlation*, are described for binary signals as follows:

1. *Balance property.* Good balance requires that in each period of the sequence, the number of binary ones differs from the number of binary zeros by at most one digit.
2. *Run property.* A *run* is defined as a sequence of a single type of binary digit(s). The appearance of the alternate digit in a sequence starts a new run. The length of the run is the number of digits in the run. Among the runs of ones and zeros in each period, it is desirable that about one-half the runs of each type are of length 1, about one-fourth are of length 2, one-eighth are of length 3, and so on.
3. *Correlation property.* If a period of the sequence is compared term by term with any cyclic shift of itself, it is best if the number of agreements differs from the number of disagreements by not more than one count.

In the next section, a PN sequence is generated to test these properties.

12.2.2 Shift Register Sequences

Consider the linear feedback shift register illustrated in Figure 12.7. It is made up of a four-stage register for storage and shifting, a modulo-2 adder, and a feedback path from the adder to the input of the register (modulo-2 addition has been defined in Section 2.9.3). The shift register operation is controlled by a sequence of clock pulses (not shown). At each clock pulse the contents of each stage in the register is shifted one stage to the right. Also, at each clock pulse the contents of stages X_3 and X_4 are modulo-2 added (a linear operation), and the result is fed back to stage X_1 . The shift register sequence is defined to be the output of the last stage—stage X_4 in this example.

Assume that stage X_1 is initially filled with a one and the remaining stages are filled with zeros, that is, the initial state of the register is 1 0 0 0. From Figure 12.7 we can see that the succession of register states will be as follows:

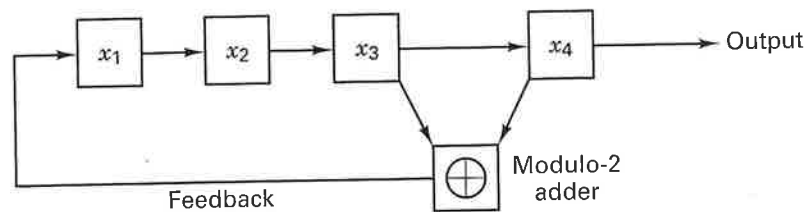


Figure 12.7 Linear feedback shift register example.

1 0 0 0 0 1 0 0 0 0 1 0 1 0 0 1 1 1 0 0 0 1 1 0 1 0 1 1 0 1 0 1
 1 0 1 0 1 1 0 1 1 1 1 0 0 1 1 1 0 0 0 1 1 0 0 0

Since the last state, 1 0 0 0, corresponds to the initial state, we see that the register repeats the foregoing sequence after 15 clock pulses. The output sequence is obtained by noting the contents of stage X_4 at each clock pulse. The output sequence is seen to be

0 0 0 1 0 0 1 1 0 1 0 1 1 1 1

where the leftmost bit is the earliest bit. Let us test the sequence above for the randomness properties outlined in the preceding section. First, the balance property; there are seven zeros and eight ones in the sequence—therefore, the sequence meets the balance condition. Next, the run property; consider the zero runs—there are four of them. One-half are of length 1, and one-fourth are of length 2. The same is true for the one runs. The sequence is too short to go further, but we can see that the run condition is met. The correlation property is treated in Section 12.2.3.

The shift register generator produces sequences that depend on the number of stages, the feedback tap connections, and initial conditions. The output sequences can be classified as either *maximal length* or *nonmaximal length*. Maximal length sequences have the property that for an n -stage linear feedback shift register the sequence repetition period in clock pulses p is

$$p = 2^n - 1 \quad (12.3)$$

Thus it can be seen that the sequence generated by the shift register generator of Figure 12.7 is an example of a maximal length sequence. If the sequence length is less than $(2^n - 1)$, the sequence is classified as a nonmaximal length sequence.

12.2.3 PN Autocorrelation Function

The autocorrelation function $R_x(\tau)$ of a periodic waveform $x(t)$, with period T_0 , was given in Equation (1.23) and is shown below in normalized form.

$$R_x(\tau) = \frac{1}{K} \left(\frac{1}{T_0} \right) \int_{-T_0/2}^{T_0/2} x(t)x(t + \tau)dt \quad \text{for } -\infty < \tau < \infty \quad (12.4)$$

where

$$K = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x^2(t) dt \quad (12.5)$$

When $x(t)$ is a periodic pulse waveform representing a PN code, we refer to each fundamental pulse as a *PN code symbol* or a *chip*. For such a PN waveform of unit chip duration and period p chips, the normalized autocorrelation function may be expressed as

$$R_x(\tau) = \frac{1}{p} \cdot \left(\begin{array}{l} \text{number of agreements less number of disagreements} \\ \text{in a comparison of one full period of the sequence} \\ \text{with a } \tau \text{ position cyclic shift of the sequence} \end{array} \right) \quad (12.6)$$

The normalized autocorrelation function for a maximal length sequence, $R_x(\tau)$, is shown plotted in Figure 12.8. It is clear that for $\tau = 0$, that is, when $x(t)$ and its replica are perfectly matched, $R(\tau) = 1$. However, for any cyclic shift between $x(t)$ and $x(t + \tau)$ with $(1 \leq \tau < p)$, the autocorrelation function is equal to $-1/p$ (for large p , the sequences are virtually decorrelated for a shift of a *single chip*).

It is now easy to test the output PN sequence of the shift register in Figure 12.7 for the third randomness property—correlation. The output sequence, as well as the same sequence with a single end-around shift, is as follows:

$$\begin{array}{cccccccccccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ \hline d & a & a & d & d & a & d & a & d & d & d & d & a & a & a \end{array}$$

The digits that agree are labeled *a* and those that disagree are labeled *d*. Following Equation (12.6), the value of the autocorrelation function for this single one-chip shift is seen to be

$$R(\tau = 1) = \frac{1}{15} (7 - 8) = -\frac{1}{15}$$

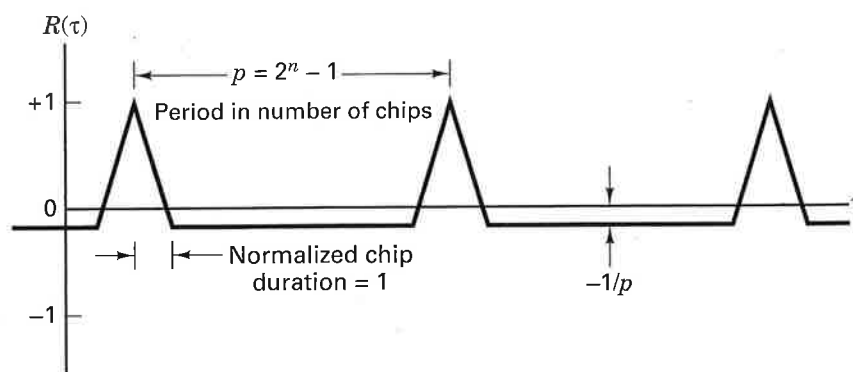


Figure 12.8 PN autocorrelation function.

Any cyclic shift yielding a mismatch from perfect synchronization results in the same autocorrelation value, $-1/p$. Hence the sequence meets the third randomness property.

12.3 DIRECT-SEQUENCE SPREAD-SPECTRUM SYSTEMS

The block diagram in Figure 12.9a depicts a *direct-sequence* (DS) modulator. "Direct sequence" is the name given to the spectrum spreading technique whereby a carrier wave is first modulated with a data signal $x(t)$, then the data-modulated signal is again modulated with a high-speed (wideband) spreading signal $g(t)$. Consider a constant-envelope data-modulated carrier having power P , radian frequency ω_0 , and data phase modulation $\theta_x(t)$, given by

$$s_x(t) = \sqrt{2P} \cos [\omega_0 t + \theta_x(t)] \quad (12.7)$$

Upon further constant-envelope modulation by the spreading signal, $g(t)$, the transmitted waveform can be expressed as

$$s(t) = \sqrt{2P} \cos [\omega_0(t) + \theta_x(t) + \theta_g(t)] \quad (12.8)$$

where the phase of the carrier is now seen to have two components: $\theta_x(t)$ due to the data and $\theta_g(t)$ due to the spreading sequence.

In Chapter 4, it was shown that ideal suppressed carrier binary phase shift keying (BPSK) modulation results in instantaneous changes of π radians to the phase of the carrier, according to the dictates of the data. We can equivalently express Equation (12.7) as the multiplication of the carrier wave by $x(t)$, an antipodal pulse stream with pulse values of $+1$ or -1 :

$$s_x(t) = \sqrt{2P} x(t) \cos \omega_0 t \quad (12.9)$$

If, like the data, the spreading sequence modulation is also BPSK, and $g(t)$ is an antipodal pulse stream with pulse values of $+1$ or -1 , Equation (12.8) can be written as

$$s(t) = \sqrt{2P} x(t) g(t) \cos \omega_0 t \quad (12.10)$$

A modulator based on Equation (12.10) is illustrated in Figure 12.9b. The data pulse stream and the spreading pulse stream are first multiplied, and then the composite $x(t)$ modulates the carrier. If the assignment of pulse value to binary value is

Pulse value	Binary value
1	0
-1	1

then the initial step in the DS/BPSK modulation can be accomplished by the modulo-2 addition of the binary data sequence with the binary spreading sequence.

Demodulation of the DS/BPSK signal is accomplished by correlating or re-modulating the received signal with a synchronized replica of the spreading signal

results in the
rd randomness

odulator. "Di-
que whereby a
ata-modulated
gnal $g(t)$. Con-
 P , radian fre-

(12.7)

$g(t)$, the trans-

(12.8)

$x(t)$ due to the

ry phase shift
adians to the
quivalently ex-
, an antipodal

(12.9)

and $g(t)$ is an
(12.8) can be

(12.10)

.9b. The data
then the com-
inary value is

lished by the
ing sequence.
elating or re-
eading signal

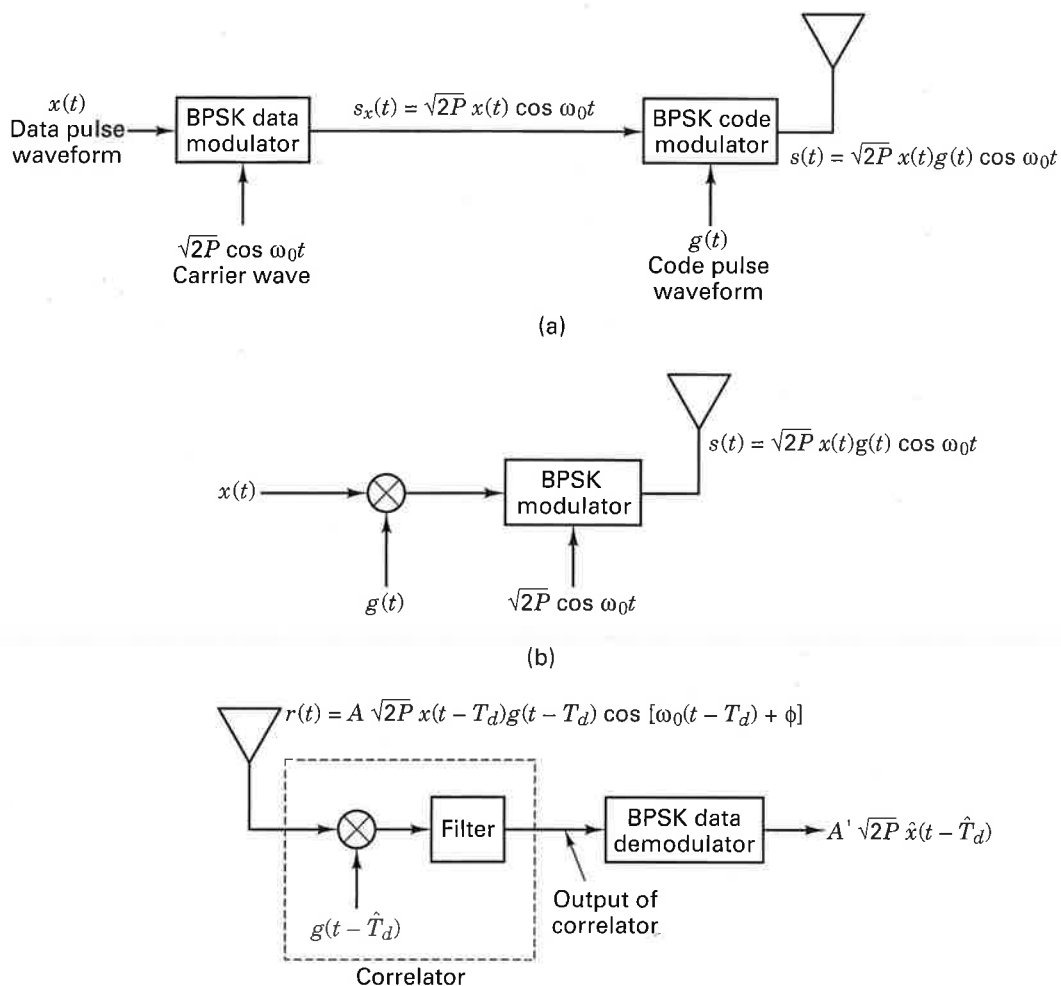


Figure 12.9 Direct-sequence spread-spectrum system. (a) BPSK direct-sequence transmitter. (b) Simplified BPSK direct-sequence transmitter. (c) BPSK direct-sequence receiver.

$g(t - \hat{T}_d)$ as seen in Figure 12.9c, where $\hat{T}_d$ is the receiver's estimate of the propagation delay T_d from the transmitter to the receiver. In the absence of noise and interference, the output signal from the correlator can be written as

$$A\sqrt{2P}x(t - T_d)g(t - T_d)g(t - \hat{T}_d) \cos [\omega_0(t - T_d) + \phi] \quad (12.11)$$

where the constant A is a system gain parameter and ϕ is a random phase angle in the range $(0, 2\pi)$. Since $g(t) = \pm 1$, the product $g(t - T_d)g(t - \hat{T}_d)$ will be unity if $\hat{T}_d = T_d$, that is, if the code signal at the receiver is exactly synchronized with

the code signal at the transmitter. When it is synchronized, the output of the receiver correlator is the despread data-modulated signal (except for a random phase ϕ and delay T_d). The despreading correlator is then followed by a conventional demodulator for recovering the data.

12.3.1 Example of Direct Sequencing

Figure 12.10 is an example of DS/BPSK modulation and demodulation following the block diagrams of Figure 12.9b and c. In Figure 12.10a are shown the binary data sequence (1, 0) and its bipolar pulse waveform equivalent $x(t)$, where the binary to pulse value assignments are the same as those described in the preceding section. Examples of a binary spreading sequence and its bipolar pulse waveform equivalent $g(t)$ are shown in Figure 12.10b. The modulo-2 addition of the data sequence and the code sequence, and the equivalent waveform of the product $x(t)g(t)$, is shown in Figure 12.10c.

For the BPSK modulation described by Equations (12.8) and (12.10), it is shown in Figure 12.10d that the phase of the carrier, $\theta_x(t) + \theta_g(t)$, equals π when the

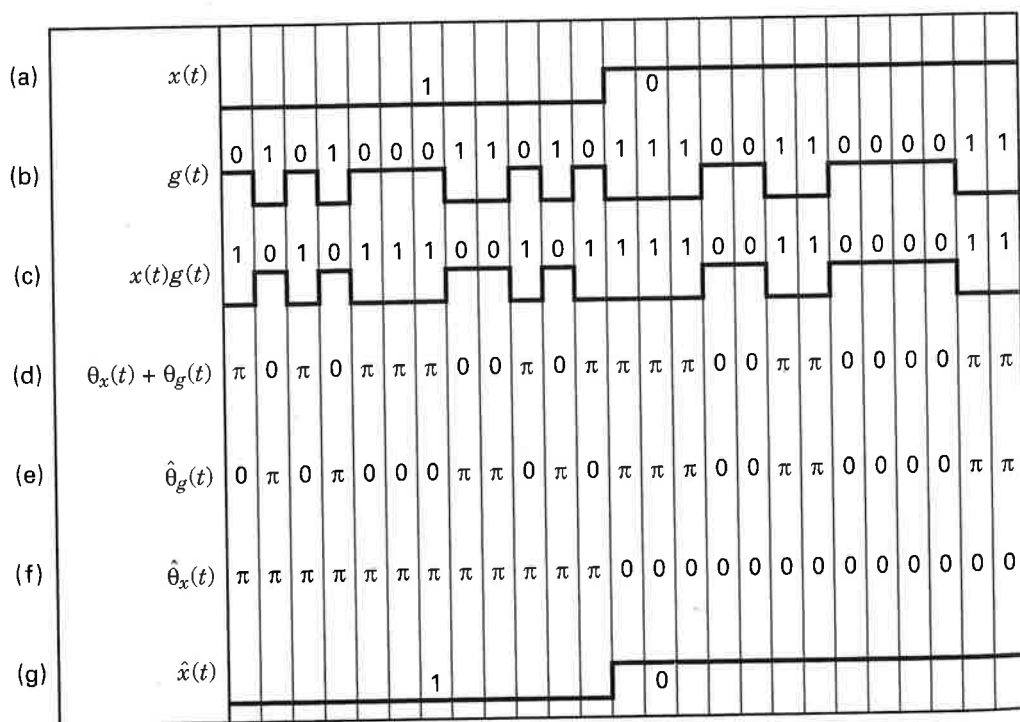


Figure 12.10 Spread-spectrum example using direct sequencing. (a) Binary data waveform to be transmitted. (b) Code sequence. (c) Transmitted sequence. (d) Phase of transmitted carrier. (e) Phase shift produced by receiver code. (f) Phase of received carrier after phase shifted (despread) by receiver code. (g) Demodulated data waveform.

output of the
or a random
followed by a

ion following
n the binary
where the bi-
he preceding
lse waveform
f the data se-
the product

(12.10), it is
ls π when the

0	0	1	1	
0	0	1	1	
0	0	π	π	
0	0	π	π	
0	0	0	0	

Binary
quence.
r code.
r code.

value of the product waveform $x(t)g(t)$ equals -1 (or the modulo-2 sum of data and code is binary 1). Similarly, the phase of the carrier is zero when the value of $x(t)g(t)$ equals $+1$ (or the modulo-2 sum of data and code is binary 0). One can appreciate the *signal hiding* property of spread-spectrum signals by comparing the code waveform in Figure 12.10b with the composite waveform in Figure 12.10c. The latter has the signal $x(t)$ "hidden" within it. Just as your eyes have difficulty finding the slowly moving data signal in the rapidly moving code signal, it is similarly difficult for a receiver to recover a slowly moving signal from a rapidly moving code without having an exact replica of the code.

As shown in Figure 12.9c, DS/BPSK demodulation is a two-step process. The first step, despreading, is accomplished by correlating the received signal with a synchronized replica of the code. The second step, data demodulation, is accomplished with a conventional demodulator. In the example of Figure 12.10 we see the code replica $\hat{\theta}_g(t)$, in Figure 12.10e, as the phase shift (either 0 or π) that is produced at the receiver by the despreading code. Figure 12.10f illustrates the resulting estimate of the carrier phase $\hat{\theta}_x(t)$, after despreading or after $\hat{\theta}_g(t)$ has been added to $\theta_x(t) + \theta_g(t)$. At this point one can recognize the original data pattern in the phase terms of the carrier wave. The final step, shown in Figure 12.10g, is to recover an estimate of the data waveform, $\hat{x}(t)$, by the use of a BPSK demodulator.

12.3.2 Processing Gain and Performance

A fundamental issue in spread-spectrum systems is *how much* protection spreading can provide against interfering signals with finite power. Spread-spectrum techniques distribute a relatively low-dimensional signal in a large-dimensional signal space. The signal is "hidden" within the signal space, since we assume that a jammer does not know which signal coordinates are being transmitted at any time. The only recourse for the jammer, intent upon communication disruption, is to jam the entire space with its fixed total power, thus inducing a limited amount of interference in each signal coordinate, or to jam a portion of the signal space with its total power, thus leaving the remainder of the signal space free of interference.

Consider a set of D orthogonal signals, $s_i(t)$, $1 \leq i \leq D$, in an N -dimensional space, where in general, $D \ll N$. Following the development in Section 3.1.3, we can write.

$$s_i(t) = \sum_{j=1}^N a_{ij} \psi_j(t) \quad \begin{matrix} i = 1, 2, \dots, D; \\ D \ll N \end{matrix} \quad 0 \leq t \leq T \quad (12.12)$$

where

$$a_{ij} = \int_0^T s_i(t) \psi_j(t) dt \quad (12.13)$$

and

$$\int_0^T \psi_j(t) \psi_k(t) dt = \begin{cases} 1 & \text{for } j = k \\ 0 & \text{otherwise} \end{cases} \quad (12.14)$$

The $\{\psi_j(t)\}$ are linearly independent functions that *span* or characterize the N -dimensional orthonormal space and are called *basis* functions of the space. For every information symbol that is transmitted, a set of coefficients $\{a_{ij}\}$ is chosen independently, using a pseudorandom spreading code, in order to hide the D -dimensional signal set in the larger N -dimensional space. The set of random variables $\{a_{ij}\}$ assume the values $\pm a$, each with a probability of $\frac{1}{2}$. The receiver, of course, has access to each set of coefficients chosen in order to perform the necessary correlation despreading. Even if the same i th symbol is sent repeatedly, the set $\{a_{ij}\}$ used to transmit it is newly selected from symbol to symbol. The energy in each signal waveform of the D signal set will be assumed equal, so that we can write the average energy for each signal as

$$E_s = \int_0^T \overline{s_i^2(t)} dt = \sum_{j=1}^N \overline{a_{ij}^2} \quad i = 1, 2, \dots, D \quad (12.15)$$

where the overbar means the expected value over the ensemble of many symbol transmissions. The independent coefficients have zero mean and correlation:

$$\overline{a_{ij}a_{ik}} = \begin{cases} \frac{E_s}{N} & \text{for } j = k \\ 0 & \text{otherwise} \end{cases} \quad (12.16)$$

The standard assumption is that the jammer has no a priori knowledge regarding the selection of the signaling coefficients $\{a_{ij}\}$. As far as the jammer is concerned, the coefficients are uniformly distributed over the N basis coordinates. If the jammer chooses to distribute its power uniformly over the total signal space, the jammer waveform $w(t)$ can be written

$$w(t) = \sum_{j=1}^N b_j \psi_j(t) \quad (12.17)$$

with total energy

$$E_w = \int_0^T w^2(t) dt = \sum_{j=1}^N b_j^2 \quad (12.18)$$

A reasonable goal for a jammer would be to devise a strategy for selecting the portions b_j^2 , of its fixed total energy E_w so as to minimize the desired signal-to-noise ratio (SNR) at the receiver after demodulation.

At the receiver, the detector output (ignoring receiver noise),

$$r(t) = s_i(t) + w(t) \quad (12.19)$$

is correlated with the set of possible transmitted signals, so that the output of the i th correlator is

$$z_i = \int_0^T r(t) s_i(t) dt = \sum_{j=1}^N (a_{ij}^2 + b_j a_{ij}) \quad (12.20)$$

The second term on the right side of Equation (12.20) averages to zero over the ensemble of all possible pseudorandom code sequences, since the set of random variables $\{a_{ij}\}$ assume the values $\pm a$, each with probability $\frac{1}{2}$. Therefore, given that $s_m(t)$ was transmitted, the expected value of the output of the i th correlator can be written, following the development in References [6, 7],

$$\mathbf{E}(z_i|s_m) = \sum_{j=1}^N \overline{a_{ij}} = \begin{cases} E_s & \text{for } i = m \\ 0 & \text{otherwise} \end{cases} \quad (12.21)$$

In Equation (12.21), the term $\mathbf{E}(z_i|s_m)$ for $i = m$ is to be interpreted as follows. Given that $s_i(t)$ is to be transmitted, N coefficients a_{ij} ($1 \leq j \leq N$) are chosen pseudorandomly (the receiver is assumed to have access to each choice of the a_{ij} for correlation despreading). Hence, in computing $\mathbf{E}(z_i|s_i)$, even though the i th information symbol is specified at the transmitter, the pattern of coefficients used to send it appears random (to the unauthorized receiver) for each transmission. Equation (12.21) presumes that the jammer had not been successful in its attempt to employ some clever tactics (described in Section 12.6).

Let us assume that all D signals are equally likely. Then the expected value at the output of any of the D correlators is

$$\mathbf{E}(z_i) = \frac{E_s}{D} \quad (12.22)$$

Similarly, using Equations (12.15) to (12.21), we compute $\text{var}(z_i|s_i)$, the variance at the output of the i th correlator, given that the i th signal was transmitted:

$$\text{var}(z_i|s_i) = \sum_{j,k} b_j b_k \overline{a_{ij} a_{ik}} \quad (12.23)$$

$$\begin{aligned} &= \sum_{j=1}^N b_j^2 \overline{a_{ij}^2} \\ &= \sum_{j=1}^N b_j^2 \frac{E_s}{N} \\ &= \frac{E_w E_s}{N} \end{aligned} \quad (12.24)$$

For completeness, the variance at the output of the i th correlator, given that the m th signal was transmitted, where $i \neq m$, can similarly be computed to be

$$\text{var}(z_i|s_m) = \frac{E_w E_s}{N} + \frac{E_s^2}{N} \quad (12.25)$$

The signal-to-jammer ratio at the output of the i th correlator can be defined as

$$\text{SJR} = \sum_{m=1}^D \frac{\mathbf{E}^2(z_i|s_m)}{\text{var}(z_i|s_m)} P(s_m) = \frac{E_s^2/D}{E_w E_s/N} = \frac{E_s N}{E_w D} \quad (12.26)$$

where the probability of the m th signal $P(s_m) = 1/D$, since the signals are assumed to occur with equal probability, and where the signal energy and the jammer energy in the i th correlator are denoted by $E^2(z_i)$ and $\text{var}(z_i)$, respectively. Because of Equation (12.21), the only terms in the summation of Equation (12.26) not equal to zero are those for which $i = m$. The result is independent of the way in which the jammer chooses to distribute its energy. Therefore, regardless of how b_j is chosen, subject to $\sum_j b_j^2 = E_w$, the SJR in Equation (12.26) indicates that spreading gives the signal an advantage of a factor of N/D over the jammer. The ratio N/D is known as the *processing gain* G_p .

Since the approximate dimensionality of a signal with bandwidth W and duration T is $2WT$, we can express the processing gain as

$$G_p = \frac{N}{D} \approx \frac{2W_{ss}T}{2W_{\min}T} = \frac{W_{ss}}{R} \quad (12.27)$$

Where W_{ss} is the spread-spectrum bandwidth (the total bandwidth used by the spreading technique) and $W_{\min}$ is the minimum bandwidth of the data (taken to be the data rate, R). For direct sequence systems, W_{ss} is approximately the code chip rate R_{ch} , and $W_{\min}$ is similarly the data rate R , giving

$$G_p = \frac{R_{ch}}{R} \quad (12.28)$$

where the *chip* is defined as the shortest uninterrupted waveform in the system. For direct-sequence spread-spectrum systems, the chip constitutes a PN-code pulse or signaling element.

Whatever the spread-spectrum application (e.g., interference rejection, fine-time resolution), processing gain is the parameter that expresses the performance advantage of the spread-spectrum system over a narrowband system. The modulation of choice for direct-sequence spread-spectrum systems is generally BPSK or QPSK. Thus, consider a single binary symbol comprising 1000 BPSK code chips. From Equation (12.28) it should be clear that the processing gain for such a signaling scheme is 1000. To appreciate that the processing gain of such a spread-spectrum system makes for more robust signaling than a similar narrowband scheme, consider the following. Imagine that the detection process makes a decision on each of the 1000 chips in our example. Of course it doesn't do that; the 1000 chips are accumulated and correlated to a code, yielding a single decision on the meaning of the bit. But, if you think of the process as 1000 individual decisions, you might say that 499 of those decisions can be incorrect, and you will still detect the bit correctly.

12.4 FREQUENCY HOPPING SYSTEMS

We now consider a spread-spectrum technique called frequency hopping (FH). The modulation most commonly used with this technique is M -ary frequency shift keying (MFSK), where $k = \log_2 M$ information bits are used to determine which

are assumed
a jammer en-
vely. Because
26) not equal
in which the
 b_j is chosen,
ding gives the
is known as

(12.27)

used by the
(taken to be
the code chip

(12.28)

n the system.
N-code pulse

ejection, fine-
performance
The modula-
ally BPSK or
K code chips.
such a signal-
ch a spread-
narrowband
makes a deci-
that; the 1000
ecision on the
decisions, you
still detect the

opping (FH).
frequency shift
ermine which

s Chap. 12

one of M frequencies is to be transmitted. The position of the M -ary signal set is shifted pseudorandomly by the frequency synthesizer over a hopping bandwidth W_{ss} . A typical FH/MFSK system block diagram is shown in Figure 12.11. In a conventional MFSK system, the data symbol modulates a *fixed frequency* carrier; in an FH/MFSK system, the data symbol modulates a carrier whose frequency is *pseudorandomly* determined. In either case, a single tone is transmitted. The FH system in Figure 12.11 can be thought of as a two-step modulation process—data modulation and frequency-hopping modulation—even though it can be implemented as a single step whereby the frequency synthesizer produces a transmission tone based on the simultaneous dictates of the PN code and the data. At each frequency hop time, a PN generator feeds the frequency synthesizer a frequency word (a sequence of ℓ chips), which dictates one of 2^ℓ symbol-set positions. The frequency-hopping bandwidth W_{ss} , and the minimum frequency spacing between consecutive hop positions Δf , dictate the minimum number of chips necessary in the frequency word.

For a given hop, the occupied transmission bandwidth is identical to the bandwidth of conventional MFSK, which is typically much smaller than W_{ss} . However, averaged over many hops, the FH/MFSK spectrum occupies the entire spread-spectrum bandwidth. Spread-spectrum technology permits FH bandwidths of the order of several gigahertz, which is an order of magnitude larger than implementable DS bandwidths [8], thus allowing for larger processing gains in FH compared to DS systems. Since frequency hopping techniques operate over such wide bandwidths, it is difficult to maintain phase coherence from hop to hop. Therefore, such schemes are usually configured using noncoherent demodulation. Nevertheless, consideration has been given to coherent FH in Reference [9].

In Figure 12.11 we see that the receiver reverses the signal processing steps of the transmitter. The received signal is first FH demodulated (dehopped) by mixing it with the same sequence of pseudorandomly selected frequency tones that was used for hopping. Then the dehopped signal is applied to a conventional bank of M noncoherent energy detectors to select the most likely symbol.

Example 12.1 Frequency Word Size

A hopping bandwidth W_{ss} of 400 MHz and a frequency step size Δf of 100 Hz are specified. What is the minimum number of PN chips that are required for each frequency word?

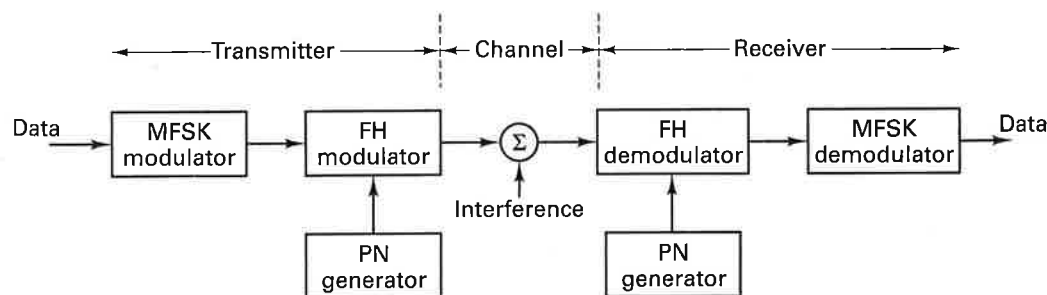


Figure 12.11 FH/MFSK system.

Solution

$$\begin{aligned} \text{Number of tones contained in } W_{ss} &= \frac{W_{ss}}{\Delta f} = \frac{400 \text{ MHz}}{100 \text{ Hz}} \\ &= 4 \times 10^6 \end{aligned}$$

$$\begin{aligned} \text{Minimum number of chips} &= \lceil \log_2 (4 \times 10^6) \rceil \\ &= 22 \text{ chips} \end{aligned}$$

where $\lceil x \rceil$ indicates the smallest integer value not less than x .

12.4.1 Frequency Hopping Example

Consider the frequency hopping example illustrated in Figure 12.12. The input data consist of a binary sequence with a data rate of $R = 150$ bits/s. The modulation is 8-ary FSK. Therefore, the symbol rate is $R_s = R/(\log_2 8) = 50$ symbols/s (the symbol duration $T = 1/50 = 20$ ms). The frequency is hopped once per symbol, and the hopping is time synchronous with the symbol boundaries. Thus, the hopping rate is 50 hops/s. Figure 12.12 depicts the time-bandwidth plane of the communication resource; the abscissa represents time, and the ordinate represents the hopping bandwidth, W_{ss} . The legend on the right side of the figure illustrates a set of 8-ary symbol-to-tone assignments. Notice that the tone separation specified is $1/T = 50$ Hz, which corresponds to the minimum required tone spacing for the orthogonal signaling of this noncoherent FSK example (see Section 4.5.4).

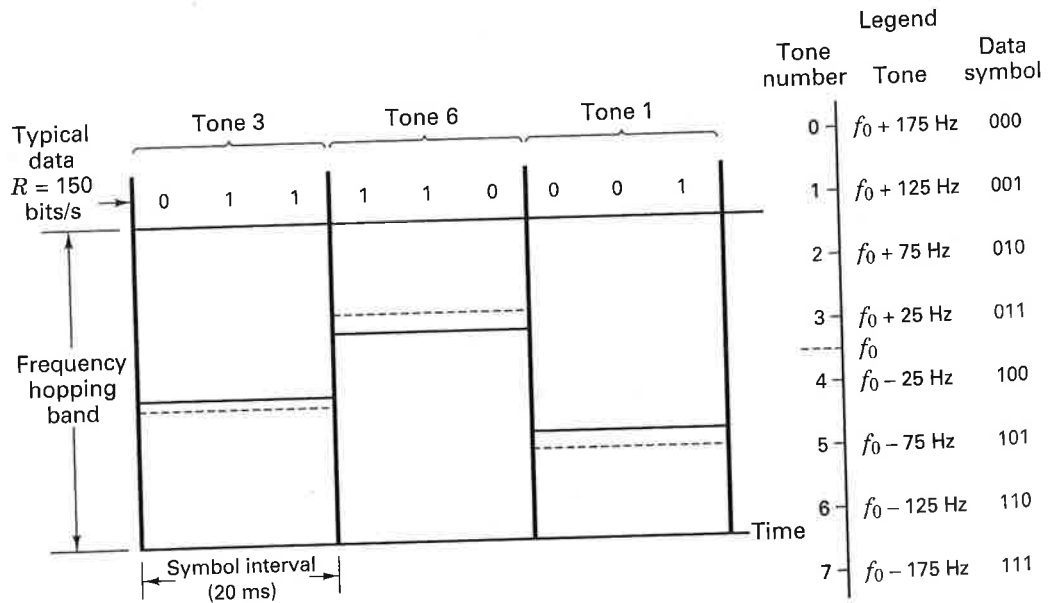


Figure 12.12 Frequency-hopping example using 8-ary FSK modulation.