A Structured Overview of Digital Communications—A Tutorial Review—Part II

BERNARD SKLAR

IN THE first part of this two-part paper, a structure and hierarchy were developed for tracing the key signal processing steps of a typical digital communications system. With the structure as a guide, formatting, source coding, and modulation transformations were examined. Also treated were potential trade-offs for power-limited and bandwidth-limited systems. In Part II, the signal processing overview continues with channel coding, multiplexing and multiple access, frequency spreading, encryption, and synchronization. To complete the overview, fundamental link analysis relationships are reviewed in the context of a satellite repeater channel.

In Part I of this paper (August 1983, *IEEE Communications Magazine*), a block diagram was introduced for a typical digital communications system; it is repeated here in Fig. 1. Also in Part I, the basic signal processing functions or transformations were classified into seven basic groups: formatting and source coding, modulation, channel coding, multiplexing and multiple access, frequency spreading, encryption, and synchronization. Figure 2 illustrates these transformations; formatting, source coding, and modulation were treated in Part I. Also treated were trade-offs among probability of bit error (P_B), bit energy per noise density (E_b/N_0), and bandwidth efficiency (R/W). The remainder of the signal processing steps outlined in Figs. 1 and 2 are treated here in Part II.

Channel Coding

Channel encoding (see Figs. 1 and 2) refers to the data transformation, performed after source encoding but prior to

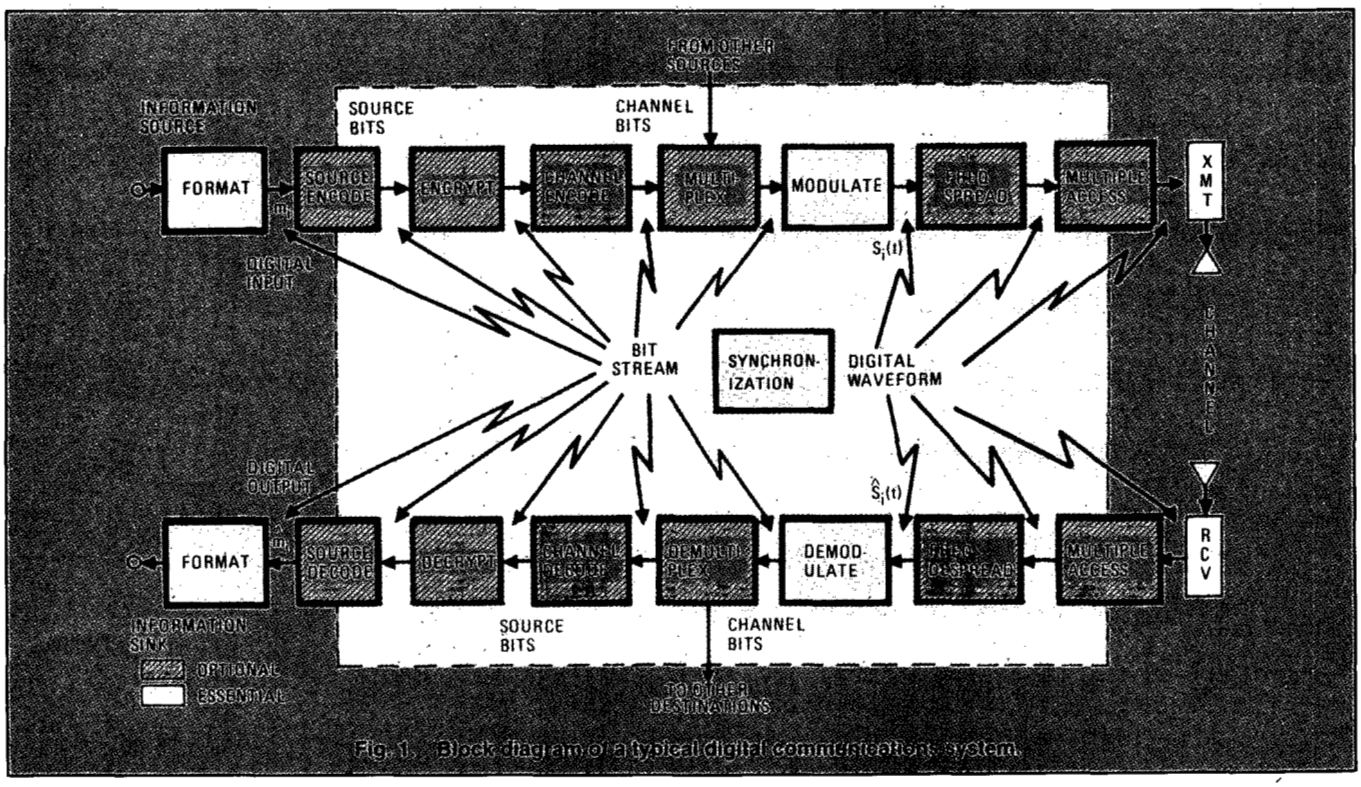


Fig. 1. Block diagram of a typical digital communications system.

0163-6804/83/1000-0006 $01.00 © 1983 IEEE

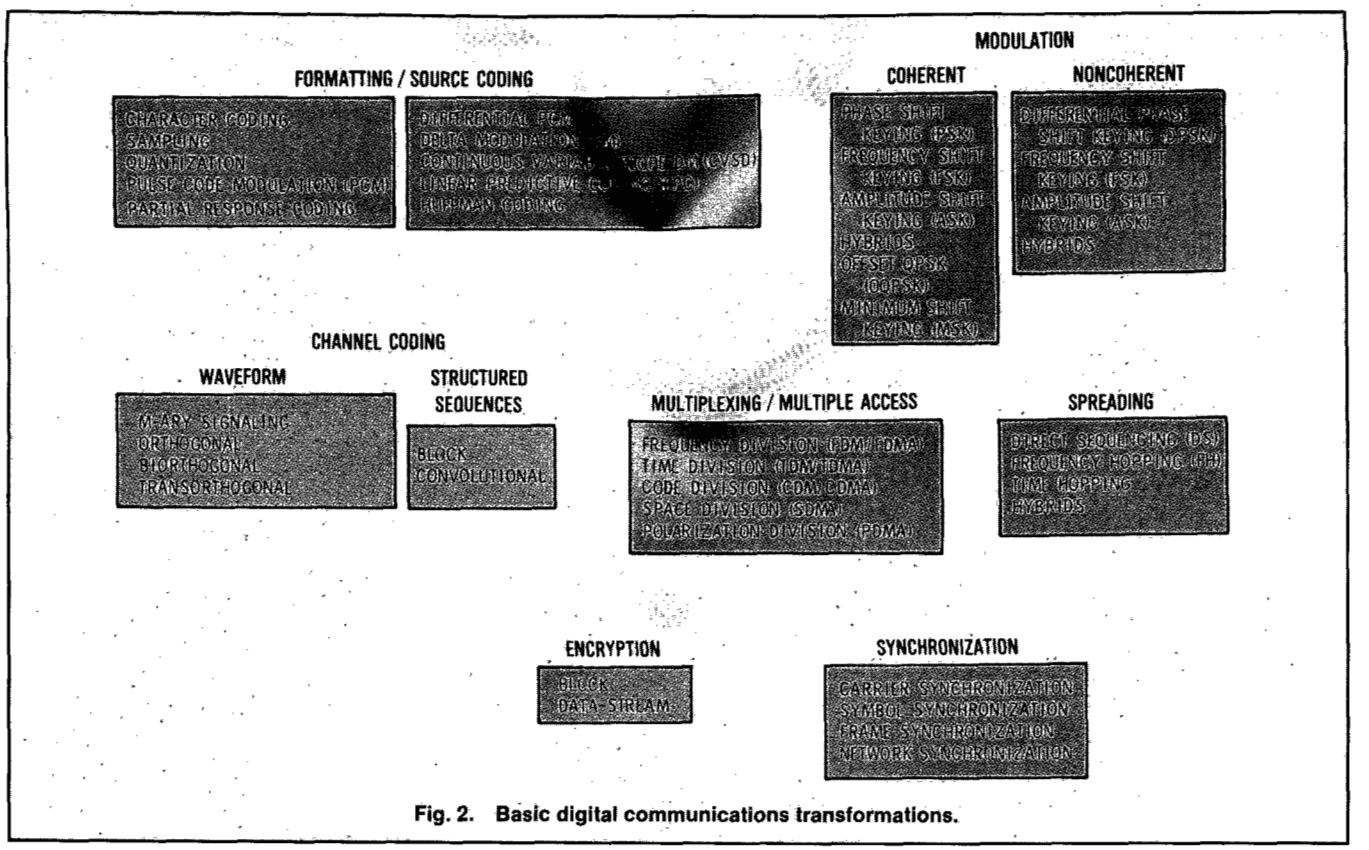


Fig. 2. Basic digital communications transformations.

modulation, that transforms source bits into channel bits. Channel coding is partitioned into two groups, waveform coding and structured sequences (see Fig. 2). Waveform (or signal design) coding is herein defined to mean any source data transformation that renders the detection process less subject to errors and thereby improves transmission performance. It can best be viewed as a transformation that demonstrates increased performance in an overall or "gestalt" sense, because the encoding produces a set of signals with better distance properties than the original signal set. The structured sequences category, by comparison, improves performance by embedding the data with structured redundancy, which may then be used to detect and correct transmission errors.

Waveform Coding

In Part I, M-ary signaling was described as a modulation or coding scheme that processes k bits at a time. The system directs the modulator to choose one of its $M=2^k$ waveforms for each k-bit sequence, where M is the symbol-set size, and k is the number of binary digits that each symbol represents. M-ary signaling alone, for the case in which $k > 1$, can be regarded as a waveform coding procedure that affects system performance. Orthogonal signals manifest improved P_B at the expense of bandwidth, as k increases; nonorthogonal signals manifest improved bandwidth efficiency (R/W) at the expense of power or P_B performance. R/W is measured in bits per second (b/s) per Hz; for binary signaling, the typical value of bandwidth efficiency is approximately 1 b/s/Hz. However, present-day multiphase shift keying (MPSK) and quadrature-amplitude modulation (QAM) systems frequently have efficiencies of 3 b/s/Hz and higher [1,2]. See Part I of this paper for a discussion of trade-offs among P_B, E_b/N_0, and R/W.

Another example of waveform coding is the use of an improved signal set as replacement for the original data symbols. The most popular of these codes are referred to as orthogonal and biorthogonal signal sets. The orthogonal signal set $s_i(t)$, where $i = 1, 2, \ldots, M$, is said to be orthogonal if and only if

$$z_{ij} = 1/E \int_0^T s_i(t)\, s_j(t)\, dt = \begin{cases} 1 & (\text{for } i = j) \\ 0 & (\text{otherwise}) \end{cases} \quad (1)$$

where it is assumed that all M signals have equal energy E, and that T is the symbol duration. Just as M-ary signaling with an orthogonal modulation format such as multifrequency shift keying (M-ary FSK, or MFSK) improves the P_B performance with increasing k, so too, coding with an orthogonally constructed signal set, prior to MPSK modulation, produces the same improvement.

A biorthogonal signal set can be obtained from an orthogonal set of $M/2$ signals by augmenting it with the negative of each signal. The biorthogonal set consists of a combination of orthogonal and antipodal signals. Since antipodal signal vectors have better distance properties than orthogonal ones (see Part I, Fig. 3), biorthogonal codes perform slightly better than orthogonal ones. With respect to

z_{ij} of (1), biorthogonal codes can be characterized as follows:

$$\begin{aligned} z_{ij} &= 1 \quad (\text{for } i = j) \\ &= -1 \quad (\text{for } i \neq j,\ |i-j| = M/2) \\ &= 0 \quad (\text{for } i \neq j,\ |i-j| \neq M/2). \end{aligned}$$

For completeness, a code generated from an orthogonal set by deleting the first digit of each code word is called a transorthogonal or simplex code. Such a code represents the minimum energy equivalent (in the P_B sense) of the equally likely orthogonal set. In comparing the performance of orthogonal, biorthogonal, and simplex codes, we can state that simplex coding requires the minimum signal-to-noise ratio (SNR) for a specified bit error rate. However, for large k, all three schemes are essentially identical in performance as they approach the Shannon limit. Biorthogonal coding requires half the bandwidth of the others. However, for each of these codes, bandwidth requirements (and system complexity) grow exponentially with the value of k.

Structured Sequences (Linear Block Codes)

Channel coding with structured sequences represents a method of inserting structured redundancy into the source data so that transmission errors can be identified. Structured sequences are partitioned into two important subcategories: block coding and convolutional coding (see Fig. 2). With block coding, the source data is first segmented into blocks of k data bits each; each block can represent any one of $M=2^k$ distinct messages. The encoder transforms each message block into a larger block of n digits. This set of 2^k coded messages is called a code block. The $(n\text{-}k)$ digits, which the encoder adds to each message block, are called redundant digits; they carry no new information. The ratio of data bits to total bits within a block, k/n, is called the code rate. The code itself is referred to as an (n,k) code.

Suppose that $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, where $c_j = 1$ or 0, is the transmitted code word, from a set of $i = 1, \ldots, M$ code words, and that $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ is the sequence received at the decoder input. If all code words have an equal likelihood of being transmitted, the optimum decoding scheme to use is called maximum likelihood decoding; it is similar to the optimum demodulation scheme under similar *a priori* assumptions. The decoder computes the conditional probability $P(\mathbf{r} \mid \mathbf{c}_i)$ for all 2^k code words. The code word $\mathbf{c}_t$ is identified as the transmitted word if $P(\mathbf{r} \mid \mathbf{c}_t)$ is the maximum of the computed probabilities.

The performance improvement possible with channel coding can be illustrated with the following example of a (15,11) single error-correcting code. Note that the notation (15,11) means that each block of 15 bits comprises 11 data bits and 4 redundant bits. Consider the following uncoded transmission. Assume BPSK modulation: $S/N_0 = 43\,776$; data rate $R = 4800$ b/s; P_B^U and P_M^U represent the uncoded probabilities of bit error and message error, respectively. P_B^C and P_M^C represent the coded probabilities of bit error and message error, respectively.

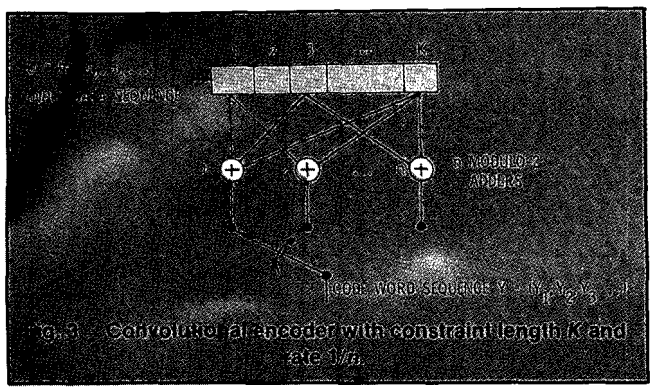


Fig. 2. Convolutional encoder with constraint length K and rate $1/n$.

Without Coding

$$\begin{aligned} E_b/N_0 &= S/RN_0 = 9.12\ (= 9.6\ \text{dB}) \\ P_B^U &= Q(\sqrt{2E_b/N_0}) = Q(\sqrt{18.24}) \\ &= 1.02 \times 10^{-5}\ (\text{see the section on Demodulation in Part I}) \end{aligned} \tag{2}$$

where

$$\begin{aligned} Q(x) &\cong (1/x\sqrt{2\pi})\exp(-x^2/2) \text{ for } x > 3 \\ P_M^U &= 1 - (1 - P_B^U)^{11} = 1.12 \times 10^{-4}. \end{aligned} \tag{3}$$

With Coding

The coded bit rate R_C is 15/11 times the data bit rate.

$$\begin{aligned} R_C &= 4800 \times 15/11 \cong 6545\ \text{b/s} \\ E_b/N_0 &= S/R_C N_0 = 6.688\ (= 8.25\ \text{dB}). \end{aligned}$$

The E_b/N_0 for the coded bit is a little less than for the uncoded bit because the bit rate has increased but the transmitter power is assumed to be fixed.

$$P_B^C = Q(\sqrt{2E_b/N_0}) = Q(\sqrt{13.38}) = 1.36 \times 10^{-4}. \tag{4}$$

It can be seen by comparing (2) and (4) that the bit error rate has degraded; more bits must be detected during the same time interval with the same available power; the performance improvement due to the coding is not yet apparent. We now compute the coded message error-rate P_M^C, as follows

$$P_M^C = \sum_{k=2}^{15} \binom{15}{k} (P_B^C)^k (1 - P_B^C)^{15-k}.$$

The summation is started with $k = 2$, since the code corrects all single errors within a block of $n = 15$ bits. A good approximation is obtained by using only the first term of the summation. For P_B^C we use the value computed in (4).

$$P_M^C \cong \binom{15}{2} (P_B^C)^2 (1 - P_B^C)^{13} = 1.94 \times 10^{-6}. \tag{5}$$

Equation (5) yields the message error-rate for the block of 15 coded bits. In this typical example, it can be seen by comparing (3) with (5) that the probability of message error has improved by a factor of 58 through the use of a block code.

Most of the research on block codes has been concentrated on a subclass of linear codes known as cyclic codes. A cyclic code word, after any number of end-around cyclic shifts, has the property of remaining a valid code word from the original

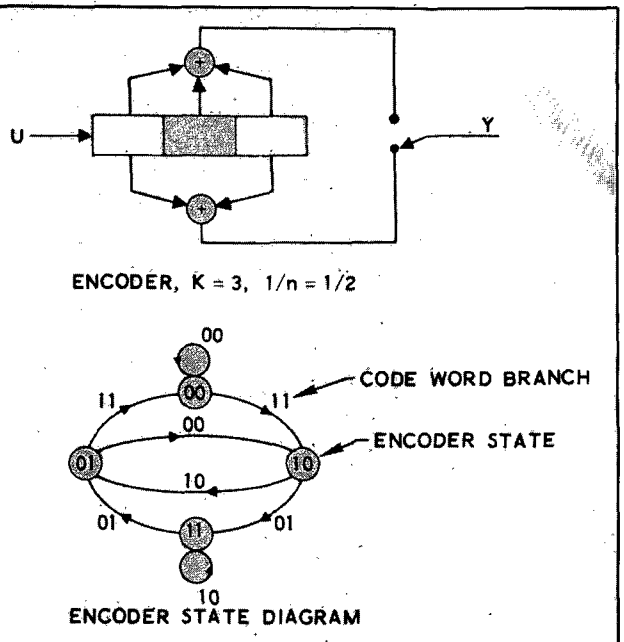


Fig. 4. Characterization of encoder by state diagram.

set of code words. Cyclic codes are attractive because they can be easily implemented with feedback shift registers. The decoding methods are simple and efficient, and generally use the same feedback shift registers as are employed for encoding [3,4].

Structured Sequences (Convolutional Coding)

A convolutional encoder convolves an input data sequence of bits with an encoding function as shown in Fig. 3. The encoder is mechanized with a K stage shift register and n modulo-2 summers, where K is called the constraint length. The source is represented by the data bit sequence $U = (u_1, u_2, \ldots, u_i, \ldots)$. At the i^{th} unit of time, data bit u_i is shifted into the first stage; all bits in the register are shifted one stage to the right, and the output of the n summers is sequentially sampled and transmitted. Since there are n code bits per data bit, the code rate is $1/n$ in this case. For the general case, k bits at a time are shifted into the register, and the code rate is k/n. The n code bits occurring at time i constitute the i^{th} branch of the code word, $Y_i = (y_{1i}, y_{2i}, \ldots, y_{ni})$. The code word Y consists of the sequence of branches: $Y = (Y_1, Y_2, \ldots)$.

Let the state of the encoder at time i be defined as $X_i = (u_{i-1}, u_{i-2}, \ldots, u_{i-K+1})$, which is the contents of the right-most $K-1$ stages of the shift register. The i^{th} code word branch Y_i is completely determined by state X_i and the present input bit u_i; thus the state X_i represents the past history of the encoder in determining the encoder output. The encoder state is said to be Markov, in the sense that the probability $P(X_{i+1} \mid X_i, X_{i-1}, \ldots, X_0)$ of being in state X_{i+1}, given all previous states, depends only on state X_i, that is, the probability is equal to $P(X_{i+1} \mid X_i)$. One simple way to represent an encoder is by a state diagram as shown in Fig. 4 for the encoder with K=3. The states of the diagram represent the possible contents of the right-most $K-1$ stages of the encoder shift register. There are two transitions from each state, corresponding to the two possible input bits, and there is a code word branch associated with each transition. The state diagram can be used to obtain a transfer function, which in turn can be used to derive error probability bounds [5].

A convenient way of incorporating encoder time history into the state diagram is through the trellis diagram shown in Fig. 5. At each time unit, the trellis shows all possible transitions between states. There are two possible paths leaving each state, corresponding to the two possible values of the input data bit. By convention, a dashed line in the trellis corresponds to input data "1" and a solid line to input data "0". The output code-word branches corresponding to the transitions are also shown for the encoder of Fig. 4; they appear as labels on the trellis branches.

An optimal decoder makes the maximum likelihood estimate of the transmitted code word Y, given the observation Z. The decoder chooses code word $\hat{Y}$ if $P(Z \mid \hat{Y}) = \max_Y P(Z \mid Y)$. Since the noise is assumed to be independent

$$P(Z \mid Y) = \prod_{i=1}^{\infty} P(Z_i \mid Y_i) = \prod_{i=1}^{\infty} \prod_{j=1}^{n} P(z_{ji} \mid y_{ji})$$

where Y_i is the i^{th} branch of code word Y, Z_i is the i^{th} branch of the received sequence Z, z_{ji} is the j^{th} code bit of Z_i, and y_{ji} is the j^{th} code bit of Y_i, each branch comprising n coded bits. The decoder problem consists of choosing a path through the trellis of Fig. 5 (each possible path defines a code word) such

that $\prod_{i=1}^{\infty} \prod_{j=1}^{n} P(z_{ji} \mid y_{ji})$ is maximized.

Binary channels are characterized by the need to make hard decisions (two-level decisions) on the received code bits. Continuous channels are characterized by the ability to make soft decisions (multilevel decisions). A multilevel decision can be thought of as a decision with a confidence factor attached. The ability to carry soft decisions along during the decoding process results in better decoding performance (approximately 2 dB) than if hard decisions are made. Ultimately, for a digital system, all decisions must be converted to hard decisions, that is, "1" or "0".

A brute-force maximum-likelihood decoder calculates the likelihood of the received data on all the paths through the

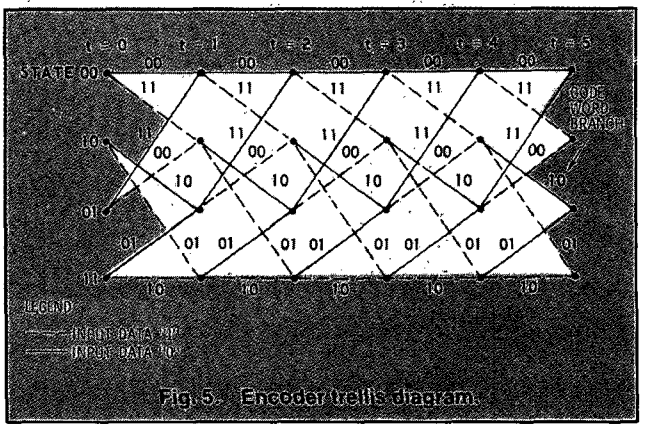


Fig. 5. Encoder trellis diagram.

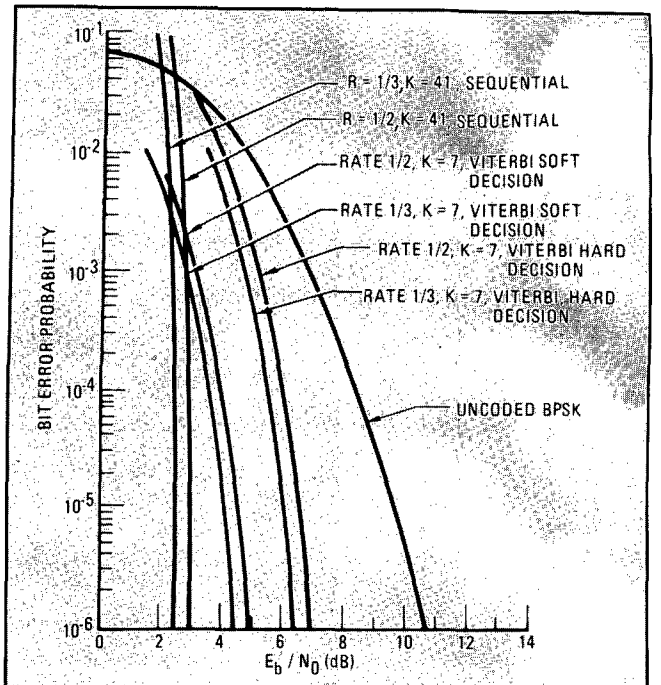


Fig. 6. Performance of Viterbi decoders and sequential decoders.

trellis. The number of paths for an L-bit information sequence is 2^L; the brute-force method becomes impractical as L increases. The Viterbi algorithm essentially performs maximum likelihood decoding; however, it reduces the computational load by taking advantage of the special structure in the code trellis. The advantage of Viterbi decoding (compared with brute-force) is that the decoder complexity is a linear rather than an exponential function of L [6]. The algorithm involves calculating a metric (measure of similarity) between the received signal (at time t_i) and all the trellis paths entering each state (at time t_i), where $i = 1, 2, \ldots$. In the event that two paths terminating on a given state are redundant, the one having the largest metric is stored (the surviving path). This selection of survivor is performed for all paths entering each of the other states. The decoder continues in this way to advance deeper into the trellis, making decisions by eliminating the least likely paths. Surviving paths need to be stored over an interval of about five constraint lengths to allow for coding delay. Storage requirements grow exponentially with constraint length; the present state of the art limits Viterbi decoders to about K=10. Viterbi decoders are very cost-effective for moderate error rates but cannot achieve very low error rates effectively. On the other hand, they are capable of very high speeds, where sequential decoders become uneconomical [7].

A sequential decoder works by generating hypotheses about the transmitted data sequence; it generates a metric between its hypotheses and the received signal. It goes forward as long as the metric remains within tolerance; otherwise, it goes backward, changing hypotheses until it finds an improved metric through a trial-and-error search. It can be implemented to work with hard or soft decisions, but soft decisions are usually avoided because they greatly increase the required storage and computations. Decoder complexity is relatively insensitive to the code constraint length; hence, constraint lengths are generally made very large (K=40), which is an important factor in providing such low P_B performance. The number of poor hypotheses and backward searches are a function of the SNR; with greater noise, more hypotheses must be generated. Because of this variability in computational load, buffer storage must be provided. Occasionally, these buffers will overflow, leaving a section of data uncoded. Therefore, an important part of a sequential decoder specification is the probability of buffer overflow.

The performance of these two popular solutions to the decoding problem, Viterbi decoding and sequential decoding, is illustrated in Fig. 6. The curves compare Viterbi decoding (rates 1/2 and 1/3 hard decision) versus Viterbi decoding (rates 1/2 and 1/3 soft decision) versus sequential decoding (rates 1/2 and 1/3). Figure 6 illustrates that, for $P_B = 10^{-5}$, coding gains of approximately 7 dB can be achieved with sequential decoders. Since Shannon's work foretold the potential of approximately 11 dB of coding gain at this error rate, it appears that the major portion of what is theoretically possible has already been accomplished.

Interleaving

Codes used for satellite channels are designed to combat independent errors; they are called random-error-correcting codes. There are also channels (for example, telephone lines, magnetic tape storage, troposcatter links, and sometimes satellite channels) on which the disturbances introduce errors that are clustered together in bursts. Use of an interleaver is a way of enhancing the random-error-correcting capabilities of a code, so that it is also useful in a burst-noise environment. The interleaver shuffles the encoded bits over a span of several block lengths (for block codes) and several constraint lengths (for convolutional codes). The span length required is determined from the need for error protection over some specified burst duration. The details of the bit redistribution pattern must be known to the receiver, as well as the transmitter, in order for the bit stream to be deinterleaved before being decoded. The overall result is to "spread out" the effect of burst noise so that induced errors appear to be independent (thereby matching the code's error-correcting capabilities).

Ramsey [8] discusses interleaver configurations that reorder a sequence of symbols so that no contiguous sequence of n_2 symbols in the reordered sequence contains any pair of symbols that were separated by fewer than n_1 symbols in the original ordering. He also shows that one such configuration is optimum in the sense of minimum possible coding delay and minimum possible storage capacity.

Multiplexing and Multiple Access

The terms multiplexing and multiple access both refer to the sharing of a communication resource (CR) (see Figs. 1 and 2). There is a subtle difference between them. With multiplexing, the system controller (which may be a human,

an algorithm, or even a wired logic board—either centralized or distributed) has instantaneous knowledge of all users' requirements or plans for CR sharing. There is no overhead needed to organize the resource allocation, and it is usually considered to be a process that takes place within the confines of a local site (for example, an assembly or a circuit board). Multiple access usually involves the remote accessing of a resource; additionally, there may be a finite amount of time required for the controller to become aware of each user's CR needs. Such time constitutes an overhead impact to the system utilization.

There are fundamentally two approaches to improving the ability of a satellite to support communications traffic. One way is to seek technological improvements toward increasing EIRP (effective radiated power referenced to isotropic), or to provide more bandwidth (there is great interest in developing the 30/20 GHz band for satellite communications). The second approach is to make the allocation of the CR more efficient. This second approach is the domain of communications multiple access.

The problem is to efficiently allocate portions of the satellite's fixed CR to a large number of users who seek to communicate digital information to each other at a variety of bit and message rates, and with various traffic requirements. A mechanism must be employed whereby the multiple signals can access the CR without creating interference to each other in the detection process. The avoidance of such interference requires that signals on one CR channel do not increase the probability of error in another channel. It should be obvious that orthogonality of the signals on separate channels suffices to avoid interference between users. Two signals are orthogonal if they can be described in the time domain by (1). Similarly, they are orthogonal if they can be described in the frequency domain by

$$\int_{-\infty}^{\infty} S_i(f)S_j(f)\,df = \begin{cases} 1 & (\text{for } i = j) \\ 0 & (\text{otherwise}) \end{cases} \qquad (6)$$

where the functions $S_i(f)$ are the Fourier transforms of some signal waveforms $s_i(t)$. Channelization characterized by (1) is called time division multiple access (TDMA), and that characterized by (6) is called frequency division multiple access (FDMA). In general, orthogonality can be achieved using code division multiple access (CDMA), a method involving both the time and frequency domains. In practice, CDMA offers user access flexibility but requires more complicated signal processing than either TDMA or FDMA [9].

A diagram of the time-frequency resource is shown in Fig. 7. We assume there are M users and that the total frequency bandwidth available is W Hz. M frequency bands of width W/M Hz are available and intervals of time duration τ_f/M seconds can be envisioned. We assume that the channel is time-synchronized so that periodic time intervals, called slots, are available. The slots are defined by

$$\text{Slot } (n,m) = \left[t:\ n\tau_f + (m-1)\tau_f/M \le t \le n\tau_f + m\tau_f/M\right]$$

where n and m are integers. The time interval $[n\tau_f, (n+1)\tau_f]$ is called a frame, and has duration τ_f seconds. The domain of the unit signal is the intersection of the slot, (n,m) and "band j" in Fig. 7. For any channelization of the CR, we assume that a modulation/coding system is chosen so that the full bandwidth W of the CR can support R b/s as the total available CR bit rate. In any subchannel having bandwidth W/M Hz, the associated bit rate will be R/M b/s.

Two additional access schemes useful for satellite communications are space division multiple access (SDMA) and polarization division multiple access (PDMA). To produce SDMA, the signals in different channels (allowed to occupy the same frequency band) are transmitted by using spot beam antennas. The spot beams produce orthogonality by physically separating the signals so they can be collected with physically separated receivers. To produce PDMA, the antennas are orthogonally polarized to separate the electromagnetic fields. A flexible implementation of SDMA, called satellite-switched TDMA (SS/TDMA), uses a microwave switch matrix in the satellite. The switching sequence of the matrix is controlled according to a programmable memory; the TDMA signals are cyclically interconnected among different antenna spot beams in rapid sequence. An earth station in the network communicates with those in other beams by transmitting TDMA bursts in proper timing to the sequence [10].

The multiple-access schemes discussed thus far would be termed fixed assignment for the case in which a user has access to the channel independent of his actual message traffic. By comparison, dynamic assignment schemes, sometimes called demand assignment multiple access (DAMA), give the user access to the channel only when he has a message to send. If the traffic from users tends to be burst-like or intermittent, then great efficiencies can be gained by using DAMA procedures to access the CR. The Intelsat IV satellite implemented a DAMA scheme called Single-channel Per carrier Access-on-Demand Equipment (SPADE) in the early 1970's. At each terminal, the SPADE subsystem responds to service requests by allocating an unused carrier frequency to the user; it then notifies the other terminals of its use through a common signaling channel. The initiating

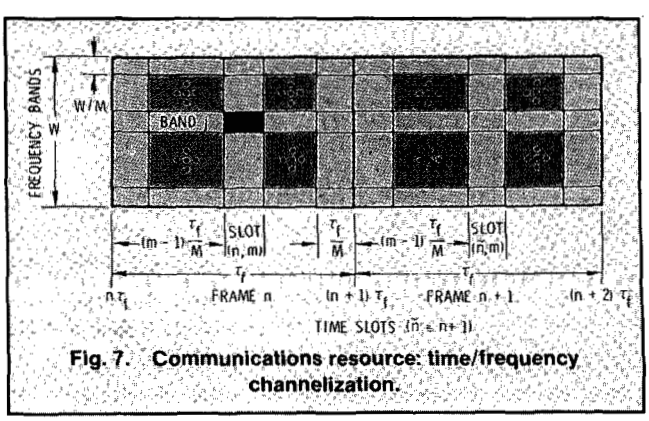


Fig. 7. Communications resource: time/frequency channelization.

terminal requests a frequency pair at random; such random selection makes it unlikely that two terminals will simultaneously request the same channel unless there are very few remaining. SPADE utilizes quadraphase shift keying (QPSK) modulation at 32 kilosymbols per second, using a signal bandwidth of 38 kHz for each 64-kb/s digitized voice channel (R/W = 1.68 b/s/Hz). This first important commercial satellite use of a DAMA scheme has resulted in more efficient use of power and bandwidth per channel than any of the fixed multiple-access schemes used earlier [11].

The development of packet switching techniques represents an important breakthrough in communications resource sharing. In circuit switched networks such as the telephone network, calls and message routing are set up prior to the commencement of message transmission. Once the route has been established, the message is transmitted on the dedicated circuit; after completion of the call, the circuit is disconnected. In packet communications, messages are packetized (partitioned into modular groups, each containing an address header). Each packet may be regarded as moving autonomously through the network, queueing at specific nodal points together with packets from other traffic. The key feature of packet switching systems is the potential for very efficient utilization of a communications or computer network, especially in the presence of bursty (high peak-to-average) traffic. Bhargava *et al* [12] present a concise summary of performance features and various access methods characterizing packet satellite networks.

For satellite communications, an important design consideration is the selection of signaling techniques suitable for the multiple access of a wideband hard-limiting repeater satellite; there are many alternatives in choosing multiple access schemes [13]. References [14] and [15] are good tutorials on the subject of multiple access for satellite systems. Nirenberg and Rubin [14] present a particularly interesting relationship between message delay and carrier power-to-noise density, as a function of bit error rate.

Frequency Spreading

Spread spectrum techniques (see Figs. 1 and 2) allow multiple signals occupying the same RF bandwidth to be simultaneously transmitted without interfering with one another. The technique is used for applications such as privacy, signal covertness, interference rejection, time delay or ranging measurements, selective addressing, and multiple access (CDMA) [16]. Figure 8 illustrates a spread spectrum system in its most general form. A carrier given by $A \cos \omega_o t$ is shown modulated with information to produce a signal $s_1(t)$, where

$$s_1(t) = A_1(t) \cos(\omega_o t + \phi_1(t)).$$

No restriction has been placed on the type of modulation that can be used. $s_1(t)$ is now multiplied by some code function $g_1(t)$. (Frequently, each code function is kept secret, and its use is restricted to a community of authorized users.) The resulting signal $g_1(t)s_1(t)$ is transmitted over the channel. At the same time, other users have multiplied their signals by other code functions. The signal present at the receiver is the linear combination of the emanations from each user

$$g_1(t)s_1(t) + g_2(t)s_2(t) + \ldots + g_n(t)s_n(t). \qquad (7)$$

Multiplication of $s_1(t)$ by $g_1(t)$ produces a signal whose spectrum is the convolution of the spectra of the two component signals. Thus, if the signal $s_1(t)$ is relatively narrowband compared with the code or spreading signal $g_1(t)$, the product will have nearly the bandwidth of $g_1(t)$. Assume that the receiver is configured to receive messages from user number 1. The first stage of the receiver multiplies the incoming signal of (7) by $g_1(t)$. The output of the multiplier will yield the following terms:

$$\begin{aligned} &\text{Wanted signal:} && g_1^2(t)\, s_1(t) \\ &\text{Unwanted signals:} && g_1(t)\, g_2(t)\, s_2(t) + g_1(t)\, g_3(t)\, s_3(t) \\ & && \quad + \ldots + g_1(t)\, g_n(t)\, s_n(t). \end{aligned} \qquad (8)$$

If the code functions $g_i(t)$, where $i = 1, 2, \ldots, n$, are chosen with orthogonal properties, then the desired signal can be extracted perfectly, and the unwanted signals yielding zero terms are easily rejected.

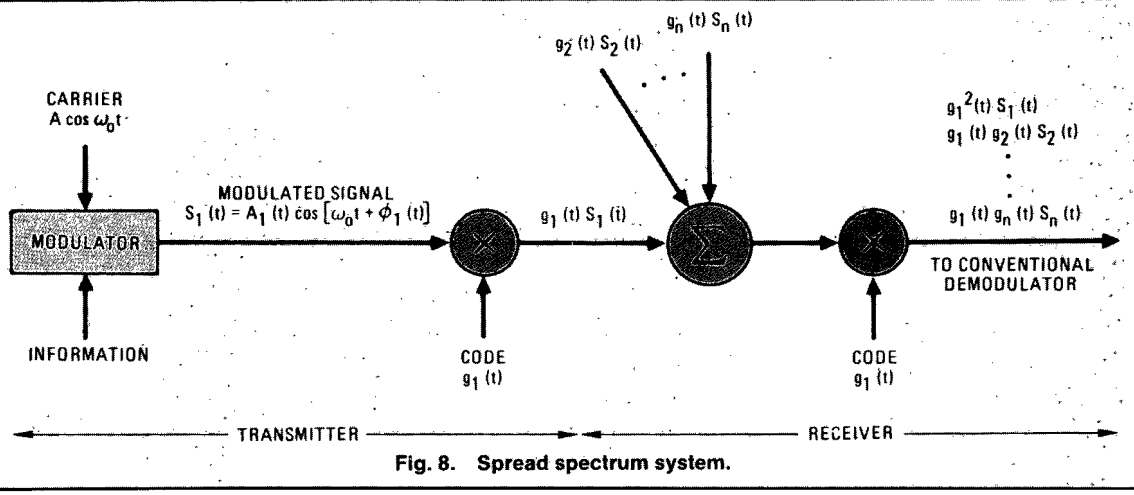


Fig. 8. Spread spectrum system.

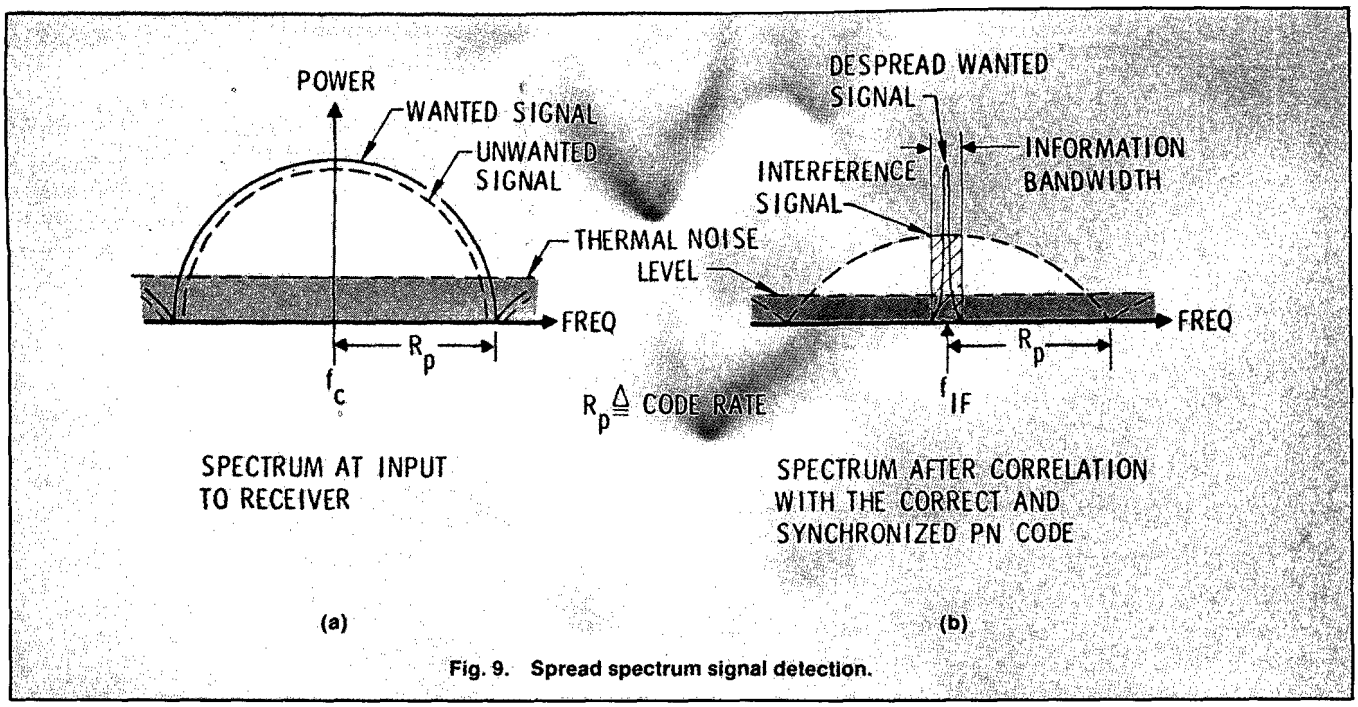


Fig. 9. Spread spectrum signal detection.

Figure 9(a) illustrates the wideband input to the receiver; it consists of wanted and unwanted signals, each spread by its own code, with code rate R_p, and each having a spectrum of the form $(\sin^2 x)/x^2$. Figure 9(b) illustrates the spectrum after correlation with the code $g_1(t)$ (despreading). The unwanted signals of (8) remain effectively spread by $g_1(t)g_i(t)$ where $i = 2, 3, \ldots, n$. Only that portion of the spectrum of the unwanted signals falling in the information bandwidth of the receiver will cause interference to the wanted signal.

If there is any jamming signal at the receiver (intentional or otherwise), the spreading signal will affect it just as it did the original signal at the transmitter. Thus, even a narrowband jamming signal in the middle of the information band will be spread to the bandwidth of the spreading signal; call it W. If the power of the jamming signal is J watts, its average density can be treated as wideband noise $J_0 = J/W$ watts/Hz. If $s_1(t)$ has power S watts and if the data rate is R b/s, the received energy per bit is $E_b = S/R$ watts-s, and the parameter E_b/J_0 which dictates the bit error rate performance in the presence of wideband noise jamming, can be written

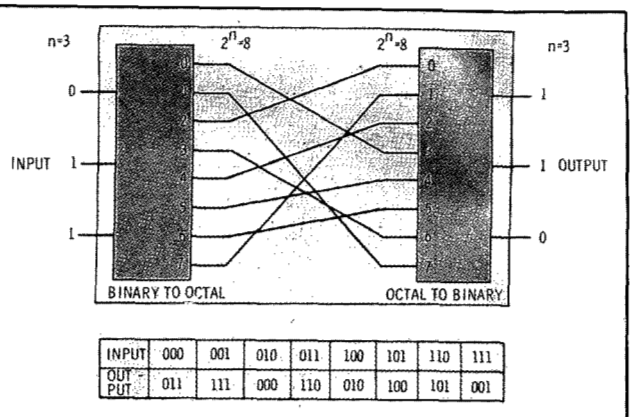
$$E_b/J_0 = (S/J)\,(W/R).$$

Thermal noise is present also, but we will assume that the jamming noise is so much greater than the thermal noise that we can neglect the latter. Hence, the ratio of jamming power to signal power is

$$J/S = (W/R)/(E_b/J_0). \tag{9}$$

This illustrates that if E_b/J_0 is the minimum ratio of bit energy to jamming noise density needed to support a given bit error rate, and if W/R is the ratio of spread bandwidth to the original data rate, also called the processing gain, then J/S is the maximum tolerable ratio of jamming power to signal power. It is commonly used as a figure of merit to describe a system's vulnerability to jamming; the larger the J/S, the greater the resistance against jamming. Another way of describing the relationship in (9) is as follows: An adversary would like to employ a jamming strategy so that the effective E_b/J_0 is as large as possible. He may try to employ pulse, tone, or partial band jamming rather than wideband noise jamming. A large E_b/J_0 would cause a small J/S for a fixed processing gain, or it would force the communicator to employ a larger processing gain for some desired J/S. The system designer strives to choose his waveform so that the jammer can gain no special advantage with a jamming strategy other than wideband noise.

There are two popular techniques for spectrum spreading (see Fig. 2). The first is called direct sequencing or pseudonoise spread spectrum. Spreading is achieved through the multiplication of the data by a binary pseudorandom sequence (discussed in the section on Encryption) whose symbol rate is many times the data rate. The second technique uses a frequency-hopping carrier. The carrier remains at a given frequency for a duration, and then hops to a new frequency somewhere in the spreading bandwidth W. Frequency hopping is generally classified as slow or fast hopping. In the case of slow hopping, there are typically several bits per hop, and the bandwidth of the transmitted signal is equal to that of the data signal. In the case of fast hopping, there are typically several hops per bit, and the bandwidth of the transmitted signal is equal to the reciprocal of the hopping duration. Figure 9, the spectral illustration of the spreading-despreading phenomenon, is an accurate rendition for each of the spreading techniques described. One important difference between direct sequencing and frequency hopping signals is that the former can be coherently demodulated. However, with frequency hopping, phase coherence is difficult to maintain; hence, it is usually demodulated noncoherently. The performance of spread spectrum systems is treated in detail in [17-20].



INPUT	000	001	010	011	100	101	110	111
OUTPUT	011	111	000	110	010	100	101	001

Fig. 10. Substitution box.

Encryption

Two reasons for using cryptosystems in communications (see Figs. 1 and 2) are as follows: 1) privacy, to prevent unauthorized persons from extracting information from the channel; and 2) authentication, to prevent unauthorized persons from injecting information into the channel. The message, or plaintext P, is encrypted with an invertible transformation E_k that produces the ciphertext $C = E_k(P)$. The ciphertext is transmitted over an insecure or public channel. When an authorized receiver obtains C, he decrypts it with the inverse transformation $D_k = E_k^{-1}$ to obtain

$$D_k(C) = E_k^{-1}(E_k(P)) = P$$

the original plaintext message.

E_k is chosen from a family of cryptographic transformations, frequently regarded as public information. The parameter k, or the key that selects the individual transformation within the family, is safeguarded; in typical cryptosystems, anyone with access to the key can both encrypt and decrypt messages. The key is transmitted to the community of authorized users over a secure channel (in some cases, a courier), and generally remains unchanged for a considerable number of transmissions. The goal of an eavesdropper or adversary (cryptanalyst) is to produce an estimate of the plaintext P, by analyzing the ciphertext obtained from the public channel, without benefit of the key.

Encryption schemes fall into two generic categories: block encryption and data-stream (or simply stream) encryption. With block encryption, the plaintext is segmented into blocks of fixed size; each block is encrypted independently from the others. A particular plaintext block will therefore be carried into the same ciphertext block each time it appears (as with block encoding). In general, however, the properties desired in a block cipher are quite different from those desired in an error-correcting code. For example, with encryption, plaintext data should never appear directly in the ciphertext; also, changing even a single bit of either the plaintext or the key should cause approximately 50% of the ciphertext bits to change.

Cryptosystems have their roots in Shannon's work [21] connecting cryptography with information theory. Shannon introduced the terms "confusion" (substitution) and "diffusion" (permutation). He suggested a method of using both of these in concert (a product cipher) to build a stronger encryption system than either method alone could produce. Figure 10 shows an example of a nonlinear substitution transformation. In general, n input bits are first represented as one of 2^n different characters (binary to octal translation in this example). A substitution is then made to one of the other characters from the set of 2^n characters. The character is then converted back to an n-bit output. It is easily shown that there are $(2^n)!$ different substitution or connection patterns possible. The cryptanalyst's task becomes computationally unfeasible as n gets large; say $n = 128$, then $2^n = 10^{38}$, and $(2^n)!$ = an astronomical number. We recognize that for $n = 128$, this substitution box (S box) represents the ideal encryption device. However, although we can identify the S box with $n = 128$ as ideal, its implementation is not feasible because it requires a unit with $2^n = 10^{38}$ connections.

Figure 11 represents an example of data permutation (a linear operation). Here the input data are simply rearranged or permuted (P box). The technique has one major disadvantage when used alone; it is vulnerable to trick messages. Such a message is illustrated in Fig. 11. A single "1" at the input and all the rest "0's" quickly reveals one of the internal connections; similar messages can be used to reveal each of the remaining connections. The product cipher of Fig. 12 illustrates the combination of substitution and permutation transformations originally suggested by Shannon. It represents the compromise solution to the difficulties

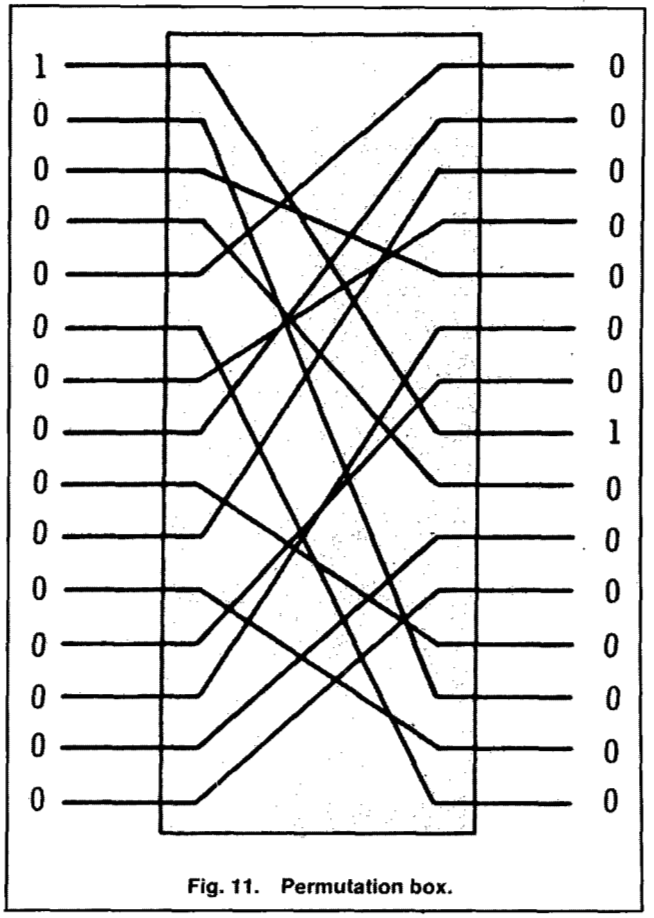


Fig. 11. Permutation box.

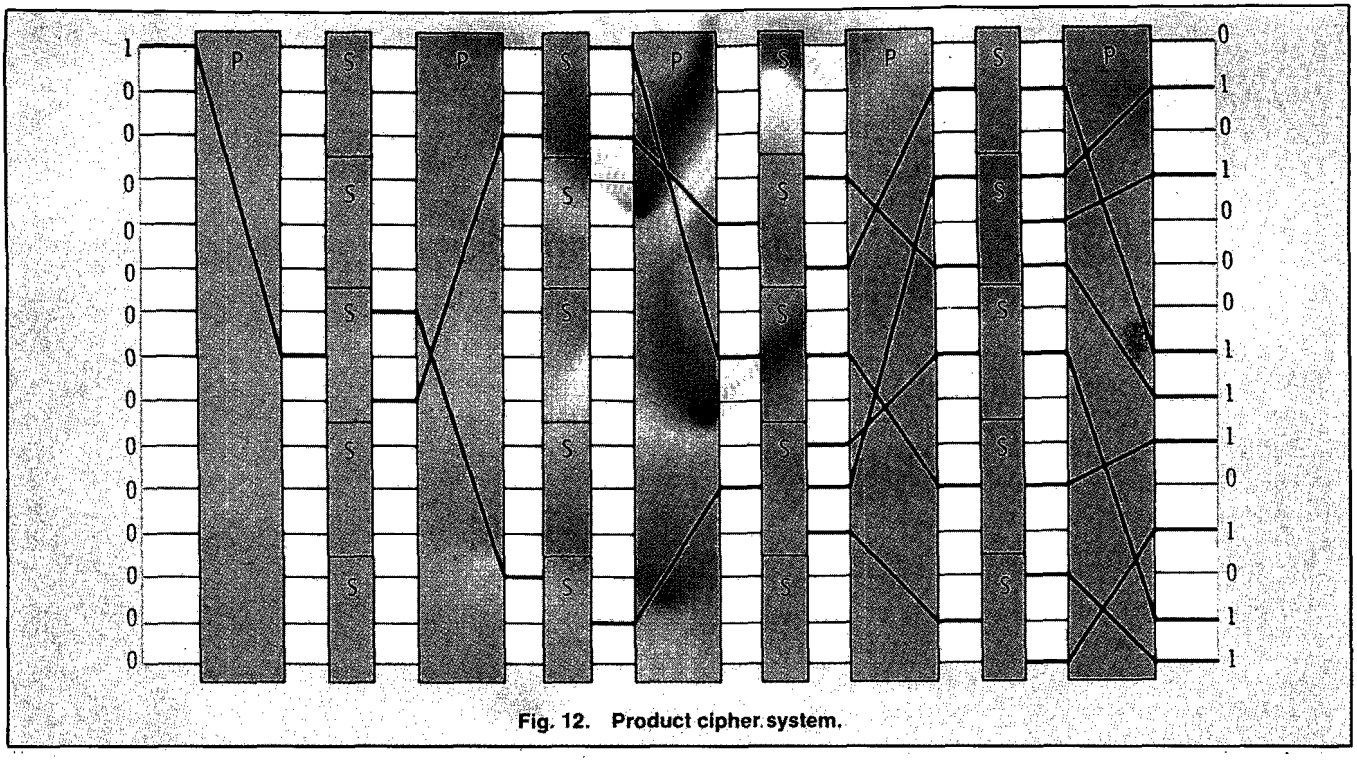


Fig. 12. Product cipher system.

with the S box or P box alone; the combination of S and P boxes yields a more powerful system than would either one alone. The basic product cipher concept was used by IBM in developing its Lucifer system. In 1977, the National Bureau of Standards adopted a modified Lucifer system as the national data encryption standard [22].

Data-stream ciphers do not treat the incoming symbols independently; encryption depends upon the internal state of the implementing device (feedback shift register). After each symbol is encrypted the device changes state according to some rule. Two occurrences of the same plaintext input will therefore typically not be encrypted into the same ciphertext. Stream encryption techniques generally employ shift registers for generating their pseudorandom key sequences. Such sequences derive their name from the fact that they appear random to the casual observer; they have statistical properties similar to the random flipping of a fair coin. However, the sequences, of course, are not random; they are precisely structured, as they must be to have any use for encryption and decryption. A shift register can be converted into a pseudorandom sequence generator by including a feedback loop that computes a new term for the first stage based on the previous n terms. An example is shown in Fig. 13, where $n = 4$, and feedback from stages 3 and 4 is modulo-2 added and returned to stage 1. If the initial stage of the register is 1 0 0 0, then the succession of states triggered by clock pulses would be 1 0 0 0, 0 1 0 0, 0 0 1 0, 1 0 0 1, 1 1 0 0, ... The output sequence is made up of the bits shifted out from the fourth register, that is, 0 0 0 1 0 0 1 1 0 1 0 1 1 1 1. Given any linear feedback shift register of degree n, the output sequence is ultimately periodic. Any output sequence achieving the maximum possible period, $p = 2^n - 1$ is called a maximum-length shift register sequence [3]. These sequences have the following randomness properties:

- *Balance property*—In each period of the sequence, the number of ones differ from the number of zeros by at most 1.
- *Run property*—Among the runs of ones and zeros in each period, one-half the runs of each kind are of length 1, one-fourth are of length 2, one-eighth are of length 3, and so on.
- *Correlation property*—If a period of the sequence is compared term by term, with any cyclic shift of itself, the

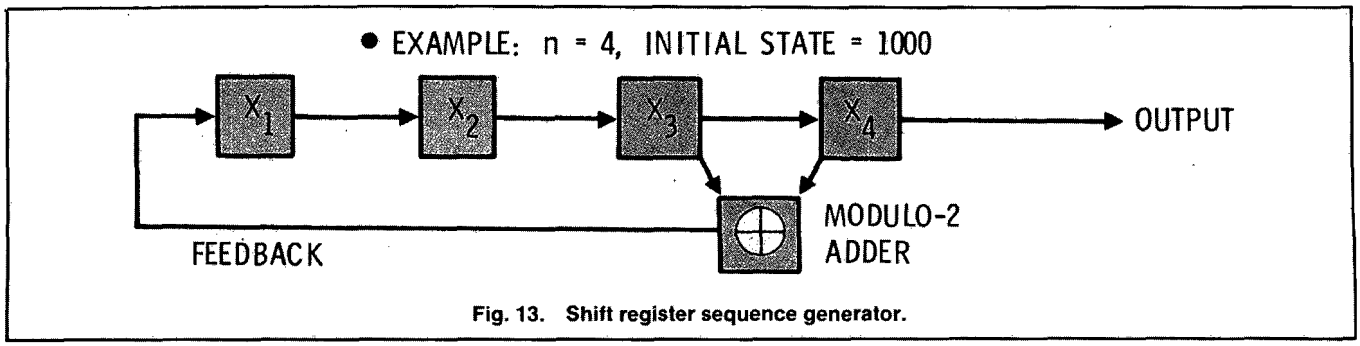


Fig. 13. Shift register sequence generator.

number of agreements differs from the number of disagreements by at most 1.

Maximum-length linear codes are not only used for compact encryption/decryption keys but also for many similar purposes, such as code sequences for CDMA schemes and code sequences for other frequency-spreading techniques. A linear shift register is very vulnerable to attack by a cryptanalyst. Even if the feedback taps are not known to him, the analyst needs only $2n$ bits of plaintext and ciphertext to learn the feedback taps, the initial state of the register, and the entire sequence of the code [23]. The use of nonlinear feedback in the shift register makes the cryptanalyst's task much more difficult, if not computationally impracticable.

Diffie and Hellman [23] provided a thorough treatment of encryption and decryption fundamentals in their 1979 paper; their bibliography and references constitute an excellent resource for further reading in the field. Kahn [24], in his 1967 book, presents a comprehensive and fascinating history of cryptography from ancient to modern times. Also, the *IEEE Communications Magazine* special issue on communications privacy is an invaluable primer [25].

Synchronization

Synchronization can be defined as the alignment of time scales of spatially separated periodic processes. In the context of digital communications, it involves the estimation of both time and frequency. The scope of this paper allows us space only to list and generally describe the synchronization requirements for digital systems. Throughout the signal processing discussions, we have dealt with the operation upon a digital symbol m_i or a digital waveform $s_i(t)$ during the time interval $[nT, (n+1)T]$, where $n = 0, 1, 2, \ldots$, are integers indexing each symbol time duration T, and $i = 1, \ldots, M$, are integers indexing individual symbols or waveforms from a finite set. An implicit assumption for each one of the processing steps (see Fig. 1) has been that the system is synchronized with respect to time and frequency, that the demodulator "knows" when to start accumulating energy for the decision-making process, and that it "knows" when to stop accumulating, when to make its decision, and when to repeat the operation. Any error in timing or frequency will result in lost energy, which effectively reduces the received E_b/N_0 and therefore degrades P_B performance.

The hierarchy of system synchronization levels (see Fig. 2) is as follows: Carrier synchronization refers to the restoration of the carrier (accurate with respect to frequency and phase) from a carrier-suppressed waveform. It is needed only for the demodulation of phase-coherent systems. Symbol (or bit) synchronization is needed for determining when the modulation may be changing state. Word, frame, or packet synchronization is needed for the proper reconstruction of the data. Network synchronization is needed for synchronizing channel access times among several users sharing the CR. Efficient signal design dictates that any discrete component of carrier or clock signal be suppressed; transmitted power is devoted exclusively to data. In this case, the system must recover the carrier and clock from a signal that contains neither one in explicit form. A survey by Franks [26] reviews some popular synchronizers and analysis techniques for implementing carrier and symbol synchronization. The survey demonstrates the application of maximum-likelihood estimation to practical signals; it also discusses commonly used circuits for approximating maximum-likelihood estimators. Other detailed treatments for implementing carrier and symbol synchronization can be found in [27] and [28].

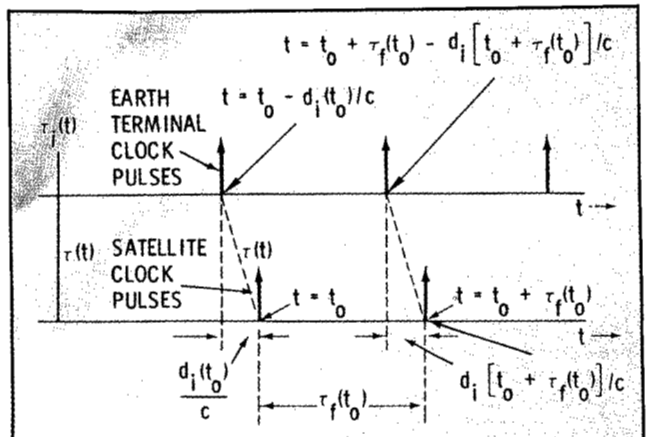


Fig. 14. Timing at satellite and at slaved terminal.

Figure 14 illustrates the general satellite and earth terminal considerations for maintaining a terminal clock slaved to the satellite clock (network synchronization). The i^{th} terminal must adjust its clock pulses and transmissions, so that they arrive in synchronism with the satellite clock pulses. Therefore, any transmission from the i^{th} terminal must be initiated early (with reference to the satellite clock) by a factor of $d_i(t)/c$, where $d_i(t)$ is the propagation distance from the satellite to the i^{th} terminal, and c is the speed of light.

Since range generally varies with time (either the satellite or the terminal may be moving with respect to one another), $\tau_i(t)$, the clock pulse sequence at the i^{th} terminal, cannot be maintained at a fixed pulse rate; therefore the frame duration $\tau_f(t)$ varies with time. For many cases, the terminal clock needs to be slaved continually to insure that it is in synchronism with the satellite clock. A good reference for synchronizing time between a satellite and an earth station is [29].

Communications Link Analysis

The communication system link budget is a balance sheet of gains and losses. It is composed of the detailed apportionment of transmission and reception resources, noise sources, and signal attenuators measured from the modulator and transmitter, through the channel, up to and including the receiver and demodulator (see Fig. 1). The budget is mainly derived from the calculation of received useful power. Some of the budget parameters are statistical, for example, RF propagation fades due to meteorological events [30, 31]. Link budget analysis is therefore an estimation technique for evaluating communications system performance.

The propagating medium or electromagnetic path connecting the transmitter and receiver is called the channel. The

concept of free space assumes a channel region free of all objects that might affect RF propagation by absorption, reflection, or refraction. It further assumes that the atmosphere in the channel is perfectly uniform and nonabsorbing, and that the earth is infinitely far away or its reflection coefficient negligible. The RF energy arriving at the receiver is assumed to be a function of distance from the transmitter (simply following the inverse square law of optics). In practice, of course, propagation in the atmosphere and near the ground results in refraction, reflection, and absorption that modify the free space transmission [32].

SNR is a convenient measure of performance at various points in the link. The definition is

$$\text{SNR} = \frac{\text{power in desired waveform}}{\text{power in interfering waveform}}.$$

The desired waveform can be an information signal, baseband waveform, or modulated carrier. A communications system primarily degrades: through the attenuation of desired waveform power relative to interfering waveform power, *or* through the increase of interfering waveform power relative to desired waveform power. These degradations are termed "loss" and "noise," respectively. Losses occur when, by some mechanism, a portion of the signal is diverted, scattered, or reflected from its intended route. Noise occurs when unwanted signal energy is injected into the link, or thermal noise is generated within the link. There are two main types of noise, thermal noise and intermodulation noise. Thermal noise is radiated into the antenna by oxygen and water vapor molecules in the atmosphere; it is also introduced by the first stages of the receiver. Intermodulation noise is caused by nonlinearities in the system; its deleterious effects are generally grouped quantitatively with other system losses, under the heading of a loss parameter. We will restrict our discussion of noise to thermal noise, in which case the power spectral density is assumed to be flat up through the GHz range; the thermal noise process in communications receivers is generally accepted to be an additive white Gaussian noise (AWGN) process [33]. A well-known physical model [34] for thermal or Johnson noise generated in dissipative components consists of a noise generator with open-circuit mean squared voltage equal to $4\,\kappa\,T°W\mathscr{R}$, where

κ = Boltzmann's constant 1.38×10^{-23} J/K
$T°$ = temperature in K
W = bandwidth in Hz
$\mathscr{R}$ = resistance in Ω

It can be shown that the maximum available thermal noise power N, coupled from the noise generator into the front end of an amplifier, is [33]

$$N = \kappa T° W \text{ (watts)}$$

and the noise density N_0 is simply

$$N_0 = N/W = \kappa T° \text{ (watts/Hz)} \tag{10}$$

Development of the fundamental link power relationships assumes an omnidirectional RF source transmitting uniformly over 4π steradians (isotropic radiator). The power density on a hypothetical sphere at a distance d from the source is related to the transmitter power P_t by

$$p(d) = P_t/4\pi d^2.$$

The power extracted with the receiving antenna can be written

$$P_r = p(d)\, A_{er} = P_t A_{er}/4\pi d^2 \tag{11}$$

where the parameter A_{er} is the absorption cross section (effective area) of the antenna defined by

$$A_{er} = \frac{\text{total power absorbed}}{\text{incident power flux density}}$$

The receiving antenna's effective and physical areas are related by the efficiency parameter, η, as follows

$$A_{er} = \eta A_{pr}$$

which accounts for the fact that the total power is not absorbed; some of it is lost through re-radiation, scattering, or spillover. Typical values for η are 0.55 for a dish and 0.75 for a horn.

A common antenna parameter that relates the power output (or input) to that of an isotropic radiator is the antenna gain G, where

$$G = \frac{\text{maximum power intensity in some fixed direction}}{\text{average power intensity over } 4\pi \text{ steradians}}.$$

Antenna gain, unlike that of an electronic amplifier, is the result of concentrating the RF flux in some restricted region less than 4π steradians; thus, the effective power radiated with respect to an isotropic source (EIRP) is defined as

$$\text{EIRP} = P_t G_t$$

where G_t is the transmitter antenna gain. To find received power for the general case in which the transmitting source manifests antenna gain over isotropic, we replace P_t with EIRP in (11)

$$P_r = \text{EIRP}\, A_{er}/4\pi d^2. \tag{12}$$

The relationship between antenna gain and antenna effective area is [35]

$$G = 4\pi A_e/\lambda^2 \qquad (\text{for } A_e >> \lambda^2) \tag{13}$$

where λ is the wavelength of the radiation. Similar expressions apply at the transmitter and receiver antennas by the reciprocity theorem [35]. Since the effect of an antenna can be expressed as a gain or area, we can replace A_{er} in (12) with $G_r\,\lambda^2/4\pi$ from (13), as follows

$$P_r = \text{EIRP}\, G_r\, \lambda^2/(4\pi d)^2 = \text{EIRP}\, G_r/L_s \tag{14}$$

In (14), the parameters $(4\pi d/\lambda)^2$ have been replaced by the term L_s, the space loss or path loss. Path loss characterizes the decrease in received power as a function of distance and frequency; it is a *definition* predicated on the use of an isotropic receiving antenna ($G_r = 1$). Hence, path loss is an abstraction that cannot be measured; it represents a

hypothetical received-power loss that *would occur if the receiving antenna were isotropic*. In a radio communications system, path loss accounts for the largest loss in signal power. In satellite systems, the path loss for a C-band (6 GHz) link to a synchronous satellite is typically 200 dB.

In evaluating system performance, the quantity of greatest interest is not the received power P_r, but the SNR. This is because the basic system constraint is our ability to detect the signal, with an acceptable P_B, in the presence of noise. Since the desired signal here is a modulated carrier waveform, we often speak of the average carrier power-to-noise ratio (C/N) or (P_r/N) as the SNR of particular interest. Into (14) we introduce P_r/N

$$\frac{P_r}{N} = \frac{\text{EIRP } G_r/N}{L_s}. \tag{15}$$

Equation (15) applies to any one-way satellite RF link. In analog systems, noise bandwidth is generally greater than signal bandwidth, and P_r/N is the main parameter for measuring signal detectability and performance quality. In digital receivers however, correlators or matched filters, where signal bandwidth is taken to be equal to noise bandwidth, are usually used. Rather than consider input noise power, a common formulation for digital links is to replace noise power with noise density. We can use (10) for rewriting (15) as follows

$$\frac{P_r}{N_0} = \frac{\text{EIRP } G_r/T^\circ}{\kappa L_s L_o} \tag{16}$$

where the system effective temperature, T°, is a function of the thermal noise radiated into the antenna and the thermal noise generated within the first stages of the receiver [36-38]. We have introduced a term L_o in (16) to represent all degradation factors due to various losses and noise sources; this term represents "other losses" not specifically addressed by the other terms of (15). It allows for a large assortment of different losses and noise sources (for example, intermodulation noise), which have been cataloged in detail [39]. Equation (16) summarizes the key parameters of any link analysis, which are: the received signal power-to-noise density (P_r/N_0), the magnitude of transmitted power (EIRP), the sensitivity of the receiver (G_r/T°), and the losses ($L_s L_o$).

If we assume that all the received power stems from the modulating signal (suppressed carrier), then we can write

$$P_r/N_0 = S/N_0 = (E_b/N_0)\, R. \tag{17}$$

If some of the received power is lodged in the carrier (a signal power loss), we can still employ (17), but we additionally represent the carrier power as a loss (within the parameter L_o of 16).

Until now, we have referred only to one kind of E_b/N_0, that value of bit energy-to-noise density *required* to yield a specified P_B. But now, to facilitate calculating a margin or safety factor M, we need to differentiate between the required E_b/N_0 and the actual or *received* E_b/N_0. From this point on we will refer to the former as $(E_b/N_0)_{\text{reqd}}$, and the latter as $(E_b/N_0)_r$. We can rewrite (17) introducing the link margin parameter M, as follows

$$P_r/N_0 = (E_b/N_0)_r\, R = M\, (E_b/N_0)_{\text{reqd}}\, R. \tag{18}$$

The difference in decibels between $(E_b/N_0)_r$ and $(E_b/N_0)_{\text{reqd}}$ yields the link margin. Consider a system specified to operate at an $(E_b/N_0)_{\text{reqd}} = 10$ dB, with $P_B = 10^{-4}$. Suppose we require a link margin of 4 dB (let us assume that the commensurate P_B for an E_b/N_0 of 14 dB is 10^{-6}). We can look upon this margin in one of two ways:

- We can state that we have 4 dB more E_b/N_0 than we actually need to meet our required P_B of 10^{-4}.
- We can state that we are operating at an E_b/N_0 of 14 dB and therefore that the actual operating P_B of the system is 10^{-6}, a margin of 100 times better error probability performance than we require.

The parameter $(E_b/N_0)_{\text{reqd}}$ reflects the differences from one system to another; these might be due to differences in modulation or coding schemes. A larger than expected $(E_b/N_0)_{\text{reqd}}$ may be due to a suboptimal RF system, which manifests large timing errors or allows more noise into the detection process than does an ideal matched filter.

Combining (16) and (18) and solving for the link margin M, yields

$$M = \frac{\text{EIRP } G_r/T^\circ}{(E_b/N_0)_{\text{reqd}}\, R\, \kappa\, L_s\, L_o}. \tag{19}$$

Since link budget analysis is typically calculated in decibels, we can express (19) as follows:

$$\begin{aligned} M\ (\text{dB}) = {} & \text{EIRP (dBW)} + G_r\ (\text{dBI}) - (E_b/N_0)_{\text{reqd}}\ (\text{dB}) \\ & - R\ (\text{dB-b/s}) - \kappa T^\circ\ (\text{dBW/Hz}) - L_s\ (\text{dB}) \\ & - L_o\ (\text{dB}). \end{aligned} \tag{20}$$

Transmitted signal power is expressed in decibel-watts (dBW); noise density is in decibel-watts per Hz (dBW/Hz); antenna gain is in decibels referenced to isotropic gain (dBI); data rate is in decibels referenced to b/s (dB-b/s); and all other terms are in decibels (dB). The values of the parameters in (20) constitute the link budget, a useful tool for allocating communications resources. In an effort to maintain a positive margin, we might trade off any parameter with any other parameter; we might choose to reduce transmitter power by giving up excess margin, or we might elect to increase the data rate by reducing $(E_b/N_0)_{\text{reqd}}$ (through the selection of improved modulation and coding). Any one of the decibels in (20), regardless from which parameter it stems, is just as good as any other decibel. It should be noted, however, that as requirements become more constrained, it may not be possible to trade or yield on some items. For example, even though binary PSK modulation outperforms binary FSK (in the P_B sense), requirements to operate in a scintillating environment would dictate the avoidance of PSK and the choice of the more robust FSK. Also, certain coverage requirements may constrain antenna dimensions, so that one might *not* have the freedom of trading-off or selecting any antenna gain one desires.

How Much Link Margin is Required

The question of how much link margin should be designed into the system is asked frequently. The answer is that if all sources of gain, loss, and noise have been rigorously detailed, and if the link parameters with large variances (for example, fades due to weather) match the statistical requirements for link availability, very little margin is needed. For satellite communications at C-band, where the parameters are well-known and fairly well-behaved, it should be possible to design a system with only 1 dB of link margin. Receive-only television stations operating with 16-ft-diameter dishes at C-band are frequently designed with only a fractional dB of margin. However, telephone communications via satellite using standards of 99.9% availability require considerably more margin; some of the Intelsat systems have 4 to 5 dB of margin. Designs using higher frequency (for example, 14/12 GHz) generally call for larger margins because atmospheric losses increase with frequency and are highly variable. It should be noted that a by-product of the attenuation due to atmospheric loss is greater antenna noise. When extra margin is allowed for weather loss, additional margin should simultaneously be added to compensate for the increase in antenna temperature (a function of thermal noise radiated into the antenna). With low-noise amplifiers, small weather changes can result in increases of 40° to 50° K in antenna temperature.

Satellite Repeaters

Satellite repeaters retransmit the messages they receive (with a translation in carrier frequency). A *regenerative* (digital) repeater regenerates, that is, demodulates and reconstitutes the digital information embedded in the received waveforms; however, a *nonregenerative* repeater only amplifies, but does not transform, the signal to its baseband format. A nonregenerative repeater, therefore, can be used with many different modulation formats (simultaneously), but a regenerative repeater is usually designed to operate with only one, or a very few, modulation formats. Link analysis for a regenerative satellite repeater treats the uplink and downlink as two separate point-to-point analyses. To estimate the performance of a regenerative repeater link, it is necessary to determine separately the bit error probability on the uplink and downlink. The overall error rate is obtained by simply summing the individual rates [40,41].

By comparison, link analysis for a nonregenerative repeater generally treats the entire "round-trip" (uplink transmission to the satellite and downlink retransmission to an earth terminal) as a single analysis. To estimate performance of a nonregenerative repeater link, the uplink and downlink values of E_b/N_0 (or P_r/N_0) are combined as follows, in the absence of intermodulation noise [40]

$$(E_b/N_0)\bar{}^{1}_{U} + (E_b/N_0)\bar{}^{1}_{D} = (E_b/N_0)\bar{}^{1}_{R}$$

where the subscripts U, D, and R indicate uplink, downlink, and resultant values of E_b/N_0 respectively.

Most conventional commercial satellites in use today are the simple nonregenerative kind. However, it seems clear that digital satellite systems of the future, which will require on-board processing, switching, or selective message addressing, will start with a regenerative repeater to transform the received waveforms to message bits. Besides the potential for sophisticated data processing, one of the principal advantages of regenerative compared to nonregenerative repeaters is the decoupling of the uplink and downlink so that the uplink noise power is not retransmitted on the downlink. There are significant performance improvements in terms of reducing the E_b/N_0 values needed on the uplinks and downlinks relative to the values needed for the conventional transponder designs in use today. Improvements as much as 5 dB on the uplink and 6.8 dB on the downlink (using coherent QPSK modulation, with $P_B = 10^{-4}$) have been demonstrated [40].

Power is severely limited in most satellite communications systems, and the inefficiencies associated with linear power amplification stages are intolerable. For this reason, many satellite repeaters employ highly nonlinear power amplifiers; the main feature here is that efficient power amplification is obtained at the cost of nonlinear distortions. The major undesirable effects of the repeater nonlinearities are:

- Intermodulation (IM) noise due to the multiplicative action of different carriers—The harm caused is twofold; useful power can be lost from the channel as IM energy, and spurious IM products can be introduced into the channel as interference [42,43].
- AM to PM conversion is a phase noise phenomenon occurring in nonlinear devices such as traveling wave tubes (TWT). Fluctuation in operating level (amplitude modulation) produces phase variations that impact the P_B performance, for systems using an MPSK modulation format [44,45].
- Signal suppression of weak signals by stronger signals [42], by as much as 6 dB.

Conventional nonregenerative repeaters are generally operated "backed-off" from their highly nonlinear saturated region; this is done to avoid appreciable IM noise and to thus allow efficient utilization of the system's entire bandwidth. However, backing off to the linear region is a compromise; some level of IM noise must be accepted in order to achieve a useful level of output power.

One set of features, unique to nonregenerative repeaters, is worth describing here; it is the dependence of the downlink SNR upon the uplink SNR, and the sharing of the repeater downlink power in proportion to the uplink power from each of the various uplink signals and noise. Henceforth, reference to a repeater or transponder will mean a nonregenerative repeater, and for simplicity we will assume the transponder is operating in its linear range.

A satellite transponder is limited in transmission capability by its downlink power, the earth terminal's uplink power, satellite and earth terminal noise, and channel bandwidth. One of these usually represents the dominant performance

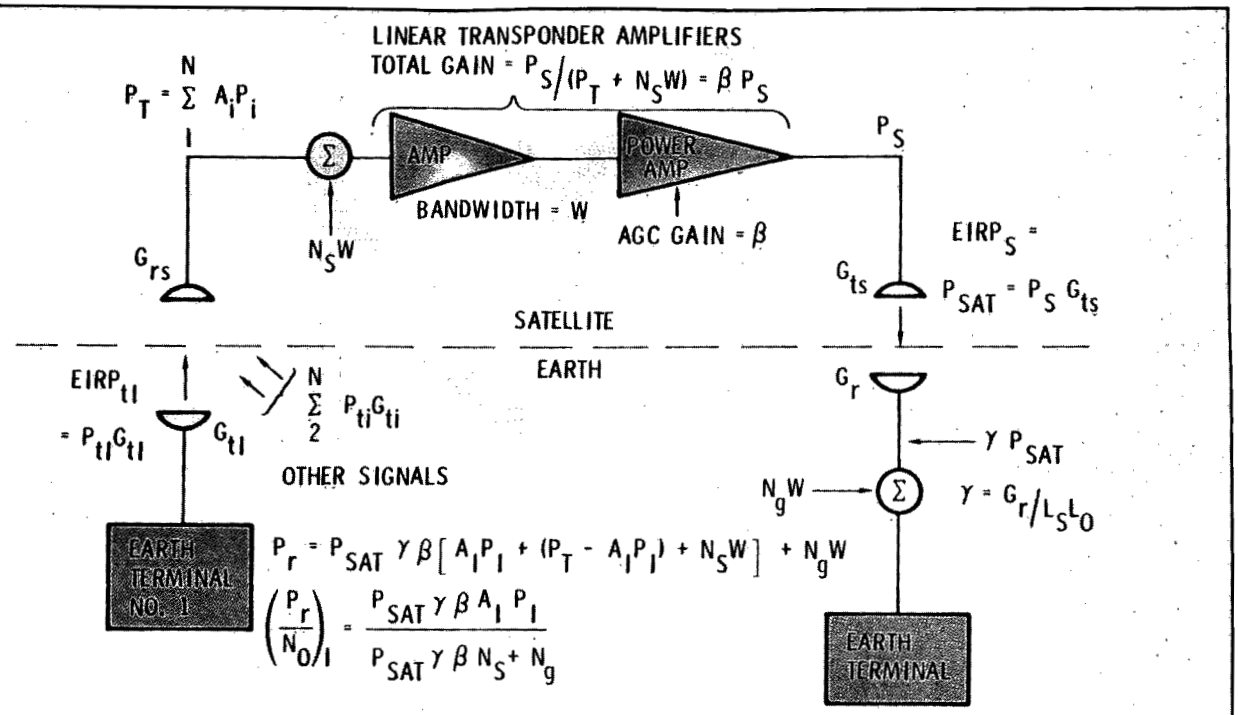


Fig. 15 Nonregenerative satellite repeater.

constraint; most often the downlink power or the channel bandwidth proves to be the major system limitation. Figure 15 illustrates the important link parameters of a linear satellite repeater channel. The repeater transmits all arriving uplink messages (or noise, in the absence of messages) without any processing beyond frequency translation. Let us assume that the multiple uplinks within the receiver's bandwidth W are separated from one another through the use of a multiple-access scheme such as FDMA or CDMA. The satellite downlink power P_{sat} is constant and, since we are assuming a linear transponder, P_{sat} is shared among the multiple uplink signals (and noise) in proportion to their respective power levels.

The transmission starts from a ground station (bandwidth $< W$), for instance, terminal number 1, with an $EIRP_{t1} = P_{t1}$ G_{t1}. Simultaneously, other signals are being transmitted to the satellite (from other terminals). At the satellite, a total signal power $P_T = \sum A_iP_i$ is received, where the A reflect the various propagation losses the different signals experience upon arrival at the satellite. Noise power N_sW is also received at the satellite, where N_s is the noise density due to thermal noise radiated into the satellite antenna and thermal noise generated in the satellite receiver. The total satellite downlink $EIRP_s = P_{sat}$ can be expressed with the following identity [46]

$$P_{sat} = P_{sat}\,\beta\left[A_1P_1 + (P_T - A_1P_1) + N_sW\right]$$

where $\beta = 1/(P_T + N_sW)$ is the AGC gain and P_T has purposely been written as $A_1P_1 + (P_T - A_1P_1)$ to separate signal number 1 power from the remainder of simultaneous signals in the transponder. Using (16), we can write the $(P_r/N_0)_{1i}$ for signal number 1 arriving at the i^{th} terminal, as follows [46]

$$\left(\frac{P_r}{N_0}\right)_{1i} = \frac{P_{sat}\,\gamma_i\,\beta\,A_1P_1}{P_{sat}\,\gamma_i\,\beta\,N_s + N_g} \tag{21}$$

where $\gamma_i = G_r/L_sL_o$ for the i^{th} terminal, and N_g is the receiver noise density for the i^{th} terminal.

When the satellite receiver noise dominates, that is, when $P_T << N_sW$, the link is said to be *uplink limited*, and most of the downlink P_{sat} is wastefully allocated to noise power. When this is the case, and when $\gamma_iP_{sat} >> N_gW$, we can rewrite (21) as

$$\left(\frac{P_r}{N_0}\right)_{1i} \cong \frac{\gamma_i\,P_{sat}\,A_1P_1/N_sW}{(\gamma_i\,P_{sat}/W) + N_g} \cong \frac{A_1P_1}{N_s}. \tag{22}$$

Equation (22) illustrates that, in the case of an uplink limited channel, the P_r/N_0 downlink ratio essentially follows the uplink SNR. The more common situation is the *downlink limited* channel, in which case $P_T >> N_sW$, and the satellite EIRP is limited. In this case, (21) can be rewritten as [46]

$$\left(\frac{P_r}{N_0}\right)_{1i} \cong \frac{\gamma_i\,P_{sat}A_1P_1/P_T}{N_g}.$$

The power of the transponder is then shared primarily among the various uplink transmitted signals; very little uplink noise is transmitted on the downlink. The repeater in this case is constrained only by its downlink parameters.

Conclusion

The purpose of this two-part paper was to generate a structure and hierarchy of key signal processing transformations. This structure, which is delineated in Figs. 1 and 2, was

the basis for an overview of digital communications systems with an emphasis on satellite links. The paper enumerated the details behind the processing steps contained in the structure. Also, properties of detection theory, principal trade-off parameters, and link analysis have been treated.

References

[1] K. Feher, *Digital Communications: Microwave Applications*, Englewood Cliffs, NJ: Prentice-Hall, 1981.

[2] P. R. Hartmann, "Digital radio technology: present and future," *IEEE Communications Magazine*, vol. 19, no. 4, pp. 10-14, July 1981.

[3] S. Golomb, Ed., *Digital Communications with Space Applications*, Englewood Cliffs, NJ: Prentice-Hall, 1964.

[4] V. K. Bhargava, "Forward error correction schemes for digital communications," *IEEE Communications Magazine*, vol. 21, no. 1, pp. 11-19, January 1983.

[5] A. J. Viterbi, "Convolutional Codes and Their Performance in Communication Systems," *IEEE Trans. Commun. Technol.*, COM-19, pp. 751-772, October 1971.

[6] J. A. Heller and I. M. Jacobs, "Viterbi Decoding for Satellite and Space Communication," *IEEE Trans. Commun. Technol.*, October 1971.

[7] G. Forney, Jr., "Coding and its application in space communications," *IEEE Spectrum*, pp. 47-58, June 1970.

[8] J. L. Ramsey, "Realization of optimum interleavers," *IEEE Trans. Inform. Theory*, IT-16, no. 3, pp. 338-345, May 1970.

[9] I. L. Lebow, K.L. Jordan, Jr., and P. R. Drouilhet, Jr., "Satellite communications to mobile platforms," *Proc. IEEE*, vol. 59, no. 2, pp. 139-159, February 1971.

[10] T. Muratani, Satellite-switched time-domain multiple access," *Record IEEE Electron. and Aerosp. Syst. Conv., (EASCON)*, pp. 189-196, October 7-9, 1974.

[11] J. G. Puente and A. M. Werth, "Demand-assigned service for the Intelsat global network," *IEEE Spectrum*, pp. 59-69, January 1971.

[12] V. K. Bhargava, D. Haccoun, R. Matyas, and P. P. Nuspl, *Digital Communications by Satellite*, ch. 10, New York: John Wiley and Sons, 1981.

[13] J. W. Schwartz, J. M. Aein, and J. Kaiser, "Modulation techniques for multiple access to a hard-limiting satellite repeater," *Proc. IEEE*, May 1966.

[14] L. M. Nirenberg and I. Rubin, "Multiple access system engineering—a tutorial," *IEEE WESCON/78 Professional Program*, Modern Communication Techniques and Applications, session 21, Los Angeles, CA, September 13, 1978.

[15] G. D. Dill, "TDMA, the state of the art," *Record IEEE Electron, and Aerosp. Syst. Conv. (EASCON)*, pp. 31-5A-31-5I, Sept. 26-28, 1977.

[16] R. C. Dixon, *Spread Spectrum Analysis*, New York: John Wiley and Sons, 1976.

[17] M. P. Ristenbatt and J. L. Daws, Jr., "Performance criteria for spread spectrum communications," *IEEE Trans. Commun.*, COM-25, pp. 756-761, August 1977.

[18] S. W. Houston, "Modulation techniques for communication, part 1: tone and noise jamming performance of spread spectrum *M*-ary FSK and 2, 4-ary DPSK waveforms," *IEEE 1975 Nat'l. Aerosp. and Electron. Conf.*, pp. 51-58, June 10-12, 1975.

[19] G. K. Huth, "Spread spectrum techniques," *IEEE WESCON/78 Professional Program*, Modern Communication Techniques and Applications, session 15, Los Angeles, CA, September 13, 1978.

[20] J. K. Holmes, *Coherent Spread Spectrum Systems*, New York: John Wiley and Sons, 1982.

[21] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, pp. 656-715, October 1949.

[22] National Bureau of Standards, "Data Encryption Standard," *Federal Information Processing Standard (FIPS)*, publication no. 46, January 1977.

[23] W. Diffie and M. E. Hellman, "Privacy and authentication: an introduction to cryptography," *Proc. IEEE*, vol. 67, no. 3, pp. 397-427, March 1979.

[24] D. Kahn, *The Codebreakers*, New York: MacMillan Co., 1967.

[25] Special Issue on Communications Privacy, *IEEE Communications Magazine*, vol. 16, no. 6, November 1978.

[26] L. E. Franks, "Carrier and bit synchronization in data communication—a tutorial review," *IEEE Trans. Commun.*, COM-28, no. 8, pp. 1107-1121, August 1980.

[27] W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, Englewood Cliffs, NJ: Prentice-Hall, 1973.

[28] W. C. Lindsey, *Synchronization Systems in Communications and Control*, Englewood Cliffs, NJ: Prentice-Hall, 1972.

[29] P. P. Nuspl, K. E. Brown, W. Seenaart, and B. Ghicopoulos, "Synchronization methods for TDMA," *Proc. IEEE*, vol. 65, pp. 434-444, 1977.

[30] R. K. Crane, "Prediction of attenuation by rain," *IEEE Trans. Commun.*, COM-28, no. 9, pp. 1717-1733, September 1980.

[31] L. M. Schwab, "World-wide link availability for stationary and critically inclined orbits including rain attenuation effects," *Lincoln Laboratory Project Report DCA-9*, January 27, 1981.

[32] P. L. Bargellini, "Principles and evolution of satellite communications," *Acta Astronautica*, Pergamon Press, vol. 5, pp. 135-149, 1978.

[33] R. Gagliardi, *Introduction to Communication Engineering*, New York: John Wiley and Sons, 1978.

[34] H. Nyquist, "Thermal agitation of electric charge in conductors," *Phys. Rev.*, vol. 32, pp. 110-113, July 1928.

[35] R. E. Collin and F. J. Zucker, *Antenna Theory, Part I*, ch. 4, New York: McGraw-Hill, 1969.

[36] P. F. Panter, *Communications Systems Design: Line-of-Sight and Tropo Scatter Systems*, New York: McGraw-Hill, 1972.

[37] D. C. Hogg and T.-S. Chu, "The role of rain in satellite communications," *Proc. IEEE*, September 1975.

[38] D. C. Hogg and W. W. Mumford, "The effective noise temperature of the sky," *Microwave J.*, pp. 80-84, March 1960.

[39] B. Sklar, "What the system link budget tells the system engineer," *Proc. Int. Telemetering Conf.*, vol. 15, 1979.

[40] S. J. Campanella, F. Assal, and A. Berman, "Onboard regenerative repeaters," *Int'l. Conf. on Commun.*, Chicago, IL., vol. 1, pp. 6.2-121-6.2-125, 1977.

[41] K. Koga, T. Muratani, and A. Ogawa, "On-board regenerative repeaters applied to digital satellite communications," *Proc. IEEE*, March 1977.

[42] J. J. Jones, "Hard limiting of two signals in random noise," *IEEE Trans. Inform. Theory*, pp. 34-42, January 1963.

[43] F. E. Bond and H. F. Meyer, "Intermodulation effects in limiter amplifier repeaters," *IEEE Trans. Commun. Technol.*, COM-18, no. 2, pp. 127-135, April 1970.

[44] O. Shimbo, "Effects of intermodulation, AM-PM conversion, and additive noise in multicarrier TWT systems," *Proc. IEEE*, vol. 59, pp. 230-238, February 1971.

[45] P. Jain, T. C. Huang, K. T. Woo, J. K. Omura, and W. C. Lindsey, "Detection of MPSK signals transmitted through a nonlinear satellite repeater," *NTC '77 Conf. Rec.*, December 5-7, 1977.

[46] J. J. Spilker, Jr., *Digital Communications by Satellite*, Englewood Cliffs, NJ: Prentice-Hall, 1977.

Bernard Sklar was born in New York, NY on September 11, 1927. He received the B.S. degree in mathematics and science from the University of Michigan, Ann Arbor, in 1949; the M.S.E.E. degree from the Polytechnic Institute of New York, Brooklyn, in 1958; and the Ph.D. degree in engineering from the University of California, Los Angeles, in 1971.

He has 30 years of experience with the aerospace/defense industry in a variety of technical design and management positions. From 1953 to 1958 he was a research engineer with Republic Aviation Corp., Farmingdale, NY; from 1958 to 1959 he was a member of the technical staff at Hughes Aircraft Co., Culver City, CA; from 1959 to 1968 he was a senior staff engineer at Litton Systems, Inc., Canoga Park, CA. In 1968 he joined The Aerospace Corp., El Segundo, CA, where he is currently employed. He is a project engineer in the MILSTAR program office, involved in the development of the joint services MILSTAR communications satellite system.

He has been involved in teaching engineering courses during the past 25 years at the University of California, Los Angeles and Irvine; the University of Southern California, Los Angeles; and West Coast University, Los Angeles. He is a past-chairman of the Los Angeles Council IEEE Education Committee, and is a Senior Member of the IEEE. ■