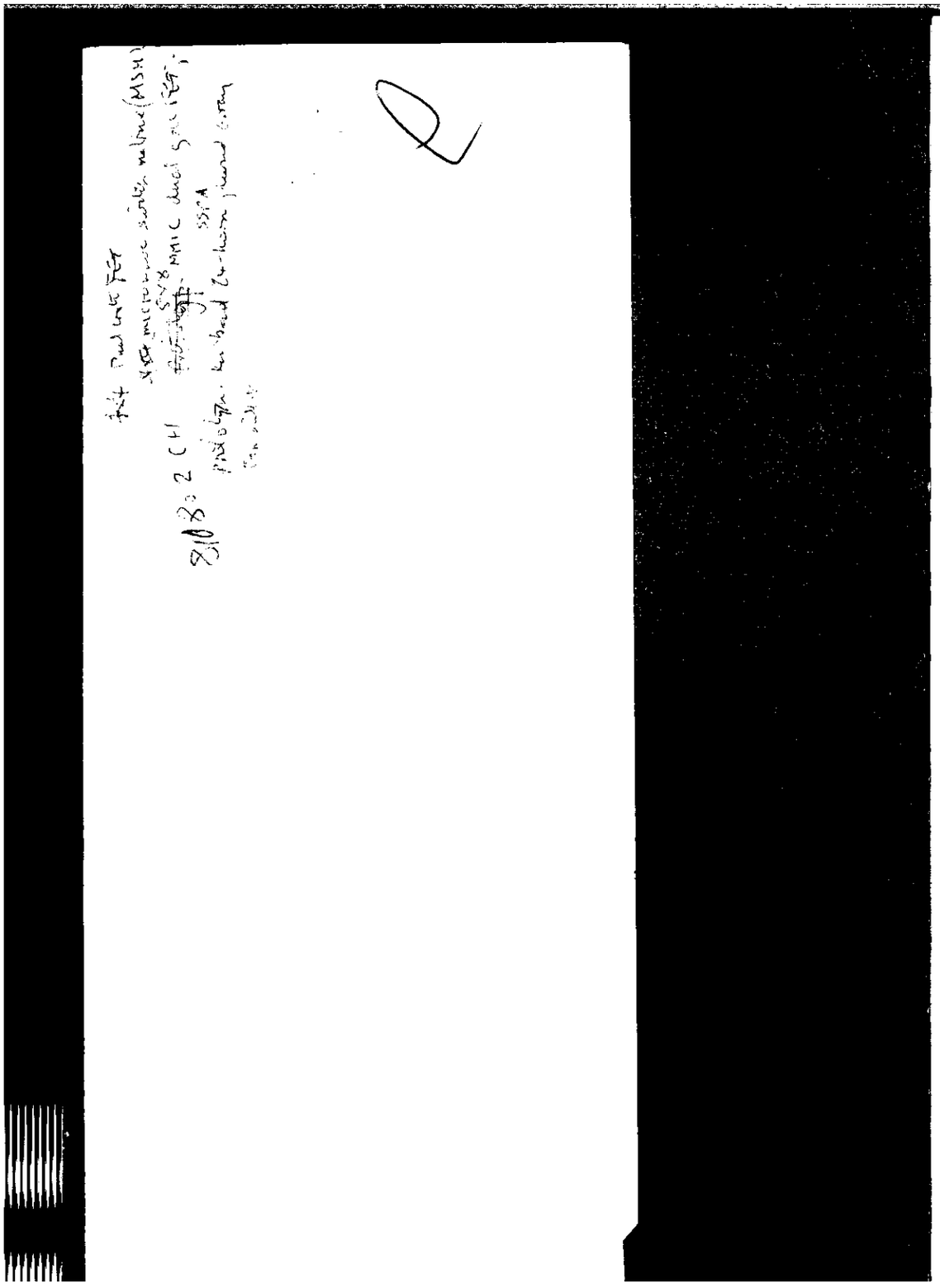


COMSAT Technical Review

Volume 2 Number 1, Spring 1972

Volume 2 Number 1, Spring 1972

The COMSAT TECHNICAL REVIEW is published twice a year by Communications Satellite Corporation, 950 L'Enfant Plaza, S.W., Washington, D. C. 20024. Editors: Lawrence Weekley, Jr., Gerald H. Bidlack, Leonard F. Smith.

Contents

Index: echoes, suppressors, telephone circuits, voice communications

Analysis of an adaptive impulse response echo canceller

S. J. CAMPANELLA, H. G. SUYDERHOUD, AND M. ONUFRY

ABSTRACT

Echo in the long-distance telephone plant is a well known, undesirable phenomenon. The conventional technique of preventing echo uses voice switched devices called echo suppressors. They are inherently limited in performance because of a design dichotomy which is exhibited when both parties talk simultaneously.

Recent techniques have made it possible to explore a radically different design approach, eliminating the dichotomy and particularly enhancing the quality of satellite communications. This paper describes the design and performance of a device called an echo canceller. The results shown are sufficient to direct further efforts toward economic feasibility.

INTRODUCTION

The problems of echo and propagation delay and the use of conventional echo suppressors in telephone communications have been extensively treated in the literature [1]-[9]. In today's telephone technology, echo is controlled almost exclusively by the use of switching-type echo suppressors. Although these echo suppressors have proved adequate on long propagation delay circuits (i.e., up to 600 ms for one-way propagation

time), they do exhibit inherent shortcomings. If these shortcomings were eliminated, even better telephone service could be provided. Consequently, alternatives are being sought. One promising alternative is the echo canceller. This paper presents an analysis of and results obtained from an experimental model of an echo canceller.

In 1966, Miura, Sato, and Nagata [10] of Nippon Electric Company described work on a digitally implemented, impulse response echo canceller, referred to as the "blockless echo suppressor." In this device, the impulse response of the telephone circuit terminating the long-distance propagation link (the echo path) was probed and stored in a digital memory. The incoming signal was fed to the terminating telephone circuit and to a convolution processor which used the stored impulse response to produce a replica of any echo that was returned from the echo path. The echo was cancelled by subtracting the modeled echo from the real echo. This method appeared to work quite well, but was limited to time-invariant echo paths. Any change in the echo path impulse response which occurred after the first probing and storing would render that stored model inaccurate. Only by again probing the terminating telephone path was it possible to re-establish an accurate impulse response model.

In 1967 Sondhi [11] reported on work performed at Bell Laboratories on an adaptive model of an echo canceller. In this device, the impulse response was modeled by an analog tapped delay line in which the coefficients of the impulse response could be automatically and continuously adjusted during the course of conversation to minimize the rms value of the echo signal appearing on the return path.

Under INTELSAT sponsorship [12] and with the technical support of COMSAT Laboratories, Nippon Electric Company developed an adaptive impulse response echo canceller using as a point of departure the concept embodied in the blockless echo suppressor, but incorporating adaptation principles similar to those suggested by Sondhi. This paper analyzes the operation of the adaptive impulse response echo canceller and presents results of experimental work performed by COMSAT to improve further the performance of the echo canceller. A novel adaptation algorithm has been developed which permits fast convergence on both speech and random noise, while maintaining good stability and adequate echo return loss enhancement. This fast convergence algorithm makes it possible to accommodate rapid changes in echo path impulse response without serious degradation of canceller performance. Experimental results, obtained on linear and nonlinear echo paths with and without phase roll, are presented.

The theoretical analysis indicates an ultimate overall echo return loss of 30 dB and a time to convergence of 512 ms for the design under investigation. Measurements made on the experimental model show an ultimate overall echo return loss of 24 dB for speech and a time to convergence of 850 ms for random noise. Incorporating a compandor within the echo canceller resulted in an ultimate overall echo return loss of 38 dB for speech and a time to convergence of 150 ms for random noise.

DESCRIPTION OF A DIGITAL ECHO CANCELLER

Figure 1 is a block diagram of a digital echo canceller. This device operates in the four-wire circuit in place of a split echo suppressor; thus, two such devices would be needed on a complete long-distance circuit, one at each end.

The method of processing is based on digital convolution over a finite window. To this end, four basic units are employed:

a. a storage register, X, of size $N = 260$ stages (32.5-ms storage capacity) for the receive-side signal samples, $x[(j - i)\ T]$;

b. a storage register, H, of size $N + 1 = 261$ stages for the model echo path impulse response, $h(iT)$;

c. a convolution processor; and

d. an adaptive control loop for obtaining and updating the model echo path impulse response.

The operation is as follows. N samples, $x[(j - i)\ T]$, of the incoming signal on the receive side are stored in the X register and recirculated every 125 μs. For every new sample, the contents shift one position, and the oldest sample is replaced by the newest sample.

During each recirculating cycle, each sample in the X register is multiplied by a corresponding sample recirculating at the same rate in the H register, and the products are summed. This process of multiply-and-add takes place in the convolution processor. The resulting sample, $\hat{y}$, is the most recent estimate of the real echo signal, y, from which it is subtracted to perform the echo cancellation.

Also, during each recirculating cycle the correction sensor in the adaptive control loop accepts the error signal, $\epsilon = y - \hat{y}$, and the samples, $x[(j - i)\ T]$, stored in the X register to produce a positive, negative, or zero correction to the values of the impulse response stored in the H register. The correction called for by the correction sensor is executed by the update unit. The function produced by the correction sensor may be represented by the product $\phi_x(x)\ \phi_\epsilon(\epsilon)$. The functions $\phi_x(x)$ and $\phi_\epsilon(\epsilon)$ are

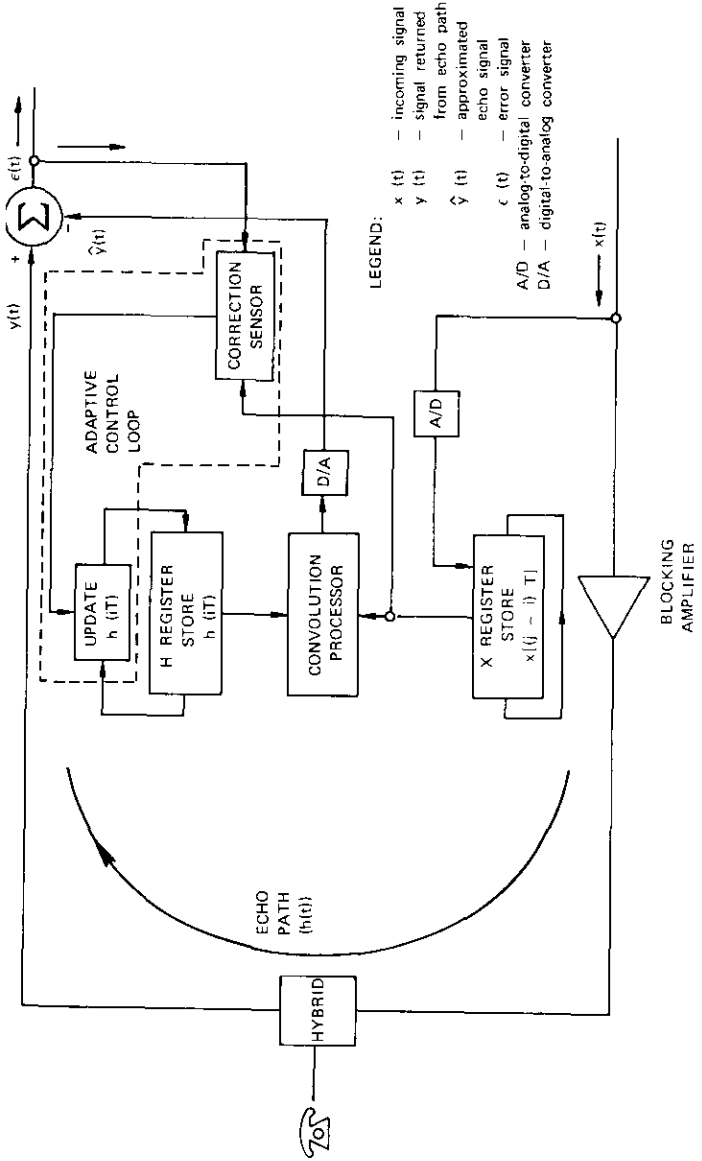


Figure 1. Block Diagram of the Digital Echo Canceller

asymmetrical, nondecreasing functions which will be discussed in greater detail in the following section. The expression $\phi_x(x)\phi_\epsilon(\epsilon)$ actually represents N products. Each product results in a correction to one of the N values of the impulse response, $h(iT)$, stored in the H register.

DERIVATION OF CONVERGENCE ALGORITHM

In this section, the derivation of the correction method used for canceller convergence is discussed. Two developments are presented; both logically lead to the algorithm used. The first is based on use of the pseudo-inverse in solving the vector equation of the convolution process. The second is based on the method of steepest descent.

The actual echo path memory function in discrete form is represented by the row vector

$$w = [w_1, w_2, \ldots, w_N] \tag{1}$$

where N is the number of sample points. If the samples are taken every T seconds, then the total duration of the sample window is NT. The simulated echo path memory function at the jth iteration is represented by the row vector

$$h^j = [h_1{}^j, h_2{}^j, \ldots, h_N{}^j]. \tag{2}$$

The input sequence on the receive side is defined in general as

$$x_1, x_2, x_3, \ldots \tag{3}$$

The input sequence involved in the jth iteration in attempts to estimate h^j is the row vector

$$x_j = [x_{j-1}, x_{j-2}, \ldots, x_{j-N}]. \tag{4}$$

Note that the elements of the preceding vector are in time reverse order to accommodate the representation of the convolution process by a scalar product.

At the time of the jth iteration, the value of the signal appearing at the output of the echo path is

$$y_j = x_j w^T \tag{5}$$

where w^T is the transpose of w. At the same time, the value of the signal appearing at the output of the simulated echo path is

$$z_j = x_j[h^j]^T \tag{6}$$

The error between z_j and y_j at the jth iteration is

$$\begin{aligned}\epsilon_j &= y_j - z_j \\ &= x_j[w - h^j]^T. \end{aligned} \tag{7}$$

A single error is thus expressed as one linear equation with N unknowns, which are the N values of the real impulse response. The information contained in this equation must be utilized to update the values $h_i{}^j$ in order to minimize ϵ_j. This is accomplished by using the pseudo-inverse which will be discussed in the following paragraph.

Assume that the updated values of $h_i{}^j$ are

$$h_i{}^{j+1} = h_i{}^j + \Delta h_i{}^j. \tag{8}$$

We wish to select values of $\Delta h_i{}^j$ which minimize

$$||y_j - x_j[h^{j+1}]^T||. \tag{9}$$

This expression may be expanded as follows:

$$\begin{aligned}||y_j - x_j[h^{j+1}]^T|| &= ||y_j - x_j[h^j]^T - x_j[\Delta h^j]^T|| \\ &= ||\epsilon_j - x_j[\Delta h^j]^T||. \end{aligned} \tag{10}$$

The pseudo-inverse solution to satisfy equation (10) is, as shown by Greville [13],

$$\Delta h^j = (x_j x_j{}^T)^{-1} x_j \epsilon_j. \tag{11}$$

The expression for $(x_j x_j{}^T)^{-1} \cdot x_j$ is

$$(x_j x_j{}^T)^{-1} x_j = \frac{x_j}{\sum_i x_{j-i}^2}. \tag{12}$$

Thus, the final solution for updating the $h_i{}^j$ coefficients is

$$\Delta h^j = \frac{x_j \epsilon_j}{\sum_i x_{j-i}^2}. \tag{13}$$

This procedure for updating the contents of the model memory function

may be considered to be a special case of convergence by the method of steepest descent.

In order to demonstrate this, we derive the steepest descent method as follows. Consider a positive even function, ϕ, of the error signal, ϵ, such that

$$\frac{d\phi(\epsilon)}{d\epsilon} = \phi_\epsilon(\epsilon) \tag{14}$$

is a monotonically nondecreasing odd function, and $d\phi_\epsilon(\epsilon)/d\epsilon \geq 0$. These properties are required to ensure convergence of the process. For steepest descent, the incrementing vector, Δh^j, must be proportional to the negative gradient of the function ϕ with respect to the coefficients $h_i{}^j$. Thus,

$$\begin{aligned}\Delta h^j &= -C\, grad_h\, \phi(\epsilon_j) \\ &= -C\, grad_h\, \phi\{x_j w^T - x_j[h^j]^T\} \\ &= C\, x_j \phi_\epsilon\{x_j w^T - x_j[h^j]^T\} \\ &= C\, x_j\, \phi_\epsilon(\epsilon_j). \end{aligned} \tag{15}$$

By letting $C = (x_j x_j{}^T)^{-1}$ and $\phi_\epsilon(\epsilon_j) = \epsilon_j$, we obtain equation (11). The updating equation can alternately be written as

$$h^{j+1} = h^j + C x_j \phi_\epsilon(\epsilon_j) \tag{16}$$

where $C > 0$, and ϕ_ϵ fulfills equation (14).

Equation (13) or (15) is approximated in the following form for the convergence algorithm used in the experimental echo canceller under discussion:

$$\Delta h^{j+1} = \eta \phi_\epsilon(\epsilon_j)\, \phi_x(x_j) \tag{17}$$

where

$$\phi_\epsilon(\epsilon_j) = \begin{cases} -1, & \epsilon_j < -\Delta\epsilon \\ 0, & -\Delta\epsilon < \epsilon_j < \Delta\epsilon \\ 1, & \epsilon_j > \Delta\epsilon \end{cases}$$

$$\phi_x(x_j) = \begin{cases} -1, & x_j < -\Delta x \\ 0, & -\Delta x < x_j < \Delta x \\ 1, & x_j > \Delta x \end{cases}$$

and η = size of the correction increment. ϕ_ϵ and ϕ_x are step functions such as those shown in Figure 2. Each possesses an inert zone and if x_j and/or ϵ_j falls in this zone, no correction is applied. Otherwise, corrections are made in accordance with the signs of the values of x_j and ϵ_j. The principal reason for selecting these functions is that they can be realized by simple digital implementations. In the following, selection of the values to be used for $\Delta\epsilon$, Δx, and η is discussed in detail. It will be seen that optimum performance can be achieved only by varying these values as a function of signal and error levels.

RATE OF CONVERGENCE

The rate of convergence of the echo canceller is largely determined by the size of the correction, η, which is made to the values in the H register. The rate of convergence is determined here by assuming the following conditions at the start of convergence:

a. the X register is fully loaded with the N most recent samples of the received input signal,

b. the H register is cleared so that all N values are zero,

c. the true echo path is characterized by a unit pulse which occupies the full range of the H register (which is a most pessimistic assumption),

d. the H register is linearly quantized using Q bits, and

e. corrections are made on μ bits, where $2^\mu = \eta$. Based on these conditions, it will take $2^{Q-\mu}$ recirculations of the values in both registers to attain the assumed impulse response. Thus, the time for convergence will be

$$T_c = T2^{Q-\mu}. \tag{18}$$

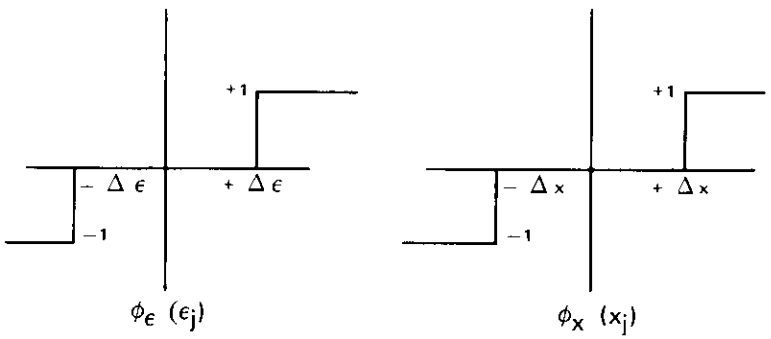


Figure 2. Plots of Functions $\phi_\epsilon(\epsilon)$ and $\phi_x(x)$ (control functions for the updating algorithm)

This estimate is based on the assumption that no errors are made in the decisions for correcting the H register. This condition will prevail when the value of $\hat{y}$ is significantly different from that of y. However, when $y - \hat{y}$ approaches the vicinity of $\Delta\epsilon$, some erroneous corrections will occur which will slow up the rate of convergence. As a consequence, the time of convergence actually observed for the experimental echo canceller will be greater than that given by the previous relationship for T_c. It was pointed out that this is a most pessimistic estimate, since it assumes that the values in the H register must reach the peak quantizing level. In actual operation, the values obtained in the H register will not always have to reach this peak, and the rate of convergence will be correspondingly smaller.

Based on the preceding formula for T_c, $Q = 12$, and $\mu = 0$, the convergence time is 512 ms. It is noted that the convergence time can be decreased by allowing μ to take on values of 1, 2, 3, . . . ; however, this will result in large corrections in $\hat{y}$ and consequent degradation in the ultimate echo cancellation capability at convergence. Thus, short convergence time (high rate of convergence) and good echo cancellation capability at convergence cannot be attained by using a constant value of μ and, thus, η. Accordingly, in the present investigation of the experimental echo canceller, the value of μ is made a function of the magnitude of the error between y and $\hat{y}$. When the error is large, the value of μ is made large; i.e., $\mu = 2$. When the magnitude of the error $y - \hat{y}$ is small, μ is set equal to 0. In this way, it has been possible to obtain a high rate of convergence and a good echo cancellation capability at convergence.

CANCELLATION AT CONVERGENCE

There are two competing mechanisms which have an effect on the magnitude of error when the system has attained the converged state. The first of these is the error caused by the width of the region within which no correction is sensed, i.e., the region $-\Delta\epsilon < \epsilon < \Delta\epsilon$. This region will be referred to as an inert zone. Since the error within this region is equally likely to assume any value, the error distribution is uniform with zero mean. Hence, in this case the pdf for ϵ is

$$P_1(\epsilon) = \frac{1}{2\Delta\epsilon}, \quad -\Delta\epsilon < \epsilon < \Delta\epsilon$$
$$= 0 \text{ otherwise.} \tag{19}$$

The error variance is thus the expected value of ϵ^2. That is,

$$\overline{\epsilon^2} = \int_{-\infty}^{\infty} \epsilon^2 P_1(\epsilon)\, d\epsilon = \int_{-\Delta\epsilon}^{\Delta\epsilon} \frac{\epsilon^2}{2\Delta\epsilon}\, de = \frac{\Delta\epsilon^2}{3}. \tag{20}$$

The second error mechanism is attributed to the process of impulse response correction. More specifically, there is a random component in the echo signal produced by the modeled impulse response. This is caused by the effects of quantization (for digital processing) and truncation of the impulse response. In addition, there is circuit noise in the real echo path which, for the sake of analysis, may be assumed to occur in the modeled echo path. The net result of these errors is a random variation in the estimate of the actual send path signal which is superimposed on the error signal, ϵ. The distribution of this variation is designated as $P_m(\epsilon)$ in Figure 3 where it is shown in relationship to the inert zone.

If the error exceeds the upper bound, a correction of magnitude $-\Delta c$ will result from a change in the impulse response sample. If the error is less than the lower bound, a correction of magnitude $+\Delta c$ will result. If the error lies inside the inert zone, no correction will take place. The probability of each of these types of correction is given by the area under the distribution $P_m(\epsilon)$. Thus, the distribution of the corrections is

$$P_1 = P(c = \Delta c) = \int_{-\infty}^{-\Delta\epsilon} P_m(\epsilon)\, d\epsilon$$
$$P_2 = P(c = 0) = \int_{-\Delta\epsilon}^{\Delta\epsilon} P_m(\epsilon)\, d\epsilon$$
$$P_3 = P(c = -\Delta c) = \int_{\Delta\epsilon}^{\infty} P_m(\epsilon)\, d\epsilon \tag{21}$$

where the error distribution $P_m(\epsilon)$ is assumed to be Gaussian, i.e.,

$$P_m(\epsilon) = \frac{1}{\sqrt{2\pi}\,\sigma_m^2} \cdot \epsilon^{-(1/2)(\epsilon^2/\sigma_m^2)}.$$

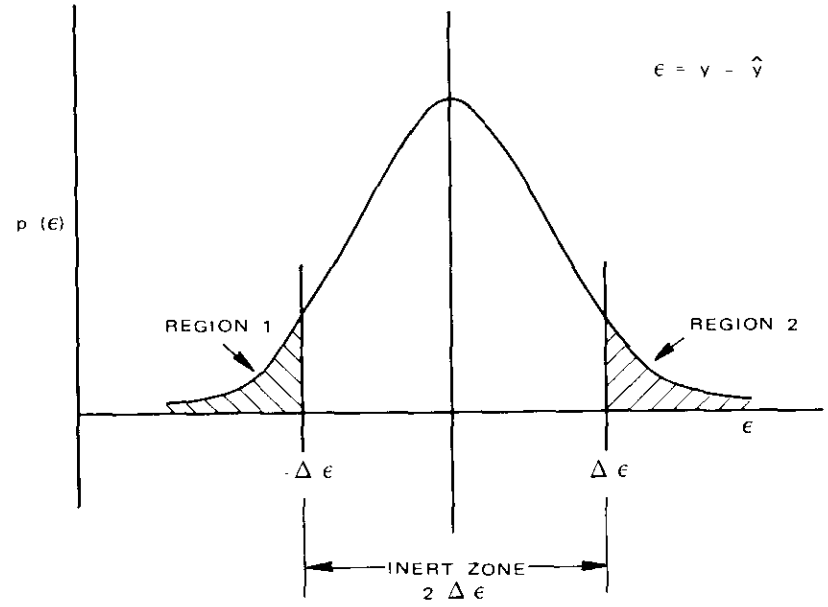


Figure 3. Distribution of Errors

The process of correction also contributes noise; the expected value of this noise is

$$\overline{c^2} = \sum c^2 P_i = \Delta c^2 \left[1 - 2\int_0^{\Delta\epsilon} P_m(\epsilon)\, d\epsilon\right] \tag{22}$$

The error, ϵ^2, due to the width of the inert zone, increases with increasing $\Delta\epsilon$, while the error, $\overline{c^2}$, due to the corrections, increases with decreasing $\Delta\epsilon$. Thus, if all other factors are constant, there will be a value of $\Delta\epsilon$ for which the error is minimum. If it is assumed that the two mechanisms of error are independent, then the total error is given by

$$\overline{\epsilon_T^2} = \overline{c^2} + \overline{\epsilon^2} = \overline{c^2} + \sigma_m^2. \tag{23}$$

To calculate $\overline{c^2}$ in the preceding expression, estimates are needed for σ_m^2 and Δc^2. As was previously pointed out, quantizing error in the

convolution processing, truncation of the impulse response, and uncorrelated noise appearing on the send side of the echo canceller contribute to σ_m^2. For the analysis conducted here, it has been assumed that the principal contributor to σ_m^2 is quantizing error. This is a reasonable assumption, since, during normal operation, the echo canceller will experience a relatively low noise level on the send side (−50 dBm0), and during double talk, the feedback correction process will be disabled. Truncation of the impulse response is unlikely to cause a significant contribution to σ_m^2 if sufficient storage time is allowed in the registers.

The quantizing error resulting from the convolution processing is estimated as follows. Each value of $\hat{y}$ produced by the impulse response model is the consequence of the aggregate of N multiplications of values of x and h. The variance σ_i^2 for any single multiplication contribution is approximated by

$$\sigma_i^2 = \overline{X}^2\sigma_H^2 + \overline{H}^2\sigma_X^2 . \tag{24}$$

In this expression, $\overline{X^2}$, which is the expected value of the squares of the values stored in the X register, is given by σ^2. $\overline{H^2}$ is the expected value of the squares of the values stored in the H register; if these values are uniformly distributed over the entire H range from -1 to 1, H^2 is equal to 1/3. σ_H and σ_X are determined from the quantizing errors in the H and X register, respectively, and are given by

$$\sigma_H^2 = \left(\frac{1}{2^Q\sqrt{12}}\right)^2; \qquad \sigma_X^2 = \left(\frac{\gamma\sigma}{2^L\sqrt{12}}\right)^2. \tag{25}$$

In equation (25), Q is the number of bits per sample admitted from the H register, L is the number of bits per sample admitted from the X register, and γ is the peak-to-rms ratio allowed in the X register. Since there are N contributions for each estimated value of $\hat{y}$, the total variance in $\hat{y}$ is $N\sigma_i^2$. Hence,

$$\sigma_m^2 = N\sigma_i^2 = \frac{\sigma^2}{2^{2Q}}\left[\frac{N}{12}\left(1 + \frac{\gamma^2}{3}\right)\right] \tag{26}$$

where it is assumed that $L = Q$.

It is also necessary to estimate Δc; the following procedure is used to obtain this estimate. Assume that the x signal is a Gaussian random noise signal with zero mean and variance σ^2. Thus, the probability distribution function for x is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x^2/2\sigma^2)} . \tag{27}$$

The Δx threshold is determined from the average value of the magnitudes of the contents of the X register. If N, the number of stages of the register, is a large sample, Δx may be computed as follows:

$$\Delta x = \int_{-\infty}^{\infty} |x|\, f(x)\, dx$$

$$= 2\int_0^{\infty} \frac{x}{\sigma\sqrt{2\pi}}\, e^{-(x^2/2\sigma^2)}\, dx = \sigma\sqrt{\frac{2}{\pi}} . \tag{28}$$

N', the average number of x samples exceeding Δx, is then given by the fraction of values of x that exceed $|\Delta x|$; i.e.,

$$\alpha = 2\int_{\Delta x}^{\infty} f(x)\, dx = 0.424. \tag{29}$$

Thus, $N' = \alpha N = 110$ for $N = 260$.

The average value of the magnitudes of x exceeding Δx is

$$\bar{x}_{\Delta x} = \int_{\Delta x}^{\infty} xf(x)\, dx \tag{30}$$

where $$f(x) = \frac{k}{\sigma\sqrt{2\pi}}\, e^{-(x^2/2\sigma^2)}$$

and k is such that

$$\int_{\Delta x}^{\infty} f(x)\, dx = 1 .$$

The value of k is found to be 1/0.212; thus,

$$\bar{x}_{\Delta x} = \frac{k}{\sigma\sqrt{2\pi}}\int_{\Delta x}^{\infty} x\, e^{-(x^2/2\sigma^2)}\, dx = \frac{k\sigma}{\sqrt{2\pi}}\int_{\Delta x^2/2\sigma^2}^{\infty} e^{-u}\, du \tag{31}$$

or $$\bar{x}_{\Delta x} = \frac{k\sigma}{\sqrt{2\pi}}\, e^{-(1/\pi)} .$$

For any given correction cycle, the expected value of Δc is the aggregate of the products of X and ΔH for all values $|x| > |\Delta x|$. Thus,

$$\begin{aligned} \Delta c &= N' \bar{x}_{\Delta x} \Delta H \\ &= \alpha N \bar{x}_{\Delta x} \Delta H \\ &= \alpha N \bar{x}_{\Delta x} (2^Q - 1)^{-1} \end{aligned} \tag{32}$$

where Q is the number of bits used to quantize the magnitude of H. This assumes that the canceller has converged and that corrections take place only with the minimum correction increment.

The estimates given in equations (26) and (32) for σ_m and Δc may be used in equation (22) to obtain estimated values of $\bar{c}^2$. Graphs of $\bar{\epsilon}^2$, $\bar{c}^2$, and $\bar{\epsilon}_T^2$ for $\gamma = 8$, $N = 260$, and $Q = 10$ are shown in Figure 4. Note that the values are normalized relative to σ^2.

In Figure 4, it should be noted that the total error is minimized for $\Delta\epsilon/\sigma = 4.8 \times 10^{-2}$ and that, at this minimum, $\epsilon_T^2/\sigma^2 = 1.1 \times 10^{-3}$. Thus, it should be expected that after cancellation the echo canceller will produce an echo level which is 30 dB below the rms signal level appearing on the send side if there is no difference in the transmission levels (TL) on the send and receive sides. It appears that, for optimum performance, the values of $\Delta\epsilon$ should be proportional to the signal standard deviation, σ.

EXPERIMENTAL RESULTS

This section deals with the performance of the experimental echo canceller described earlier, which was designed principally for operation in a self-adaptive mode. Experimental results are presented showing the effects on canceller performance, using random noise and speech input signals, as various improvements were made to the basic experimental model. More specifically, results are shown for the following canceller models:

a. basic model with fixed $\Delta\epsilon$ and Δx thresholds and single minimum size correction, η;
b. improved model in which the correction steps, η, are proportional to the error, ϵ;
c. further improved model which includes a variable Δx proportional to the received signal level; and
d. model using syllabic compandors as part of the cancellation processing.

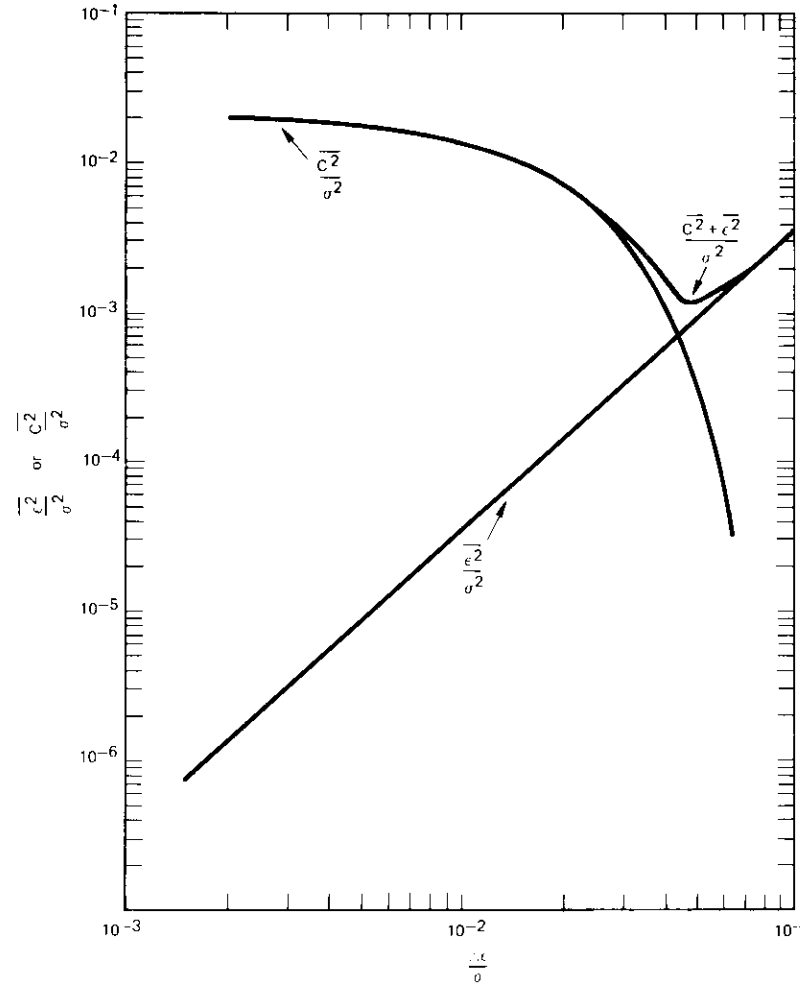


Figure 4. Error at Convergence as a Function of $\Delta\epsilon/\sigma$

It will be seen that, for random noise, the convergence time improves from about 850 ms for case a to about 150 ms in case d. The echo cancellation achieved is about 20 dB for cases a, b, c, and about 33 dB for case d.

Measurement Description

The two parameters of importance are speed of convergence and echo return loss enhancement (*ERLE*). The speed of convergence is defined as the time required to attain 90 percent of the ultimate *ERLE* in dB, when, at the onset of convergence, the H register is cleared (all zeros). *ERLE* is defined as

$$ERLE = -10 \log \frac{\overline{\epsilon^2}}{\overline{y^2}}$$

where ϵ is the error signal and y the uncancelled signal. All measurements of *ERLE* were made using band-limited noise as the test signal.

At this point it is appropriate to define two additional terms: overall echo return loss (*OERL*) and echo return loss (*ERL*). For equal *TL*s, *OERL* designates the overall loss from the receive-in side to the send-out side of the canceller, and *ERL* designates the loss of the actual echo path, i.e., from receive-out to send-in. The following obvious relationship exists among these various losses:

$$OERL = ERL + ERLE. \tag{34}$$

In the following discussion, both *OERL* and *ERLE* values will be given. *OERL* may also be expressed as follows:

$$OERL = -10 \log \frac{\overline{\epsilon^2}}{\sigma^2}. \tag{35}$$

A special tape recording of toll telephone quality was prepared to assess the canceller's performance for speech. The following short message, representing a typical opening segment of a telephone conversation, was recorded on tape: "Hello, this is Mr. Bird at COMSAT Laboratories. May I speak to Mr. Smith, please? This is in reference to purchase order number two-four-seven-three, placed last April." This tape was used to generate a second tape in which the original message was broken up into successively larger segments. The first segment started with the first word only; then one word was added in each consecutive segment. A calibrated test tone for level alignment was inserted at five-segment increments.

This second tape was used to provide the receive input signal to the canceller under test. Each segment was played back to allow the canceller to converge on the speech signal, the correction loop was then disabled, and the *ERLE* was measured. At the completion of each measurement, the H register was zeroed to guarantee the same initial conditions for each speech segment. Finally, curves showing convergence on speech in terms of *ERLE* versus number of spoken words were made from these data.

Photographs showing convergence on noise were made by applying the noise signal at the receive input and photographing the signal at the send output of the canceller, thus recording the decreasing error signal as $h(iT)$ approached the true impulse response of the echo path.

Case I—Basic Model

Case I was limited to a flat 6-dB echo path as shown in Figure 5. Figure 6 is a photograph showing convergence on random noise. The measured convergence time, T_c, was about 850 ms, which is longer than was predicted by equation (18). Three principal reasons for the longer convergence time are as follows:

a. Because of the thresholding of x, the error correction was not applied to every $h(iT)$ sample during each updating cycle.

b. Because an analog multiplier and integrator were used, their operation was easily perturbed by digital noise, which in turn produced errors in $\hat{y}$.

c. The predicted convergence time given by equation (18) assumed perfect convolution processing and ignored processing error. However, erroneous corrections occurred as the condition of convergence was approached.

An overall echo return loss of 27 dB was measured. This compares favorably with the value of 30 dB predicted from the model discussed in this paper. To obtain the *ERLE* it is necessary to reduce the overall echo return loss by the real echo path loss of 6 dB. Thus, the *ERLE* is 21 dB. Failure to attain the performance predicted by the model is attributed to the errors introduced in the analog circuit parts of the convolution processor.

Figure 7 shows the canceller's performance for the simple and complex echo paths shown in Figure 5 and Table I summarizes the performance.

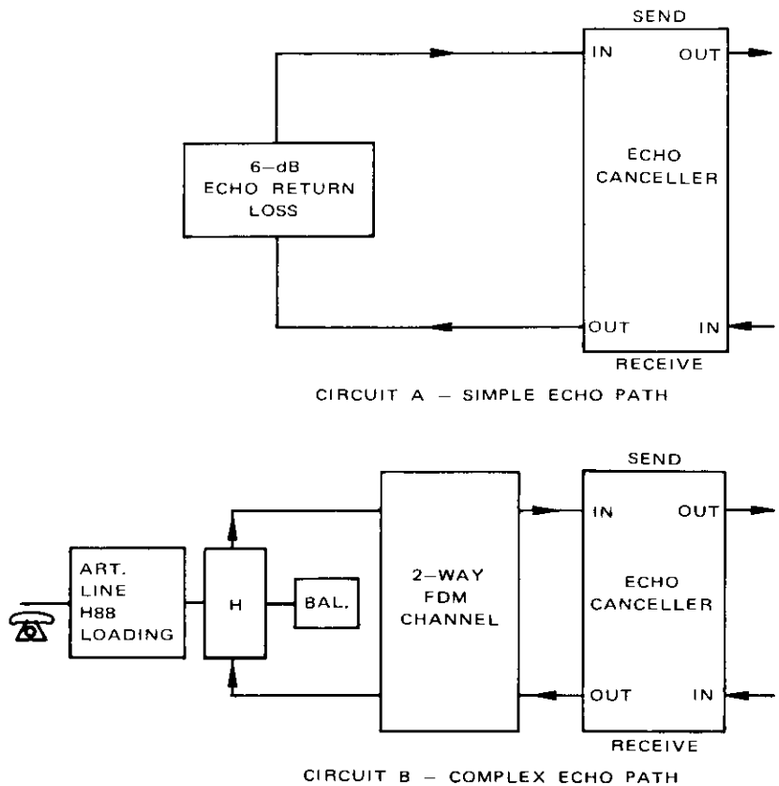


Figure 5. Two Echo Paths Used for Tests

TABLE 1. *ERLE* AND RATE OF CONVERGENCE ON SPEECH IN THE BASIC ECHO CANCELLER MODEL

Path Type	Final *ERLE* (dB)	Final *OERL* (dB)	Number of Words to Reach 90 Percent of Final dB Value
A (Simple)	18	24	15
B (Complex)	15	21	16

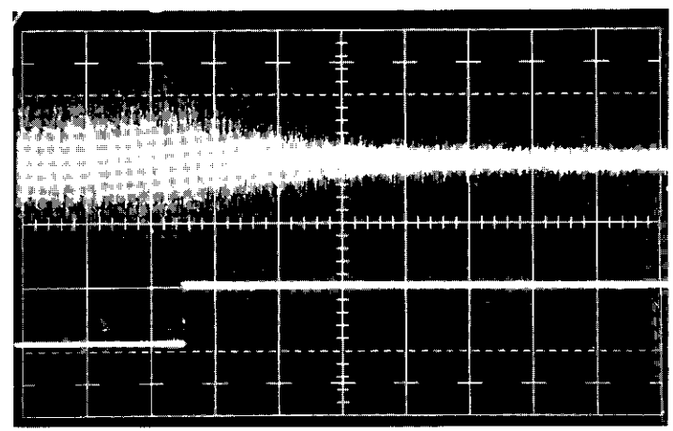


Figure 6. Photograph of Basic Model Convergence on Noise for a Simple Echo Path

Case II—Model with η Proportional to ϵ

As discussed earlier, the rate of convergence is expected to be enhanced when the $h(iT)$ correction step, η, is made proportional to the error, ϵ. To approximate this type of operation, a three-level error threshold detection function was implemented corresponding to three Δh correction increments with $\mu = 0$, 1, and 2. Table 2 shows the error ranges and the corresponding Δh correction increments.

Figure 8 shows convergence on random noise for case II. Observe that the convergence time is now 350 ms, which is better than a 2:1 reduction of the value obtained in case I. Figure 9 shows convergence on speech and Table 3 summarizes the performance for noise and speech excitation for both echo paths. Note again the improved performance when compared to the basic model.

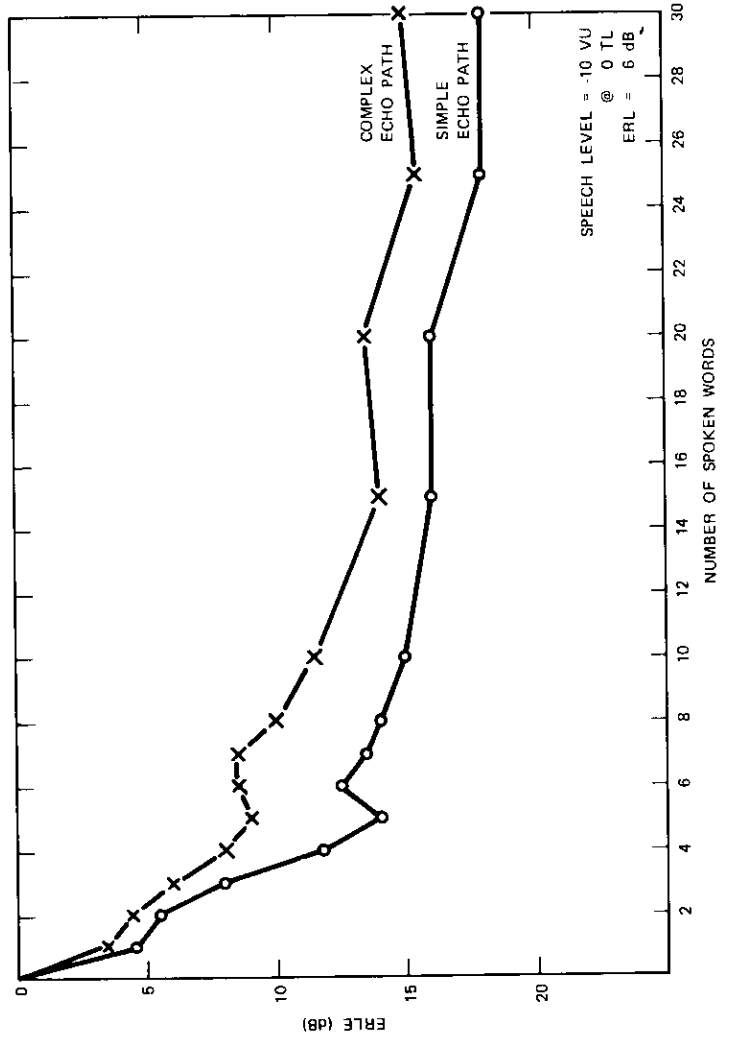
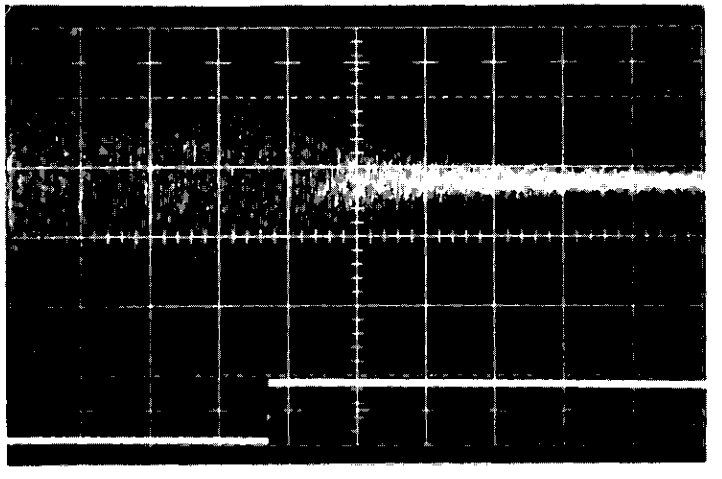
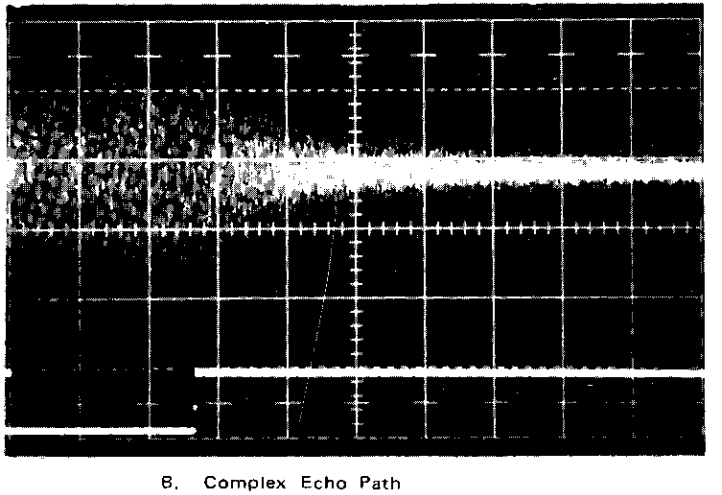


Figure 7. Basic Model Convergence on Speech for Simple and Complex Echo Paths



A. Simple Echo Path
ERLE = 19.5 dB
Horizontal Sweep = 100 ms/div



B. Complex Echo Path
ERLE = 20 dB
Horizontal Sweep = 100 ms/div

Figure 8. Photographs of the Canceller Performance on Noise with η Proportional to ϵ

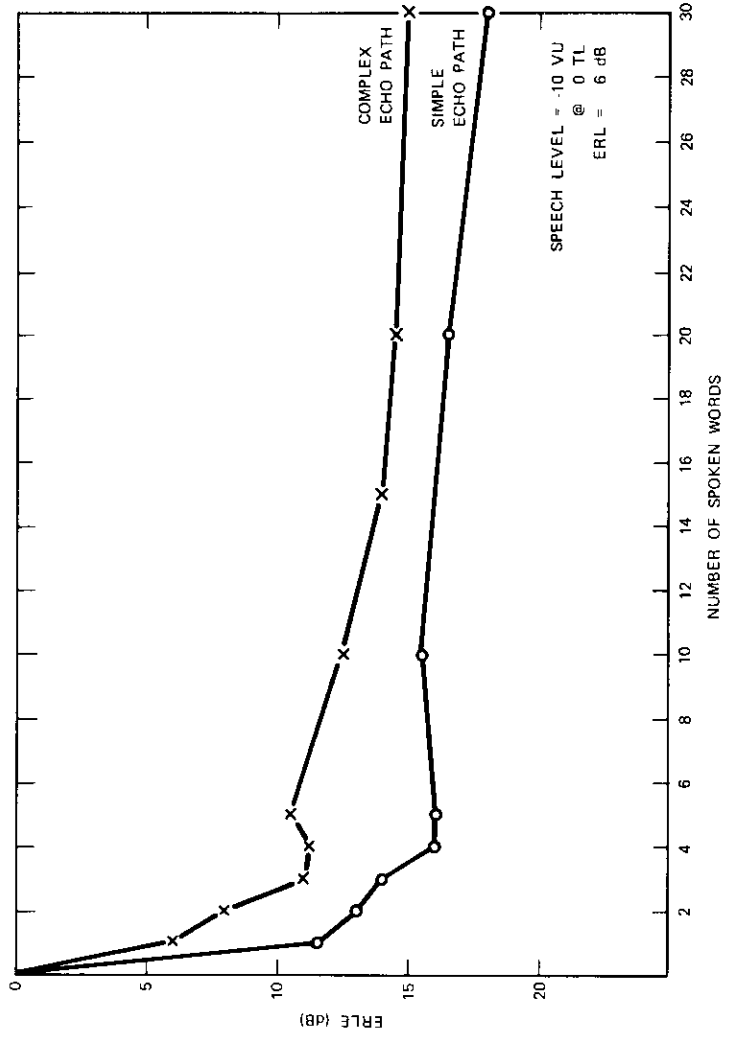


Figure 9. Canceller Convergence on Speech with η Proportional to ϵ for Simple and Complex Echo Paths

TABLE 2. η CORRECTIONS PROPORTIONAL TO ϵ

Range of ϵ	Correction Size, η ($\mu_1 > \mu_2 > \mu_3$)
$\epsilon < \Delta\epsilon$	0
$\Delta\epsilon_a < \epsilon < \Delta\epsilon_b$	$2^{\mu_1} - Q$
$\Delta\epsilon_b < \epsilon < \Delta\epsilon_c$	$2^{\mu_2} - Q$
$\Delta\epsilon_c < \epsilon$	$2^{\mu_3} - Q$

TABLE 3. *ERLE* AND TIME TO CONVERGENCE FOR NOISE AND SPEECH ($\eta \propto \epsilon$)

Echo Path Type	Input Signal	Final *ERLE*	Final *OERL*	Convergence to 90 Percent of Final Value
A (Simple)	Noise	19.5 dB	25.5 dB	350 ms
B (Complex)	Noise	20.0 dB	26.0 dB	400 ms
A (Simple)	Speech	17.5 dB	23.5 dB	6 words
B (Complex)	Speech	15.0 dB	21.0 dB	8 words

The frequency response of the overall echo path from the receive-in to send-out ports of the echo canceller is obviously of interest. Figures 10 and 11 show such responses, obtained by using the experimental echo canceller, for convergence on random noise and speech, respectively. Figure 10 shows the echo path attenuation versus frequency characteristics before and after convergence on random noise. Similarly Figure 11 shows these characteristics before and after convergence on speech. Note that the converged response is characterized by ripples of periodicity $(2NT)^{-1}$, which is the reciprocal of twice the length of the processing window. Because of the different spectral distributions of speech and random noise, the overall response is flatter after convergence on noise than after convergence on speech.

Case III—Model with Δx Proportional to $|x|$

The data given thus far were obtained by using a receive-side signal level of -10 *VU* at 0 *TL*. This is a relatively high signal level. The performance of the basic unmodified canceller degrades significantly for low-

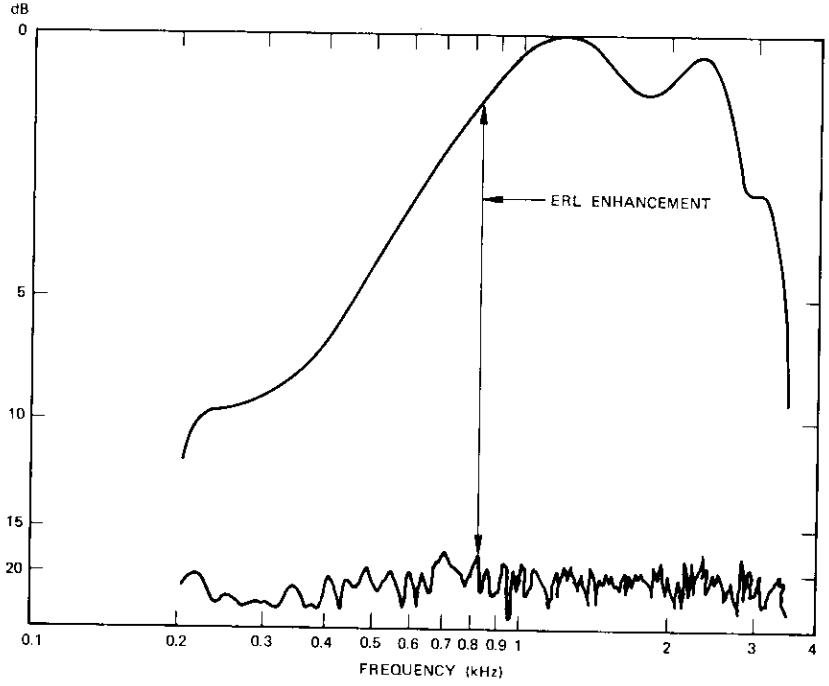


Figure 10. ERL Enhancement vs. Frequency After Convergence on Noise for Circuit B

level inputs of the received signal. The limiting factor is the Δx threshold in the correction sensor of the adaptive control loop. This threshold acts as a receive signal detector, disabling the adaptive loop when a low-level receive signal is present. The function of Δx is to establish the threshold which, when exceeded, allows $h(iT)$ to be corrected if $|\epsilon(t)| > \Delta\epsilon_a$. If the receive signal level is so low that most of the $|x(j - i)\ T|$ values are less than Δx, very few corrections are permitted and performance in terms of convergence time and cancellation at convergence is degraded.

One way to overcome this problem is to decrease the fixed value of Δx. However, if this is done for high-level receive-side signals, too many corrections occur; consequently noise is increased. This leads to reduced cancellation.

This problem is overcome by making the threshold, Δx, a function of

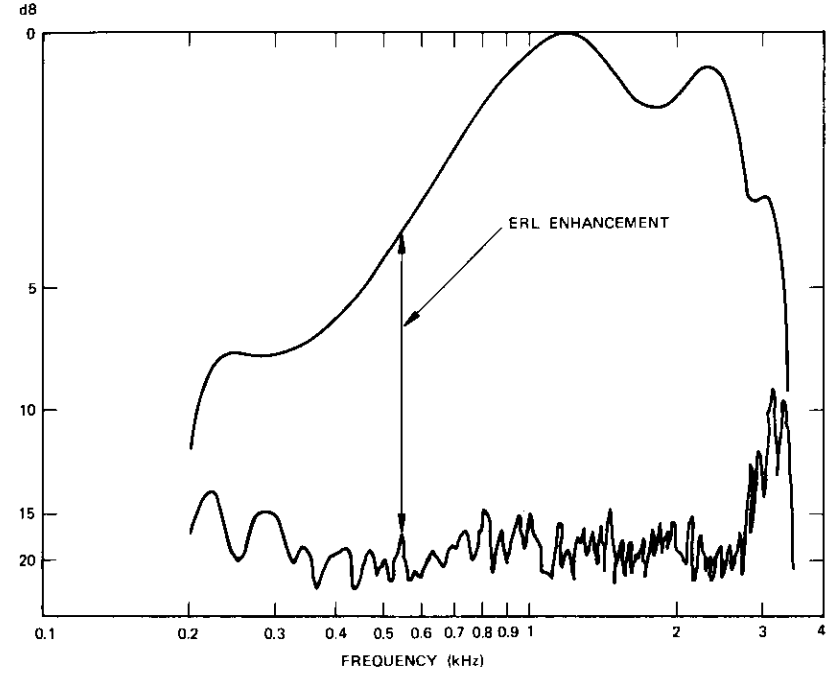


Figure 11. ERL Enhancement vs. Frequency After Convergence on Speech (30 words) for Circuit B

the received signal level, $|x|$; i.e., $\Delta x = \phi_x(|x|)$. It has been found practical to equate the threshold to the average of the absolute values of the contents of the X register. (Recall that this is also the value obtained in Equation (28)). Of course, a minimum threshold is still needed to prevent operation on noise-like signals and to retain the disabling functions in the absence of receive-side speech. Hence, the modifications incorporated in case II are also incorporated here..

As shown by the results, this implementation allows good convergence over a wide dynamic range of the received signal, x. Figure 12 shows the performance of convergence on speech down to -30 *VU* at 0 *TL*. Also shown for comparison are performance curves without the variable Δx threshold. Even though the attained cancellation is not as high for the low signal level, it should be remembered that less *ERLE* is needed to

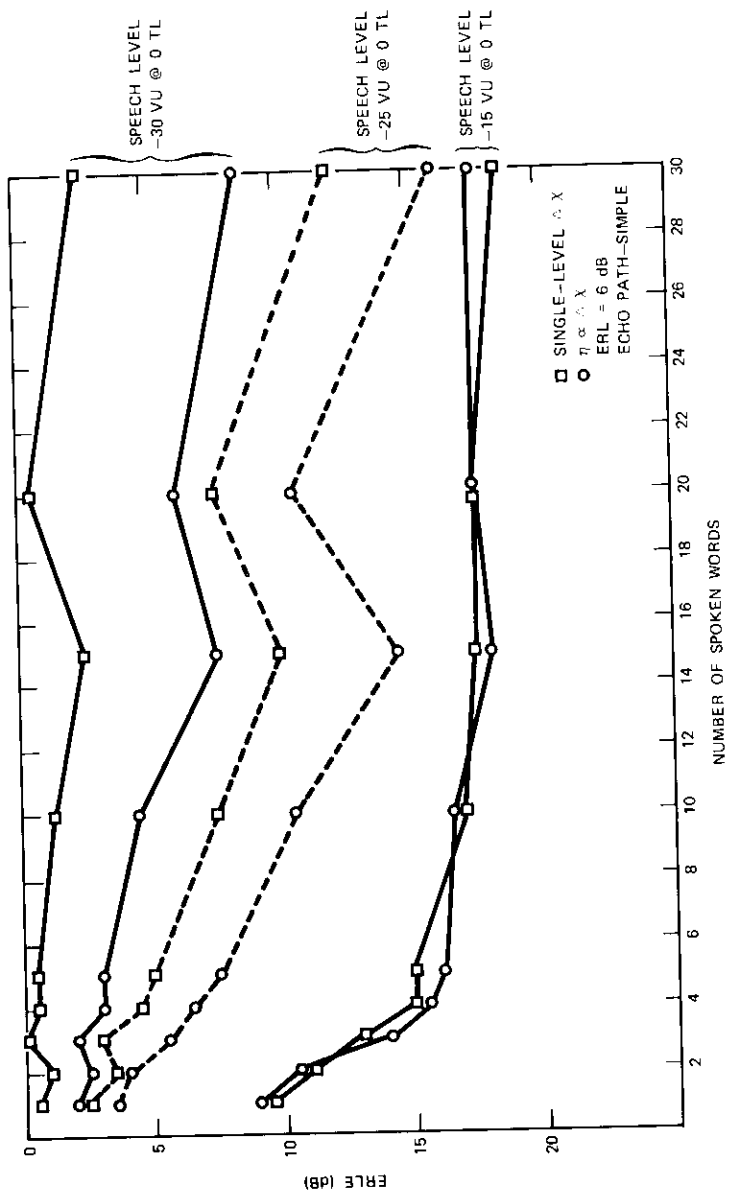


Figure 12. Canceller Convergence as a Function of Speech Level with η Proportional to ε and Δx Proportional to |x|

satisfy good echo tolerance requirements at this level. The results are also shown in tabular form in Table 4. Note that the *OERL* and *ERLE* are significantly enhanced for a receive-side input of −30 *VU*.

TABLE 4. *ERLE* AND RATE OF CONVERGENCE FOR SPEECH IN THE ECHO CANCELLER WITH Δx PROPORTIONAL TO $|x|$ FOR 6-dB *ERL* ECHO PATH A

Input Signal Level at 0 *TL* (*VU*)	Final Values of *OERL* and *ERLE* (dB)				Convergence to 90 Percent of Final Value (words)	
	Single-Level Δx		Δx Proportional to $\|x\|$		Single-Level Δx	Δx Proportional to $\|x\|$
	ERLE	*OERL*	*ERLE*	*OERL*		
−15	18.5	24.5	17.5	23.5	7	5
−25	12.0	18.0	16.0	22.0	22	19
−30	2.5	8.5	8.5	14.5	26	22

Case IV—Model with Syllabic Compandors

During the course of this investigation, it was consistently noted that the canceller performed best for high receive-side signal level and low echo path loss. This is attributed to the quantizing errors in the H and X registers. The magnitude of the quantizing error in each of these registers is independent of the magnitudes of the stored quantized values and dependent only on the number of quantizing levels; i.e., the number of bits. Thus, it should be expected that the effects of quantizing error will be least for large magnitudes of the stored values. This is particularly true in the X register when a high-level signal occurs on the receive-side of the canceller, and in the H register when the echo return path loss is small; i.e., approximately 6 dB. When either the receive-side signal is small or the echo path loss is high, serious performance degradation caused by quantizing distortion occurs.

A scheme to alleviate this problem is shown in Figure 13. It incorporates judicious use of two syllabic compressors and a syllabic expander. Compressor 1 is located at the canceller receive-side input and causes the values stored in the X register to be increased by G_1 dB. As the level of

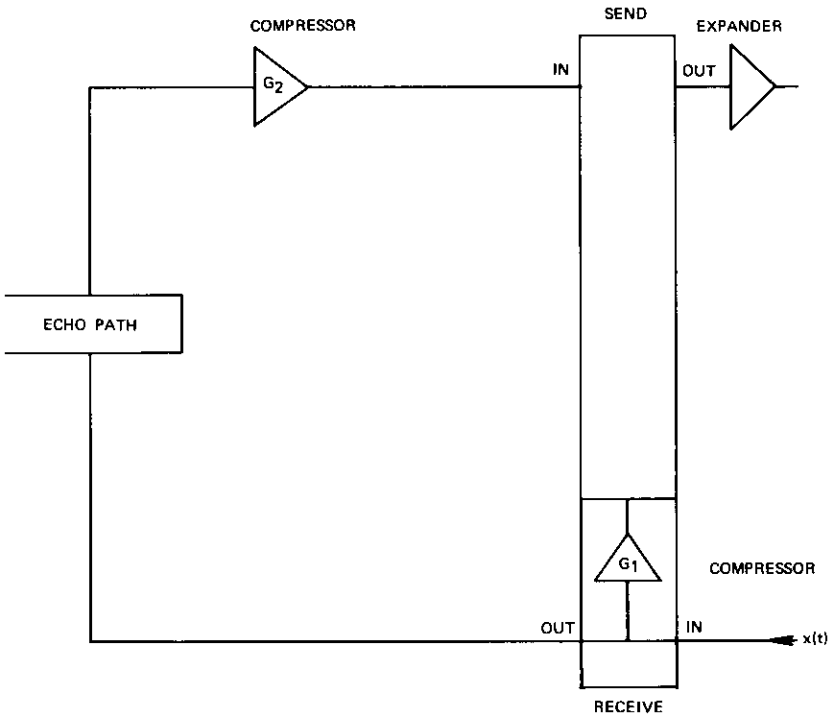


Figure 13. Compandored Echo Canceller

$x(t)$ varies, G_1 varies so that the level variations admitted to the X register are diminished. The compressor used has a 2:1 input/output dB characteristic. Thus, a 20-dB variation at the x input results in a 10-dB variation in the values stored in the X register.

A similar compressor is placed in the send-in path of the canceller. It attempts to reduce the level changes occurring in the echo path. If its gain is G_2, then the effective echo path loss, including the compressor gain, is given by

$$\text{effective } ERL = ERL - G_2.$$

For a high-level receive-side input signal so that G_1 is equal to zero, G_2 is quite small and the effective *ERL* is approximately equal to the true *ERL*. However, if the true *ERL* decreases significantly, then G_2 increases, thus diminishing the effective *ERL*. Since the echo path loss determines the magnitude of the values stored in the H register, the inclusion of compressor 2 tends to diminish variation in the stored h values.

The function of the expander at the canceller output is to restore as closely as possible the input level on the send path for the near-end talker. In doing so, the expander will treat the residual echo as a low-level noise, and a signal-to-idle-noise advantage typical of compandored telephone circuits will also be realized.

Figures 14 and 15 are photographs showing convergence on noise using the simple echo path of Figure 5. In Figure 14, a noise level of −8 dBm0 was used with an *ERL* of 6 dB. Convergence was completed in about 150 ms with a resulting *ERLE* of 36 dB, thus providing a residual echo level of −50 dBm0. In Figure 15, the noise level was dropped to −28 dBm0. Convergence was completed in 300 ms, with a 29-dB *ERLE* and a residual echo level of −63 dBm0. Table 6 summarizes the results.

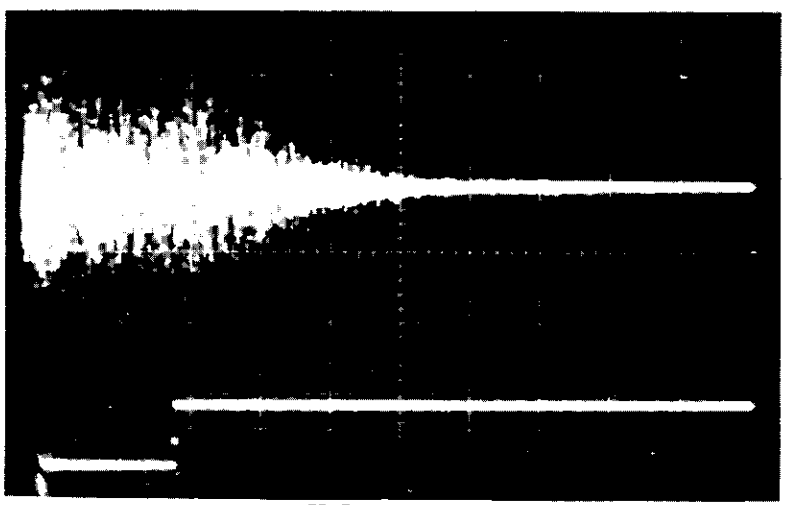


Figure 14. Compandored Echo Canceller Convergence on Noise at −8 dBm0 for 6-dB Echo Path

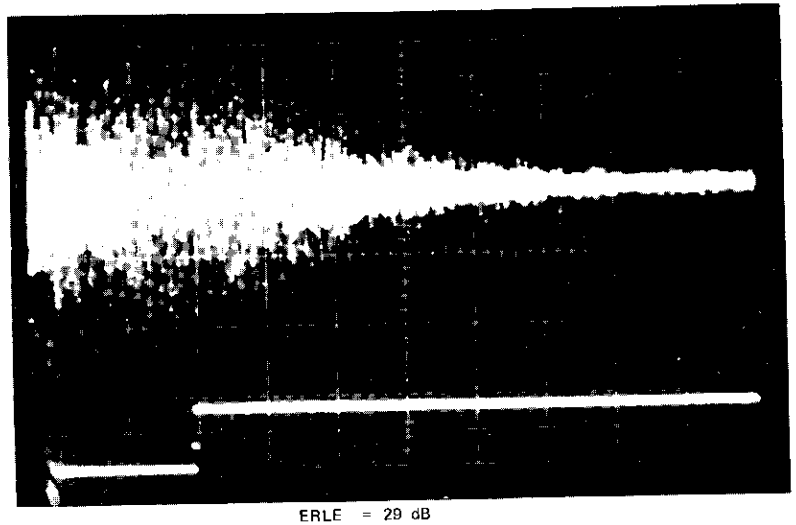


Figure 15. Compandored Echo Canceller Convergence on Noise at −28 dBm0 for 6-dB Echo Path

TABLE 6. *ERLE* AND CONVERGENCE TIME ON RANDOM NOISE FOR COMPANDORED ECHO CANCELLER

ERL (dB)	Input Signal Level (dBm0)	*ERLE* (dB)	*OERL* (dB)	Residual Echo Level (dBm0)	Convergence Time (ms)
6	−8	36	42	−50	150
6	−28	29	35	−63	300
15	−8	32	47	−55	200
15	−28	31	46	−74	350

Figure 16 shows the *ERLE* versus the number of spoken words for various values of speech level. It is interesting to note that, after the single word "hello," the residual echo levels for −10- and −20-*VU* receive speech were −40 dBm0 and −36 dBm0, respectively; these values are

below the −31-dBm0 sensitivity threshold of conventional switching-type echo suppressors. The maximum attained *ERLE* of 32 dB produced a residual echo level of less than −48 dBm0 for speech.

Figure 16 presents the performance of the compandored canceller in an echo path with compandored FDM-derived channels. As may be seen by comparing these results with those of Figure 17, the overall performance of the echo canceller was only slightly degraded by the complex compandored echo path. A summary of the performance achieved for the different types of echo paths is given in Table 7.

TABLE 7. COMPANDORED ECHO CANCELLER PERFORMANCE ON SPEECH

Echo Path	*ERL* (dB)	Input Signal Level at 0 *TLP* (*VU*)	*ERLE* (dB)	*OERL* (dB)	Residual Echo Level (dBm0)	Convergence to 90 Percent of Final Value (words)
Simple Path A	6	−10	32	38	−48	4
Simple Path A	6	−20	32	38	−58	4
Simple Path A	6	−50	14	20	−70	8
Compandored Complex Path B	9	−10	26	35	−45	3
Compandored Complex Path B	9	−20	26	35	−55	3
Compandored Complex Path B	9	−50	16	25	−75	24

Effect of Phase Roll

Phase roll is the phenomenon caused by carrier frequency offset in transmit and receive modems at one end of a carrier channel. It makes the impulse response of the echo path a time-variant function, which can in turn significantly affect the canceller's performance.

In order to investigate this phenomenon, phase roll was introduced by modem carrier frequency offset in the echo path used to obtain the results of Figure 16. *ERLE* measurements were made by using receive-side random noise levels of −10 and −30 dBm0 as the test signal. The results are shown in Figure 18, where the degradation in *ERLE* is given as a function of phase roll for two levels of the test signal.

It is seen that, for phase roll rates as great as 1.2 rad/s, relatively high *ERLE* values still result. It must be remembered, however, that, during a

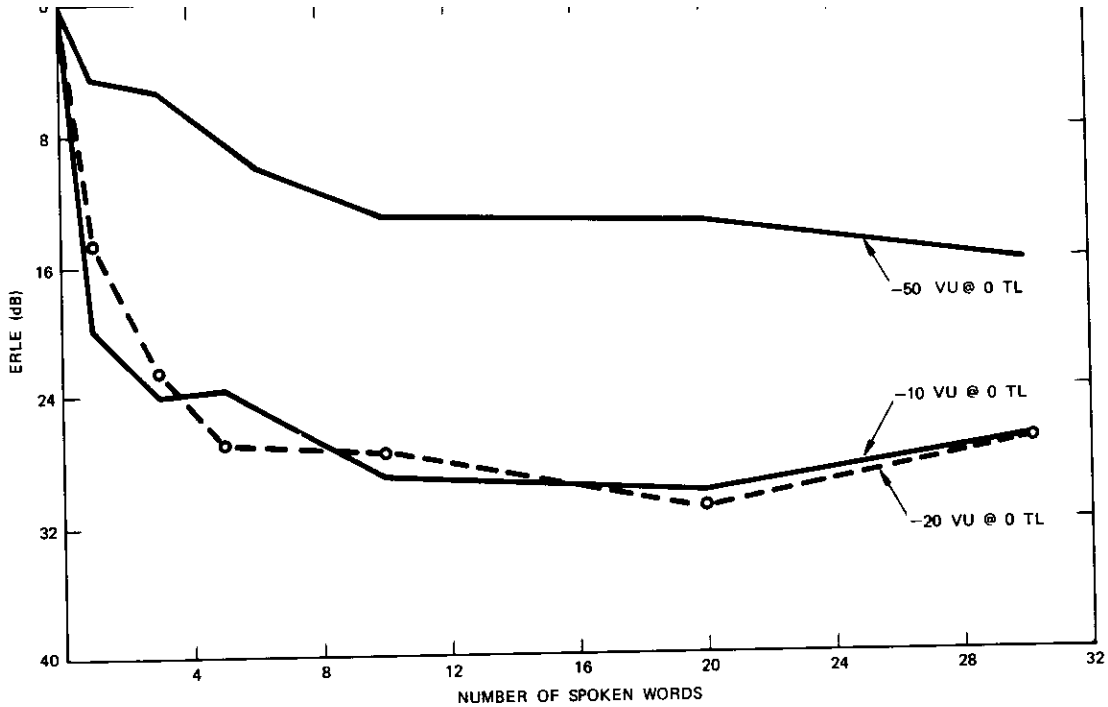


Figure 16. Compandored Echo Canceller Performance on Speech for a Compandored Echo Path

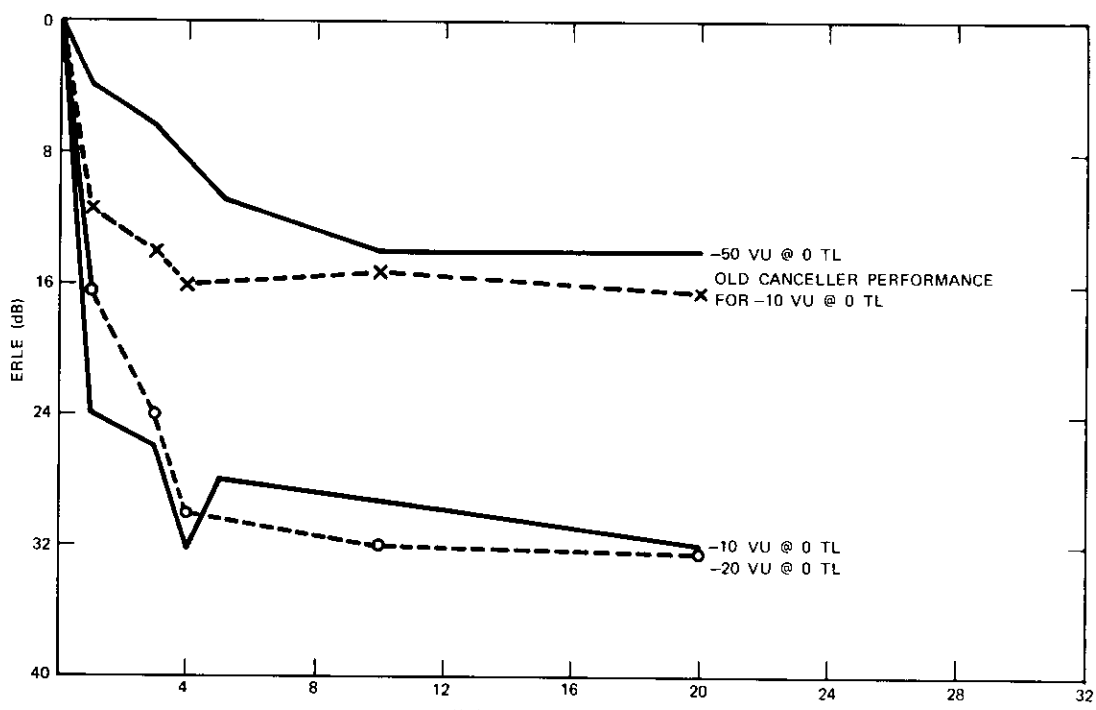


Figure 17. Compandored Echo Canceller Performance on Speech for a Simple 6-dB Echo Path

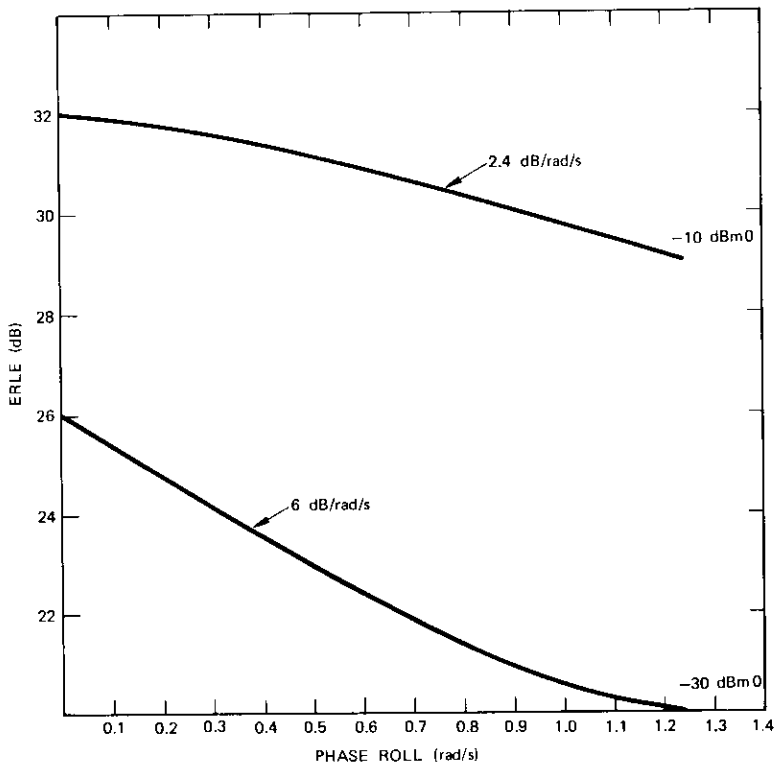


Figure 18. Effect of Phase Roll on ERLE

pause in the speech, convergence is not maintained. If the phase roll rate is high this will necessitate reconvergence at every onset of receive-side speech. For slow phase roll rates this may not constitute a problem; however, at high roll rates, serious performance degradation may occur. It is likely that the best means to cope with this problem is to eliminate the source of the phase roll.

CONCLUSIONS

The incorporation of additional adaptive features and appropriate use of compressors and expanders has made it possible to demonstrate an

experimental echo canceller possessing an overall echo return loss capability of 38 dB for receive-side voice levels ranging from 0 to −20 *VU* at 0 *TL*. The residual echo level is never greater than −48 *VU* at 0 *TL* for received signal levels below −10 *VU* at 0 *TL*. The time to convergence for random noise for this configuration is between 150 and 200 ms. For speech, convergence is accomplished on the utterance of four to five words. Experiments with this modified echo canceller on a laboratory circuit with a phase roll of 0.5 rad/s exhibit a degradation of only 3 dB in the echo return loss enhancement.

List of Symbols

N	number of samples used for convolution processing
T	sampling interval
j	index to indicate any sampling instant
i	index to indicate any of the N samples
ϕ_x, ϕ_ϵ	asymmetrical, non-decreasing functions of x and ϵ, respectively
x	incoming speech signal
y	echo signal
$\hat{y}$	modelled echo signal
ϵ	error signal
pdf	probability density function
Q	number of bits used for quantizing samples in H register
L	number of bits used for quantizing samples in X register
TL	transmission level designation
dBm0	power in dBm referred to the 0 *TL*
VU	volume unit
ERL	echo return loss
ERLE	echo return loss enhancement
OERL	overall echo return loss (*ERL* + *ERLE*)

REFERENCES

[1] O. B. Blackwell, "The Time Factor in Telephone Transmission," *Bell System Technical Journal*, Vol. 11, No. 1, Jan. 1932, pp. 53–66.

[2] J. W. Emling and D. Mitchell, "The Effects of Time Delay and Echoes on Telephone Conversations," *Bell System Technical Journal*, Vol. 42, No. 6, November 1963, pp. 2869–2891.

[3] R. R. Riesz and E. T. Klemmer, "Subjective Evaluation of Delay and Echo Suppressors in Telephone Communications," *Bell System Technical Journal*, Vol. 42, No. 6, November 1963, pp. 2919–2941.

[4] S. J. Campanella, H. G. Suyderhoud, and M. Onufry, "Subjective Evaluation of Dedicated Multiple-Hop Satellite Communications for Government and Military Users," *IEEE Transactions on Communications Technology*, Vol. COM-18, No. 5, Oct. 1970, pp. 663–672.

[5] H. G. Suyderhoud, "A Survey of Echo Suppressor Progress," *Telephony*, Vol. 180, No. 13, March 29, 1971, pp. 30–32.

[6] P. T. Brady, "Effects of Transmission Delay on Conversational Behavior in Echo-Free Telephone Circuits," *Bell System Technical Journal*, Vol. 50, No. 1, January 1971, pp. 115–134.

[7] D. L. Richards, "Transmission Performance of Telephone Connections Having Long Propagation Times," *Het PTT Bedrijf* [Netherlands], Vol. XV, No. 1–2, May 1967, pp. 12–24.

[8] E. T. Klemmer, "Subjective Evaluation of Transmission Delay in Telephone Conversations," *Bell System Technical Journal*, Vol. 46, No. 6, July-August 1967, pp. 1141–1147.

[9] R. G. Gould and G. K. Helder, "Transmission Delay and Echo Suppression," IEEE *Spectrum*, Vol. 7, No. 4, April 1970, pp. 47–54.

[10] A. Miura, R. Sato and K. Nagata, "Subjective Evaluation of Delay and Echo Suppressors in Telephone Communication," *Journal of the Acoustical Society of America*, Vol. 40, 1966, p. 1255.

[11] M. M. Sondhi, "An Adaptive Echo Canceller," *Bell System Technical Journal*, Vol. 46, No. 3, March 1967, pp. 497–511.

[12] Unpublished INTELSAT Contract Final Report, February 1970.

[13] T. N. E. Greville, "The Pseudoinverse of a Rectangular or Singular Matrix and its Application to the Solution of Systems of Linear Equations," *SIAM Review*, Vol. 1, No. 1, Jan. 1959, pp. 38–43.

ACKNOWLEDGMENT

The authors wish to acknowledge Mr. Y. Kato of the Nippon Electric Company who was responsible for development and construction of the basic breadboard model of the echo canceller.

Dr. S. J. Campanella received a B.S.E.E. from Catholic University of America, a M.S.E.E. from the University of Maryland, and the Ph.D. in electrical engineering from Catholic University of America. Before joining COMSAT, he was Manager of the Electronics Research Center at Melpar, Inc.

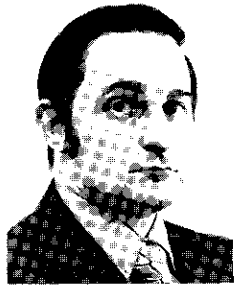
Presently, he is Manager of the Baseband Processing Branch of the Communications Processing Laboratory, COMSAT Laboratories, where he is responsible for communications signal processing research and development.

Dr. Campanella is a member of IEEE, Sigma Xi, Phi Eta Sigma, and the Acoustical Society of America.

H. G. Suyderhoud was born in Borneo, Indonesia. He received the Ingenieur degree from Delft Technological University in the Netherlands in 1955, and the M. S. degree in Applied Mathematical Statistics from Rutgers in 1966.

He is presently head of the Baseband Techniques Section in the Communications Processing Laboratory of COMSAT Laboratories, where he is responsible for the development of analog and digital speech operated devices. From 1957 to 1968, he worked for Bell Telephone Laboratories.

Mr. Suyderhoud is a member of the IEEE and the American Statistical Association.



Michael Onufry obtained a B.S. from the Pennsylvania State University in 1962 and has done graduate work at both Northeastern University and George Washington University. He is a member of the IEEE.

Mr. Onufry is a Member of the Baseband Signal Processing Branch in the Communications Processing Laboratory of COMSAT Laboratories. His work has included designing a digital echo canceller and conducting experiments to study the effects of delay and echo on telephone conversations. Before joining COMSAT, he was engaged in research and development for both RCA and Fairchild Hiller Corporation.

Index: analog to digital converters, color television, communication satellites, digital techniques, digital to analog converters

A 150 Mbps A/D and D/A conversion system

O. A. HORNA

ABSTRACT

An analog-to-digital and digital-to-analog converter was developed for digital transmission of color TV signals over communications satellites. Five different building blocks on standard printed circuit cards are used, providing eight-bit precision with a total conversion time less than 50 ns.

No operational amplifiers are used inside the A/D conversion loop; all voltage adding and subtracting is done directly at the input of the comparators. The A/D loop contains a digital feedback circuit which can correct the quantization errors in the front stage (or stages), thus reducing the required settling time and required accuracy of the first-stage approximation.

Temperature-compensated current switches with a low-resistance ladder network for summing and weighting are used in D/A converters. A unique current summing configuration is used for the two most significant digits in order to decrease the precision requirement, the power dissipation, and the transient voltage during the transition from the 01 to the 10 state or *vice versa*.

A method for measuring the exact relationship between the analog input and the digital output has been developed. The measurements have shown an eight-bit conversion accuracy ($\pm 2 \times 10^{-3}$) with a 15-MHz sampling rate.

INTRODUCTION

A new pulse-code modulation (PCM) terminal developed at COMSAT Laboratories for transmission of color television signals over communications satellites uses several A/D and D/A converters with different numbers of quantization levels (2^4 to 2^8) and conversion times (100 ns to 2 μs). Since none of the D/A and A/D conversion equipment presently on the market satisfies the special requirements of the terminal, a new modular system using five different building blocks and standard inexpensive integrated circuits was developed. Each building block is on a double-sided printed circuit card. These building blocks are:

a. a precision sample-and-hold circuit with an aperture time of less than 1 ns,

b. an eight-level comparator circuit with partial decoding for a three-bit output and a response time of less than 20 ns,

c. a decoder for five bits with a five-bit memory and a digital correction feedback circuit, and

d. two precision D/A converters with settling times shorter than 20 ns. Both of these converters are of the same design; one is intended for D/A conversion at the receiving end, and the other is for the feedback loop in A/D converters.

From these five cards, different parallel and series-parallel converters are assembled. Of all A/D conversion techniques [1], [2], [3], parallel conversion has the shortest conversion time for a given speed of hardware; it is used for less than five bits. Series-parallel conversion is a compromise between speed and complexity and is therefore used beyond six bits.

D/A Converter

Since speed was the prime objective, the circuit in Figure 1 was chosen for D/A conversion [1], [2], [3]. It consists of a resistive ladder network and $n + 1$ electronic switches, S_0 to S_n, which feed currents to different nodes of the network according to the logic signals at $n + 1$ parallel digital inputs [4].

The current-driven ladder network is built in "closed form" with discrete components in order to keep parasitic inductances, L_s, to a minimum. The network time constant, L_s/R, which is less than 2 ns, results in a settling time of 15 ns to reach 99.8 percent of the final value. This low-resistance network can be loaded by a capacitance, C_X (see Figure 1),

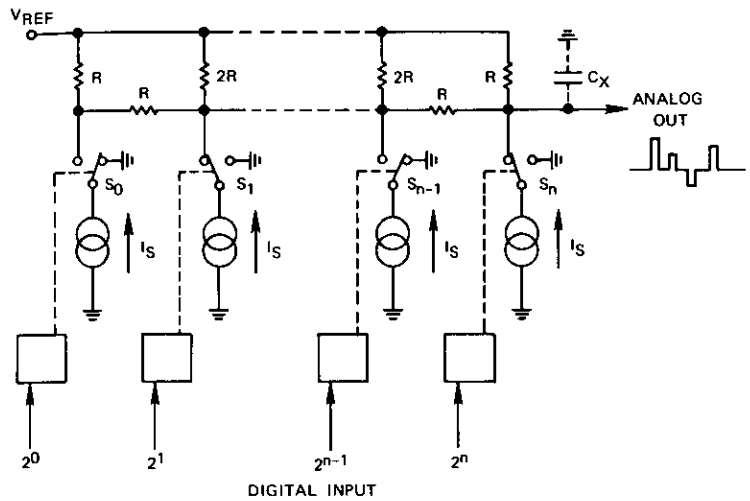


Figure 1. Block Diagram of the D/A Converter with a Current-Driven Ladder Network

of several pF without reducing the speed of the response. It can therefore drive comparator inputs without a buffer amplifier, which would introduce additional delay and settling time. For values of R in the range of 50 to 500 ohms, and for a full analog output voltage swing, $\Delta V = 5V$,* the current I_s is between 7.5 and 75 mA, which is within the optimal range for fast transient response of the bipolar silicon transistors developed for the current switches and used in the emitter coupled logic (ECL) circuits.

Figure 2 is a circuit diagram of an electronic switch (S_n) with a current source. The fast ECL gate is used as a line receiver and switch driver. Its output transistor, Q_1, and an external transistor, Q_2, form the switch, S_n, and the source of current, I_s, with an equivalent internal resistance of several megaohms, which has negligible effects on the low-resistance ladder network.

If the voltage at the base of transistor Q_1 is lower than the bias voltage,

* The value of ΔV is determined from the parameters of the comparators.

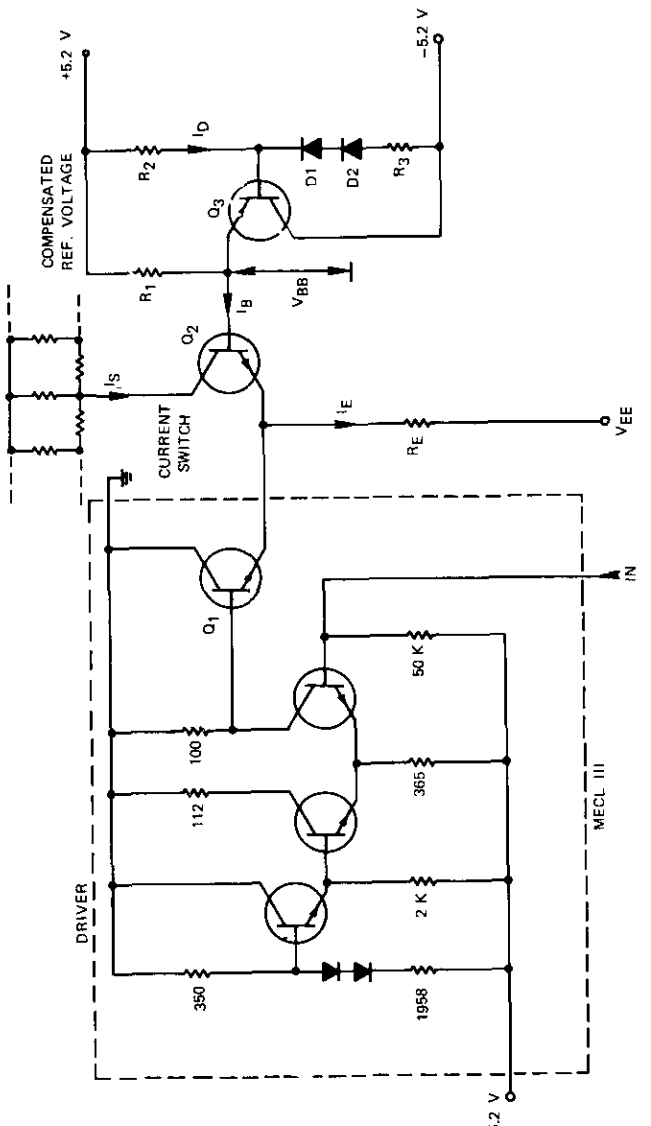


Figure 2. Current Switch and Compensated Current (I_S) Source for D/A Converter

V_{BB} (see Figure 2), Q_1 is cut off and I_E is routed through Q_2 in the ladder network. The current I_S is

$$I_S = I_E - I_B = \frac{h_{FE}}{1 + h_{FE}} I_E = \frac{h_{FE}}{1 + h_{FE}} \frac{V_{EE} - V_{BB} - V_{EB}}{R_E} \tag{1}$$

where h_{FE} is the DC gain of Q_2, and V_{EB} is its emitter-to-base voltage. Both quantities are temperature dependent. By definition, the temperature coefficient of h_{FE} is

$$\theta_{FE} = \frac{1}{h_{FE}} \frac{dh_{FE}}{dT_j} \tag{2}$$

and the temperature coefficient of V_{EB} is

$$-\theta_{EB} = \frac{1}{V_{EB}} \frac{dV_{EB}}{dT_j} . \tag{3}$$

The temperature coefficient of the current I_S is therefore

$$\theta_S = \frac{dI_S}{dT_j} = \frac{\theta_{FE}}{1 + h_{FE}} + \frac{V_{EB}\theta_{EB}}{V_{EE} - V_{BB} - V_{EB}} \tag{4}$$

In a typical circuit with $h_{FE} + 1 = 100$, $V_{EE} - V_{BB} - V_{EB} = 25$ V, $\theta_{FE} = 7 \times 10^{-3}/°C$, and $V_{EB}\theta_{EB} = 1.8$ mV/°C, the temperature coefficient [5] of the current I_S is

$$\theta_S = 7 \times 10^{-5} + 7.2 \times 10^{-5} = 0.142 \times 10^{-3} \left[\frac{1}{°C}\right]. \tag{5}$$

With temperature change $\Delta T_j = 14°C$, the change in I_S is $\Delta T_j \cong 2.10^{-3}$, or one-half of the least significant bit (LSB) in an eight-bit converter ($1/2^8 \approx 4.10^{-3}$).

The junction voltage change of the NPN transistor, Q_2, is compensated by an equal junction voltage change of the PNP transistor, Q_3, which supplies the base voltage, V_{BB}. The part of θ_S which is caused by changes of h_{FE} is compensated by the temperature-dependent voltage, V_{BB}, i.e., by two diodes, D_1 and D_2, in the divider for the base voltage of Q_3. A nearly perfect compensation of θ_{FE} can be achieved by properly choosing the divider current, I_D.

Because θ_S is relatively small, this compensation is required only for the two or three more significant bit (MSB) switches. To reduce the power and number of precision resistors in the ladder network, the three less significant bits are converted in one resistor, R, by switching currents I_s, $I_s/2$, and $I_s/4$, as is shown by transistors Q_0, Q_1, and Q_2 in Figure 3.

Digital Resampling

The analog output (see Figure 1) of the D/A converter is usually resampled in order to eliminate the transients and to adjust the width of the amplitude-modulated output pulse (PAM) according to the required frequency characteristic [6].

A transient appears at the output of the ladder network when a new digital input changes the state of different switches; the primary cause of this transient is the timing differences at the digital input. Such a transient can be reduced by more than an order of magnitude when a D-type flip-flop (Figure 3) is used to drive the current switch Q_2. After all of the logic input voltages settle down, a short "encode command" pulse (see Figure 3) simultaneously changes the states of all switches according to the new logic input. Another pulse fed to the reset input of the flip-flops resets all switches, Q_0 to Q_n, to the nonconducting state; i.e., the analog output is reset to the zero level. The width of the pulse at the analog output can be adjusted by adjusting the timing of the "encode command" and "reset" pulses.

It is the transition from the 0111 . . . state to the 1000 . . . state which causes the greatest transient error. The analog output voltage changes only by the smallest quantization level, but the two MSB switches change their state. For 95 percent of the time [7], the amplitude of the video signal is between 25 and 75 percent of the maximum value; therefore, the transition from 0111 . . . to 1000 . . . is frequent.

The configuration in Figure 3 reduces the transient by at least a factor of two and also reduces the current I_S by one-half for the given voltage swing and resistance of the ladder network. The two MSB are decoded to switch the currents $I_1 = I_2 = I_3$ into one node of the ladder network according to the table in Figure 3. During the transition from 01 . . . to 10 . . . , only a current I_2 is added to I_3. With the circuit in Figure 3, the amplitude of all transients is held so low that an additional output resampler is unnecessary.

To shift the output voltage level according to the particular use of the D/A converter, another compensated constant current source Q_4, with a

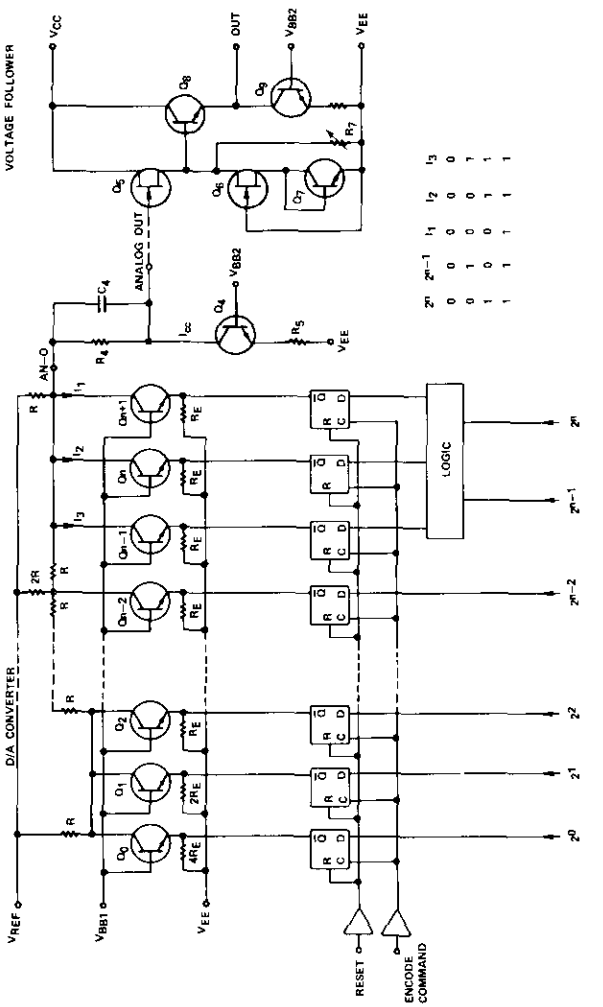


Figure 3. D/A Converter with Command and Reset Inputs and a Temperature-Compensated Voltage Follower with an FET Input

resistor R_4 (which can be the voltage divider for comparators, as shown by resistors R_x, which equal 10 Ω in Figure 4) is connected to the output. When the converter is used at the receiving end, a voltage follower, Q_5 to Q_9, is added to drive a 75 Ω line.

Transistors Q_5 and Q_6 are precisely matched to have the same drain current with −0.75 V gate bias. Because the same current must flow through Q_5 and Q_6, Q_5 must have the same bias as Q_6, i.e., the emitter-to-base voltage, V_{EB}, of transistor Q_7 which is in the same package as Q_8. There is no DC offset between gate Q_5 and emitter Q_8, and the change of $V_{EB}(Q_8)$ with temperature is compensated. A zero temperature drift can be achieved by adjusting resistor R_7. If the amplification factor of transistor Q_8 is $h_{FE} \geq 80$, the gain, A, of the emitter-follower stage is greater than 0.98. The 10- to 90-percent rise time, t_r, is less than 2.5 ns; the settling time (to 99.8 percent) is approximately 15 ns.

Comparators

For conversion times less than 50 ns, only the parallel or series-parallel A/D conversion technique can be used with available components [1], [2], [3]. The analog voltage is directly compared with reference voltages in a circuit using a chain of comparators [7], [8], [9] (see Figure 4). The Motorola MC1650 comparator presently has the required response time, which is under 25 ns. MC1650 consists of a differential amplifier and an ECL gated D-type flip-flop (see Figure 4).

According to the manufacturer's data, the maximum analog input voltage swing is ±2.5 V. The hysteresis was found to be less than ±10 mV, but a mutual influence (crosstalk) between both comparators was detected. This can be substantially reduced if one comparator is connected with interchanged inputs and outputs instead of according to manufacturer's recommendations.

When fully loaded, the MC1650 comparators dissipate approximately 400 mW; therefore, an effective heat-sink is required (see Figure 5). Since the higher junction temperature substantially increases the current gain of all transistors, the operational input bias current I is substantially less than 3 μA, which is the value given for 25°C in the manufacturer's data. For an absolute precision of 2.5 mV, the total resistance of the reference voltage divider does not need to be lower than 160 ohms, even when it is loaded with 16 comparators.

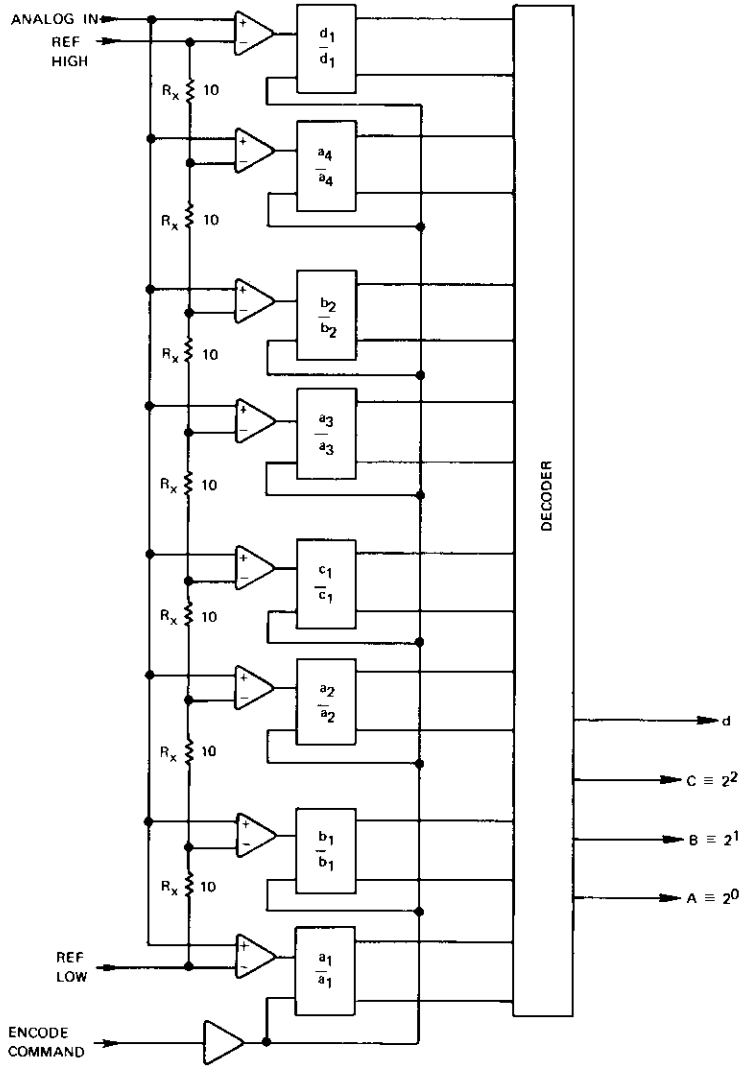


Figure 4. Block Diagram of the A/D Basic Building Block: Eight Comparators with a Decoder with Three-Bit Output

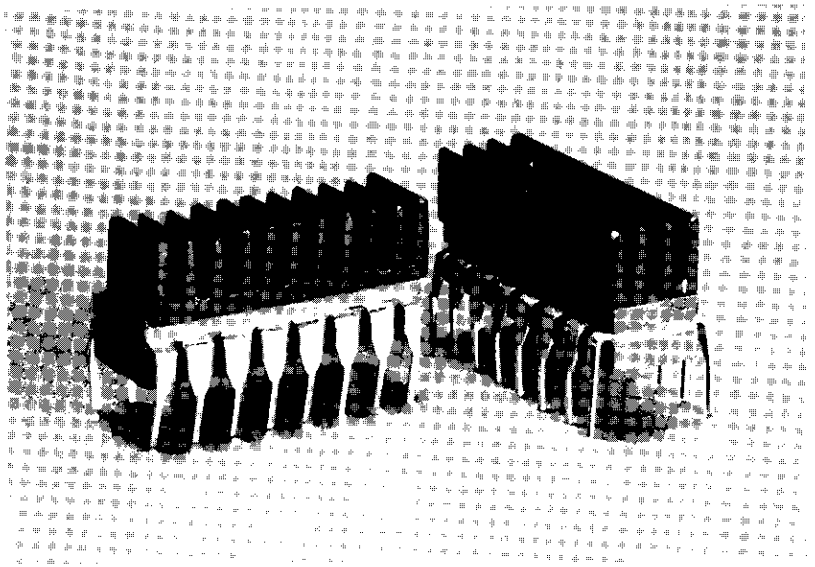


Figure 5. Two Types of Heat-Sinks Used in the A/D-D/A System for Comparators and Other IC

Parallel A/D Converter

Eight comparators (four MC1650) with a reference voltage divider and a logic circuit to decode the output into a three-digit binary number, CBA, form the basic building block for A/D conversion. Two such circuits and a master-slave D-type flip-flop for each bit (memory element) form a four-bit A/D converter for sampling rates over 35 MHz.

To sample the analog signal, the gate input of the comparator is used. The fall time of the gate pulse must be short enough to meet the requirements for quantization precision. A fall time of less than 2 ns can be easily achieved with fast ECL gates connected as a DC coupled monostable.

When a "convert command" pulse is received, the pulse generator (Figure 6) generates a 15-ns positive-going pulse which drives the clock inputs, C_i, of the flip-flops, and the digital information at D_i inputs is transferred to Q_i memory outputs. After this transfer has been completed,

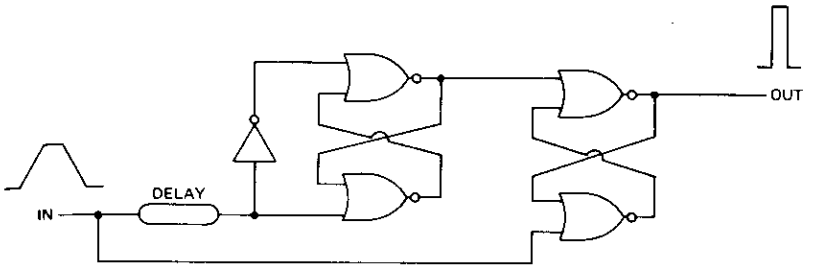


Figure 6. DC Coupled Pulse-Forming Monostable Circuit with Fast Rise and Fall Times Using Standard IC NOR Gates

the delay pulse opens the gates of the comparators (see the "encode command" input in Figure 4) and the comparator's flip-flops change to a new state according to the amplitude of the analog input. The clock inputs of the output flip-flops return to zero, and therefore the D inputs accept this new digital information from the comparator's logic network (see Figure 7). These states will be transferred to the Q_i outputs when another "convert command" pulse is received.

The same timing sequence is used in all converters. It delays the results of the conversion by one full sampling period, but it makes full use of the available conversion time without elaborate circuitry.

Series-Parallel A/D Converter

The parallel-type technique (Figure 7) is also used for the five-bit A/D converter; i.e., four comparator building blocks are connected in series. Because the number of comparators, k, is related to the number of bits, n, by the relationship

$$k + 1 = 2^n \tag{6}$$

the six-bit parallel converter requires 63 comparators and the eight-bit converter would require 255 comparators. This scheme is impractical to implement for obvious reasons; therefore, for $6 \leq n \leq 8$ bits, the series-parallel configuration is used [1], [3], [8], [9]. The six-bit, two-stage A/D converter (Figure 8) is an example of this technique.

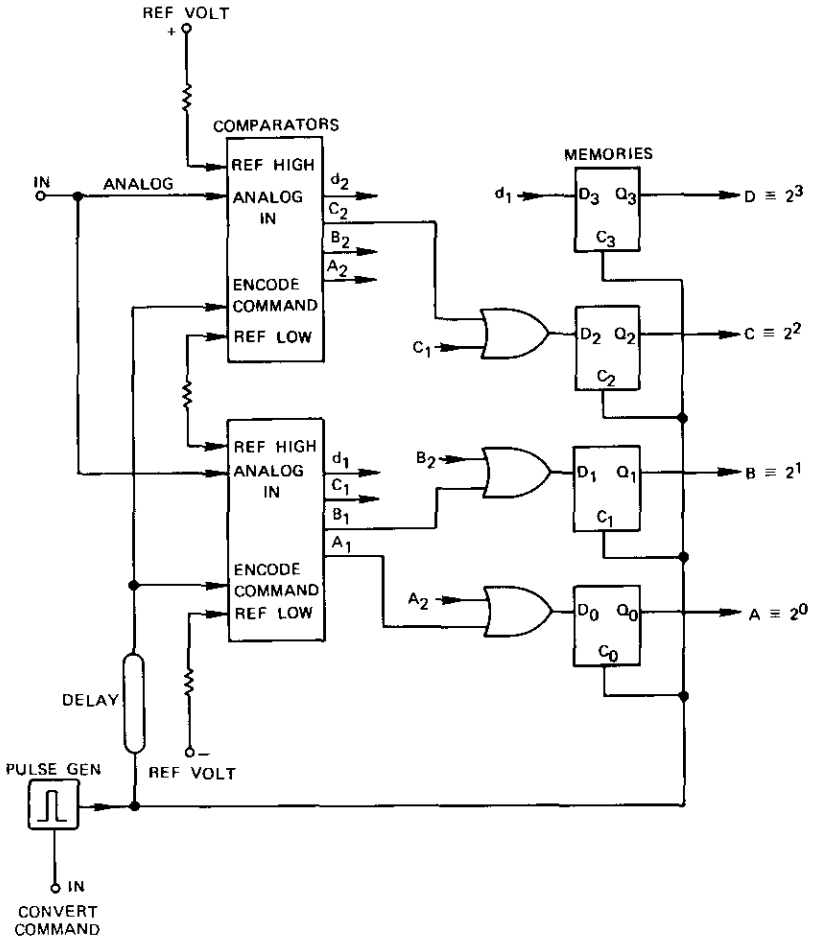


Figure 7. Block Diagram of a Four-Bit A/D Converter

Because the quantization process comprises two successive steps, a sample-and-hold device is needed. The voltage at the output of sampler V_S is quantized with three-bit precision in the first stage. The comparator output is decoded, stored in memories, and fed to a three-bit D/A converter. The output voltage, V_C, which has eight discrete values equal to the eight quantization levels of the first-stage comparators, is subtracted

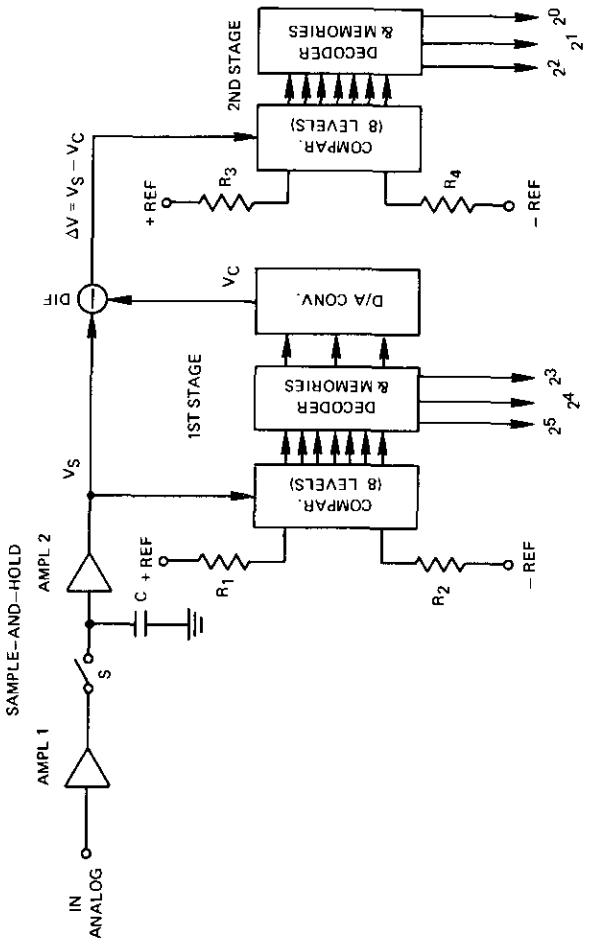


Figure 8. Block Diagram of a Two-Stage, Series-Parallel, A/D Converter with Six-Bit Output

from V_S and the difference voltage, $\Delta V = V_S - V_C$, is quantized in the second stage with three-bit precision. The whole precision is therefore six bits. This approach requires only 14 comparators compared to 63 for a parallel converter.

Nevertheless, the converter contains additional hardware: a sample-and-hold circuit, a D/A converter, and a device to produce a difference voltage, ΔV. Also, if a differential amplifier is used to amplify the difference voltages, ΔV, the ultimate conversion precision is still determined by the properties of the first-stage comparator; i.e.,

$$2^n \leq \frac{V_i}{V_h} \tag{7}$$

where V_i is the maximum input voltage swing (5 V for MC1550) and V_h is the hysteresis voltage (less than 20 mV). This is shown in Figure 9.

Suppose that the quantized voltage, V_{S1}, is less than one quantization level higher than the nearest reference voltage, V_{15}. If because of a slight inaccuracy in the first stage, the fifth comparator (with reference V_{15}) does not switch and remains in the 0 stage, only V_{14} will be subtracted from V_{S1} and the difference will be $\Delta V_1 = V_{S1} - V_{14} > V_{28}$. That is, ΔV_1 is out of the range of the second stage, and, therefore, the output of the comparators is decoded to $V_{S1} \sim 100111$, which is one quantization step in error from the true value, $V_{S1} \sim 101000$.

This example shows that the precision of the two-stage converter in Figure 8 is determined by the absolute precision of the first stage, and that the comparison can only be made after the transients have settled down to a value required by the precision of the whole device. This limits the speed and the precision of the converter with given comparators and sample-and-hold circuit.

Digital Corrective Feedback

All properties of a series-parallel configuration can be substantially improved by using digital corrective feedback as shown in Figure 10. The second stage has added quantization levels (e.g., four in Figure 10) and a logic circuit which detects overrange and underrange signals in the second stage and sends a corrective signal to the adder, which is located between the output of the decoder and the input of the memories of the first stage. Therefore, if the signal ΔV in the second stage is overrange, a 1000 is added to the numerical output of the first stage; if ΔV is under-

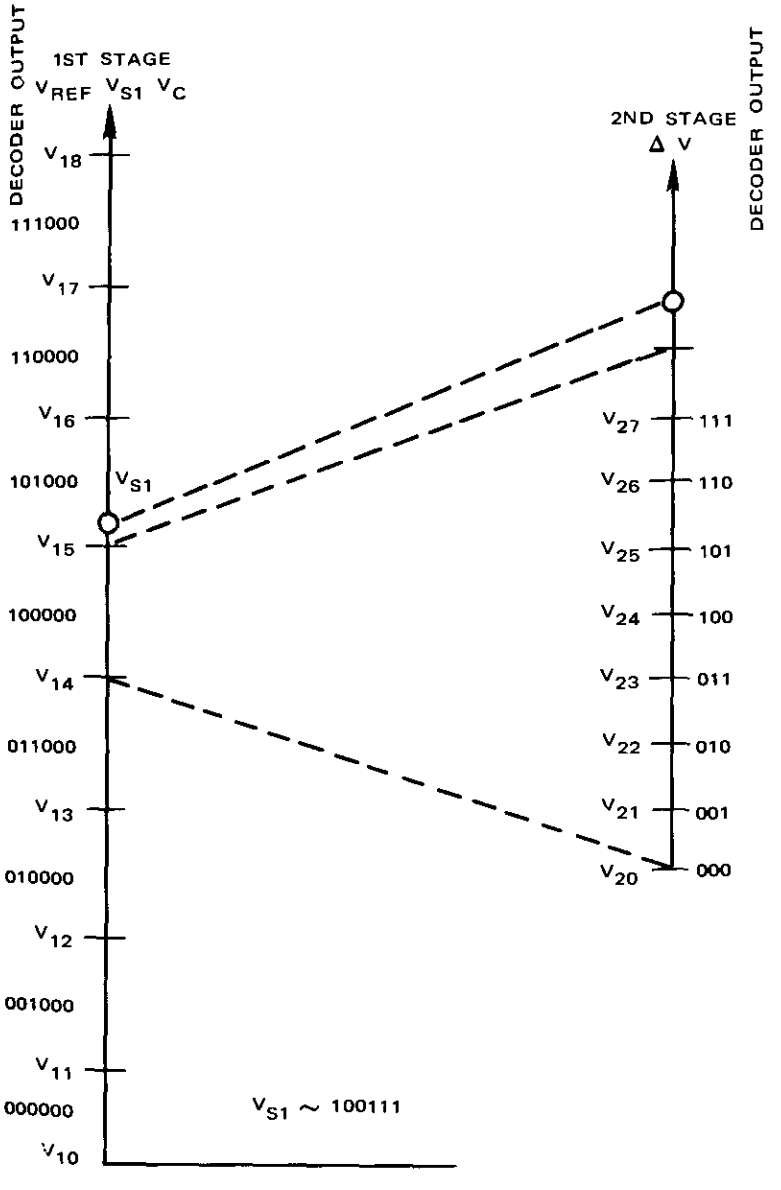


Figure 9. Graphical Representation of the Results of an Erroneous Decision in the First Quantization Stage

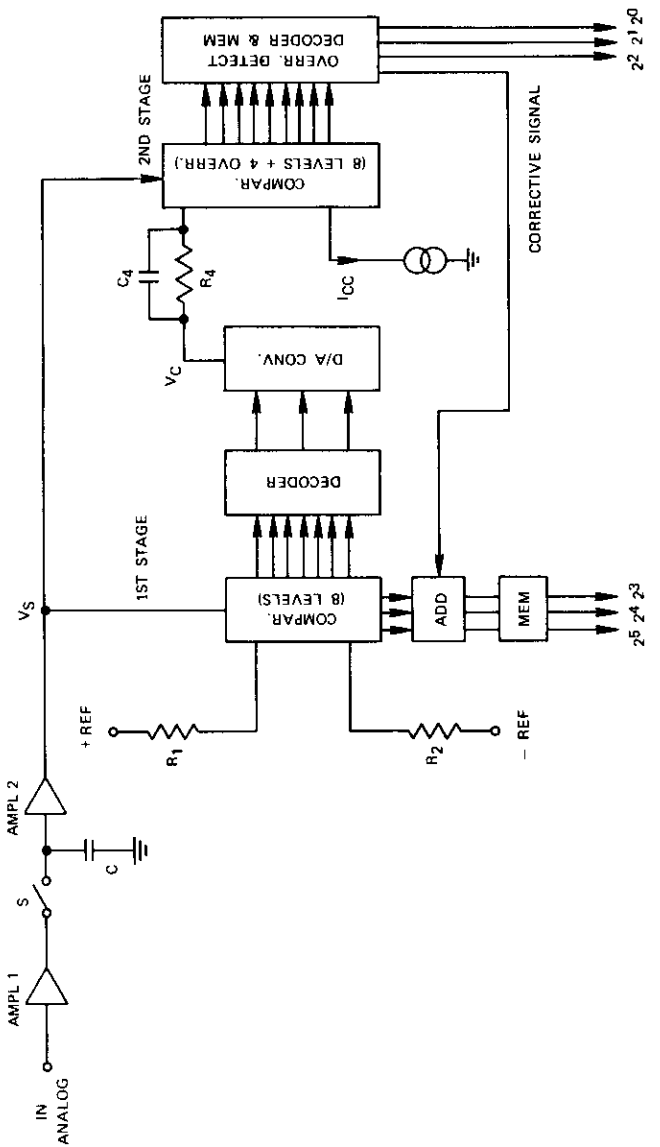


Figure 10. Six-Bit, Series-Parallel, Two-Stage Quantizing Process with Corrective Digital Feedback Loop

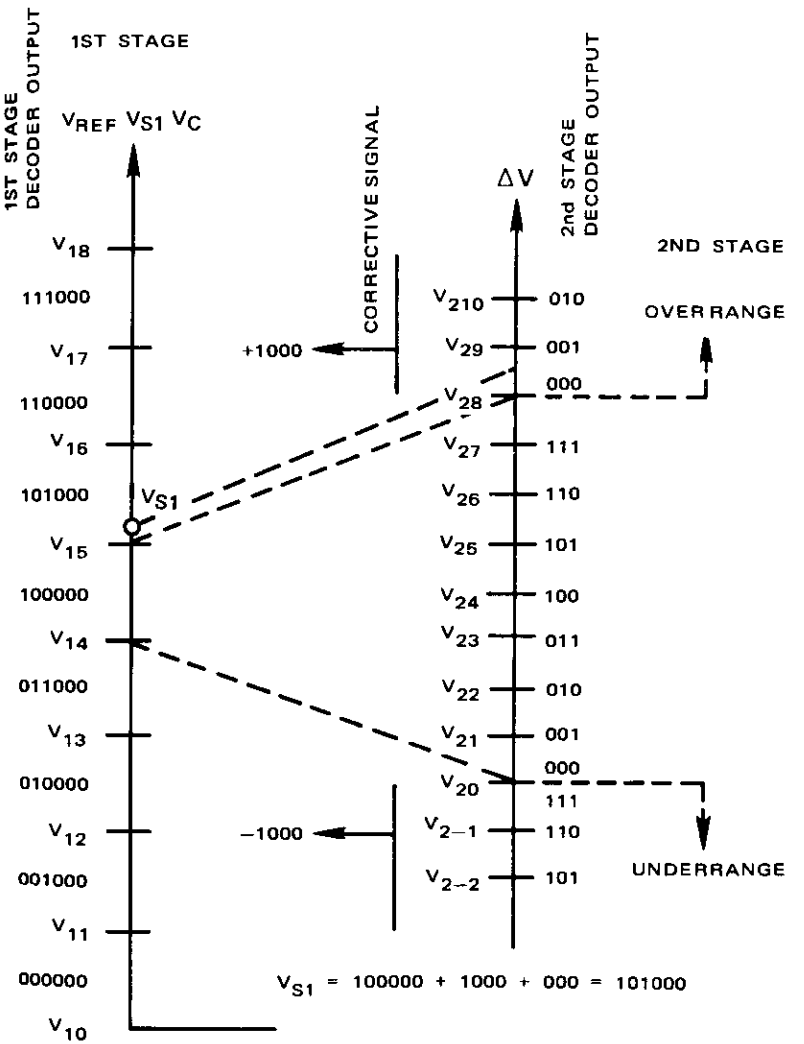


Figure 11. Graphical Representation of a Two-Stage Quantizing Process with Corrective Signal 1000 Added

range, a 1000 is subtracted from the output of the first stage. Hence, as shown in Figure 11, the situation in Figure 9 can never lead to an erroneous conversion, and it is unnecessary to quantize precisely in the first stage.

The range of correctable first-stage errors is determined by the number of additional comparators in the second stage, as shown in Figure 11. A suitable compromise between the settling time and relative precision of the first stage and/or additional hardware (comparators, logic) in the second stage can be found. Usually the span of the overrange and underrange levels should be between 25 and 50 percent of the whole range of the stage (i.e., $V_{28} - V_{20}$ in Figure 11).

In order to simplify the overrange detector (second stage) and the adder (first stage, Figure 10), it is possible to increase the different reference voltages, V_{Ri}, $i = 1, 2, 3, \ldots$ (see Figure 12), of the first stage so that they are higher than the corresponding output voltages, V_{Ci}, $i = 1, 2, 3, \ldots$, of the D/A converter; i.e.,

$$V_{Ri} = V_{Ci} + \frac{V_{ovr}}{2} \tag{8}$$

where V_{ovr} is the whole overrange and underrange span of the second stage as shown in Figure 12. With these values of reference voltage, V_{Ri}, the comparators can change their states only when the input voltage is higher than the corresponding voltage V_{C2}. If $V_{ovr}/2$ is chosen to be higher than all possible errors of the first-stage quantization process, the voltage $\Delta V = V_S - V_{Ci}$ can only be overrange, i.e., $\Delta V > V_{20}$, and if $\Delta V > V_{28}$, only a +1000 corrective signal can exist. This substantially simplifies the whole corrective feedback circuit.

Because speed was the main objective, the inputs of the second-stage comparator have been used to form the difference voltage, ΔV; i.e., the chain of reference voltage resistors has been connected in series with resistor R_4 and source Q_4 of current I_{CC} (Figure 3) of the D/A converter. The second-stage comparators therefore have a sliding reference voltage (Figure 10). This limits the precision of the whole converter to the value given in equation (7) (e.g., $n = 8$ for MC1650), but eliminates the voltage difference-forming passive or active network, DIF, with its attenuation, delay, and/or settling time.

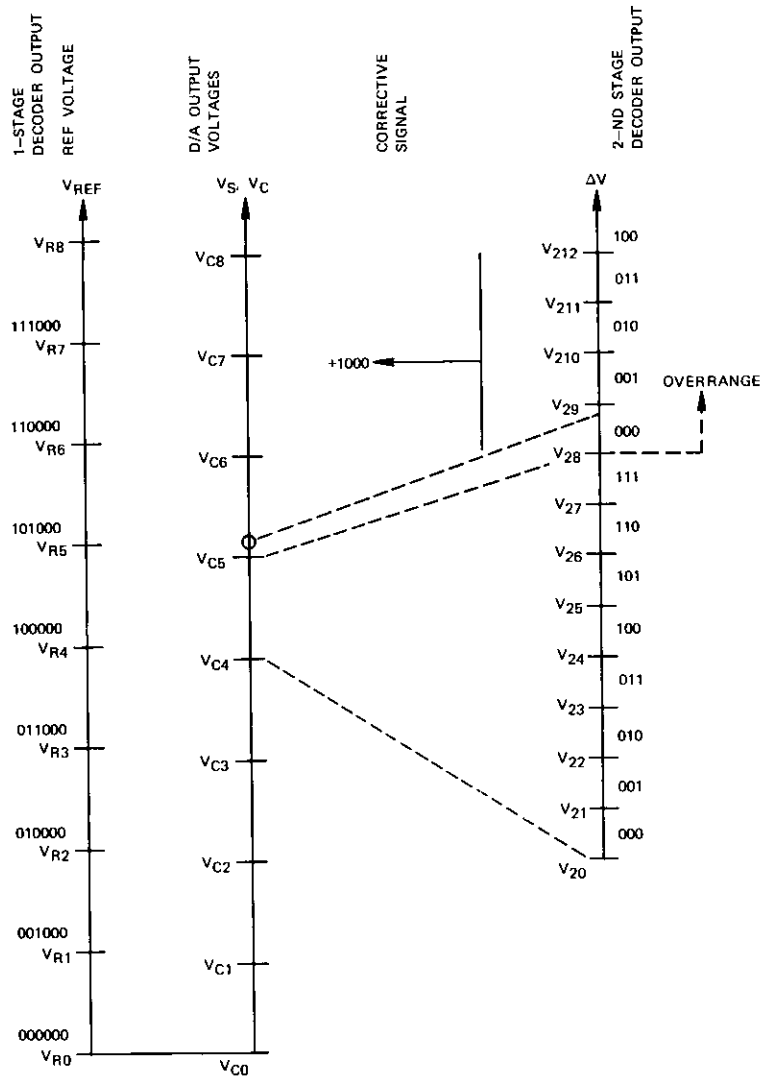


Figure 12. Two-Stage Quantizing Process with Corrective Signal and Different Reference, V_{Ri}, and D/A Output Voltage, V_{Ci}

Sample-and-Hold Circuit

Figure 13 is an equivalent circuit diagram of a sample-and-hold device. The switch, S, closes periodically (sampling period t_S) for a time t_C, during which capacitor C is charged to voltage V_S which has a given relationship to the input, $V_{IN}(t)$. Then the switch is opened and V_S is held for the time $t_S - t_C$ during which the A/D conversion is completed. There is always a certain error, t_e, in the timing of the opening of switch S. This error is caused mainly by the finite rise time of the opening signal.

If the input signal, $V_{IN}(t)$, is frequency limited within an equivalent bandwidth B, its "tangential rise time" [6], [10] is

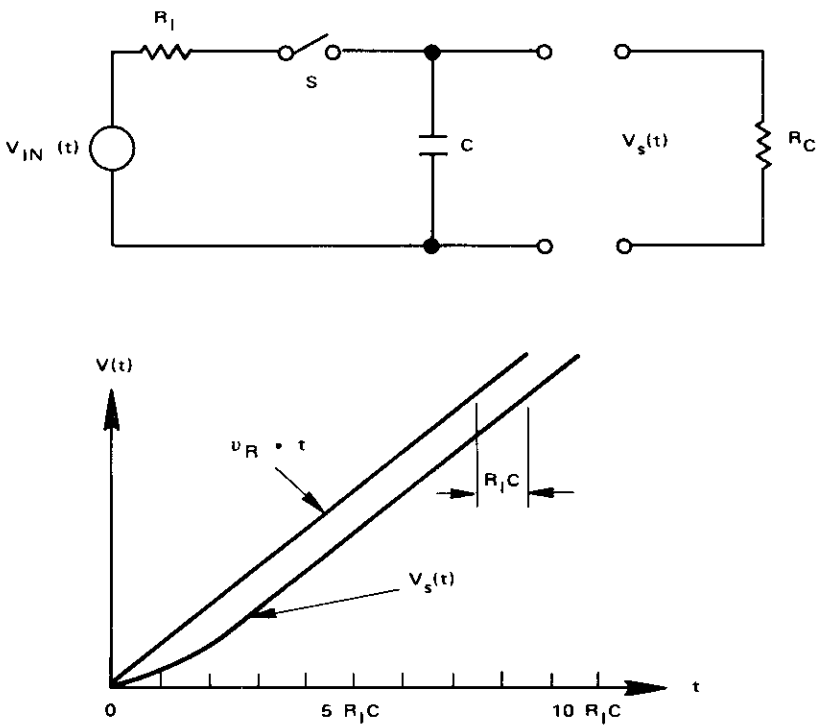


Figure 13. Equivalent Diagram of a Sample-and-Hold Circuit and Its Transient Response

$$t_{rt} = \frac{1}{2B} \tag{9}$$

which is the maximum possible rate of change of the input signal. If the maximum input signal is quantized to 2^n levels, a change of one level can occur within the time interval

$$\Delta t_{rt} = \frac{t_{rt}}{2^n} = \frac{1}{2^{n+1}\,B}\,. \tag{10}$$

To retain n-bit sampling precision, the time error, t_e, must be $t_r \leq t_{rt}/2$; i.e.,

$$t_e \leq \frac{t_{rt}}{2^{n+1}} = \frac{1}{2^{n+2}\,B}\,. \tag{11}$$

For $n = 8$ bits and $B = 5$ MHz, the timing of the switching of S should have a precision better than 0.2 ns; therefore the rise time of the switching signal must be less than 1 ns.

Because the sampling period, t_S, must be less than or equal to t_{rt} [6], [9], [11] and the closing time, t_C, must be much less than t_S, during time t_C it is possible to approximate $V_{IN}(t)$ by a linear ramp, $V_{IN}(t) = v_r t$, as shown in Figure 13. If the initial voltage, $V_S(t)$, is equal to zero for $t < 0$, and the switch is closed at $t = 0$, then during the period $0 < t < t_C$, the output voltage, $V_S(t)$, is

$$V_S(t) = v_r\left[t - R_I C\left(1 - \exp - \frac{t}{R_I C}\right)\right] \tag{12}$$

For $t \gg R_I C$, voltage $V_S(t)$ is a replica of $v_r t$ delayed by $t_D = R_I C$. The time-dependent exponential term should therefore be as small as possible at instant $t = t_C$ (opening of the switch) in order to make $V_S(t)$ reasonably independent of possible changes of sampling time, t_C. A good compromise between conflicting requirements is

$$\exp - \frac{t_C}{R_I C} = \frac{1}{2^{n-1}} \tag{13}$$

i.e.,

$$t_C = 0.7(n - 1)R_I C. \tag{14}$$

For $t_C = 15$ ns, n = 8 bits, and C (the memory capacitor) = 50 pF, R_I, the internal resistance of the signal source, must be less than 60 ohms.

During the time $t_S - t_C$, the memory capacitor, C, is discharged through the input resistance of the amplifier and the parasitic resistance of the open switch. At the end of the conversion cycle $t = t_S - t_C$, this relative error must be smaller than half of the quantization step; i.e., the time constant R_eC must be [12]

$$R_eC \geq 2^{n+1}(t_S - t_C). \tag{15}$$

For $t_S - t_C = 1$ μs and $C = 50.10^{-12}$, the minimum equivalent discharge resistance, R_e, is 10 MΩ.

Diode Switch

A switch capable of meeting the preceding criteria is the diode bridge using Schottky barrier diodes. The optimal operating conditions of the diode bridge switch shown in Figure 14 are as follows:

a. When nonconducting between points a-b, the diodes must be backward biased for every possible voltage difference $\Delta V_{ab} = V_{IN} - V_S$, but this bias ($V_{OF}$) must be as low as possible to limit the diodes' backward current.

b. In the conducting state, the diodes must be matched to have the same voltage drop when conducting the constant current $I_{ON}/2$. Under this condition, matching changes substantially less with temperature than with diodes biased in the conducting state while driven from a low internal impedance source.

Figure 15 is a simplified circuit diagram of the sample-and-hold circuit. The amplifiers AMP-1 and AMP-2 are the same voltage followers used in the D/A converter (see Figure 3). Their field-effect input transistor has a gate leakage current several orders of magnitude smaller than the diode switch backward current. The constant current pulse generator is a combination of the monostable multivibrator and two fast current switches with a rise and fall time less than 1 ns.

Each "sample command" pulse generates a $t_C = 15$-ns pulse at the output of gate C_2 which switches the circuit current, I_{ON}, from transistor Q_1 to Q_2. This current flows through the secondary winding and charges capacitors C_X to the voltage at which diodes D_{2i} and D_{4i} can conduct the whole current, I_{ON}. Because of the high collector impedance of transistors Q_1 and Q_2, the current I_{ON} flowing through the diode bridge remains reasonably constant during the conducting period, t_C, if the transformer is properly designed.

At the end of the sampling time, the current I_{ON} switches back to Q_1

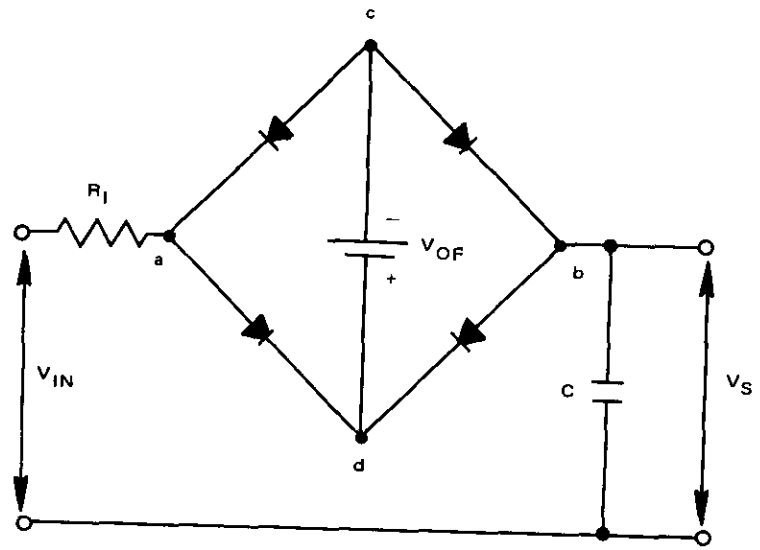


Figure 14a. Diode Bridge Switch in the Nonconducting State

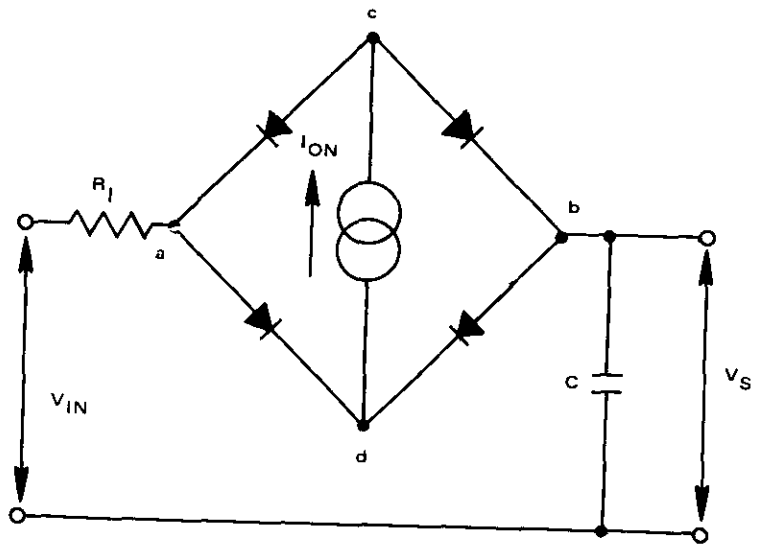


Figure 14b. Diode Bridge Switch in the Conducting State

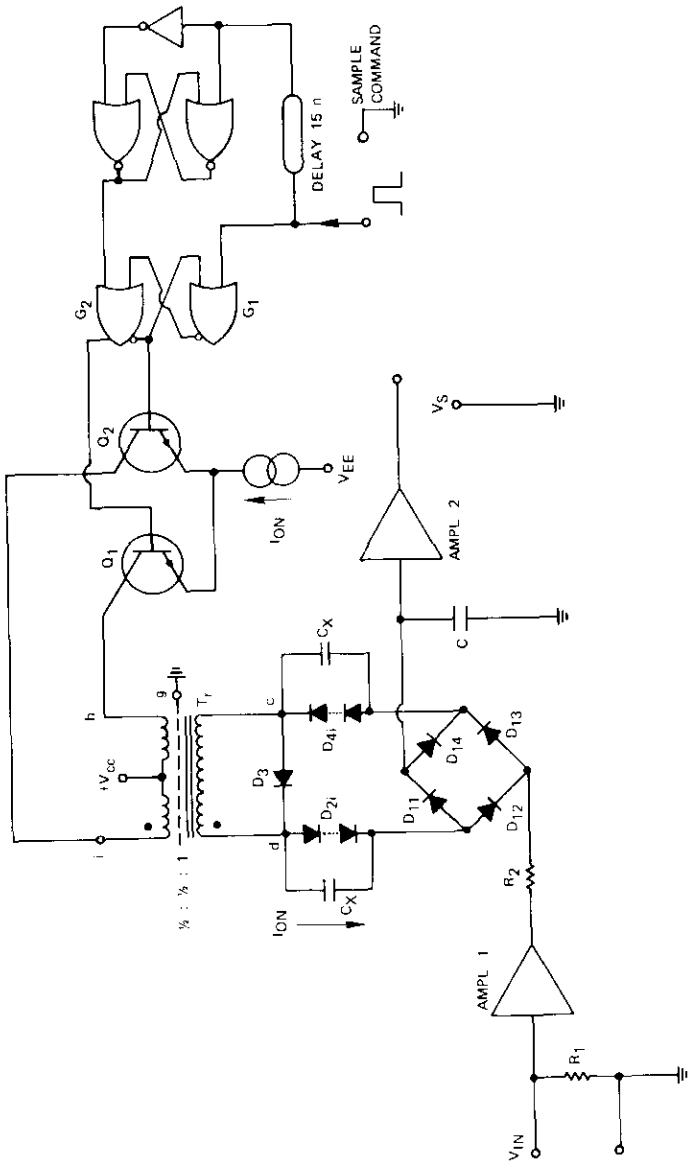


Figure 15. Simplified Diagram of the Sample-and-Hold Circuit with a Diode-Bridge Switch

and the polarity of the secondary voltage is reversed. Now diode D_3 is conducting and, during the OFF period, which is substantially longer than the ON period, t_C, the voltage at points a-b only changes from 700 mV (diode D_3 conducting the full current, I_{ON}) to 0 (i.e., the transient current in the secondary dies out). The voltage on C_X is relatively stable because of the nonlinear resistance of the diodes. The 50-mV voltage drop reduces the diode current one order of magnitude. The bias voltage, V_{OF}, is therefore nearly constant and independent in a wide range of sampling rates $1/t_S$. Hence, the current transformer must be designed to transmit the current pulse, I_{ON}, with a permissible droop (0.5 to 5 percent) and with the maximum width, t_C.

Figure 16 is an equivalent diagram for the ON period. For $t \ll L_M/R_D$, the exponential change of I_D can be approximated by a tangent [12]. Then, for the end of the ON period, i.e., $t = t_C$,

$$\frac{\Delta I}{I_{ON}} = \frac{I_{ON} - I_D}{I_{ON}} = \frac{R_D}{L_M} t_C \tag{16}$$

where L_M is the inductance of the secondary of Tr, and $R_D = 0.025\ d/I_D$ is the dynamic resistance of the diode string containing d diodes. Because $I_D \rightarrow I_{ON}$, equation (16) can be simplified and solved for L_M as follows:

$$L_M = \frac{0.025\ d}{\Delta I} t_C.$$

For $t_C \leq 50$ ns, $d = 10$, and $\Delta I = 0.25$ mA, the value of L_M is 50 μH.

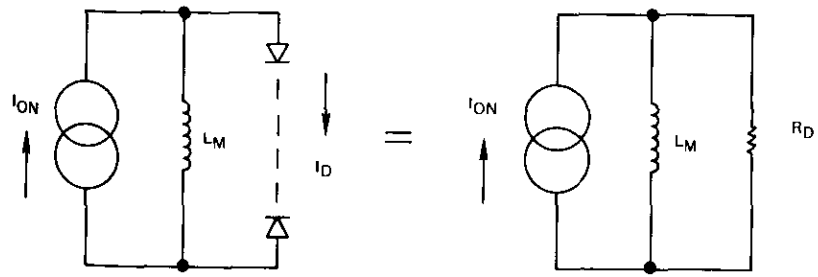


Figure 16. Equivalent Diagram of the Transformer During the Period When the Diode Switch is Conducting

For a rise time less than 1 ns, the stray inductance L_S of the transformer must be less than 5×10^{-9} H and the ratio L_M/L_S must be greater than 10^4.

The parasitic capacitive coupling between primary and secondary must be kept to an absolute minimum. For a 5-V pulse with a 1-ns rise time, a 1-mA current can flow through 0.2-pF capacitance and unbalance the diode bridge.

To fulfill the previous design criteria, a toroid transformer using shielded coaxial cable ("triaxial") for winding was designed [13]. The inner conductor is used as a secondary, the outside shield as a primary, and the internal braid as a shield between both windings. The concentricity of the windings reduces the stray inductance to a value which cannot be measured directly. Connection of the outside and inside braid and the external shield according to Figure 17 reduces the capacitive coupling to a value which does not influence the stability of the diode bridge balance.

Measurement

Signal-to-noise ratio measurements, often used to evaluate A/D and D/A converters, have several limitations, especially with sampling rates

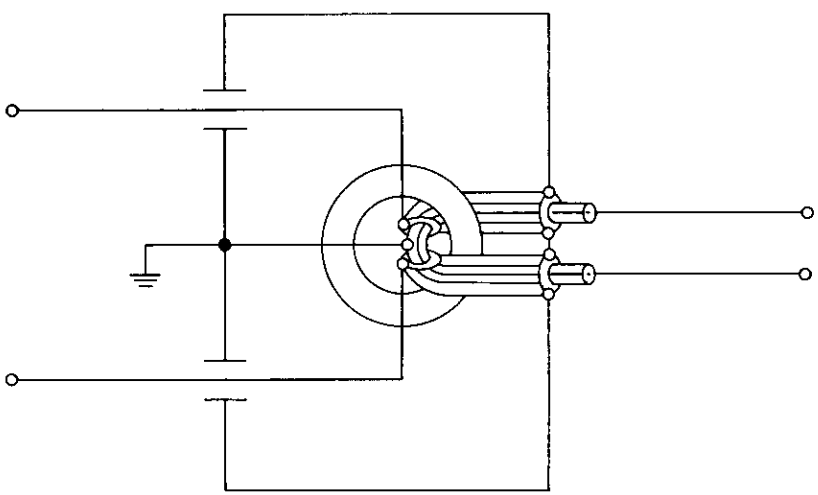


Figure 17. Coupling Transformer with Triaxial Winding and External Shield

over 1 MHz. They are also unsuitable for the diagnosis and detection of sources of errors, such as nonlinearity, switching transients, hysteresis, insufficient sensitivity of comparators, DC level shifts, and defects in logic.

More precise information about A/D converter behavior can be obtained from static and pseudodynamic measurements of the transfer characteristics (see Figures 18 and 19). In the static test (Figure 18), the A/D converter command input is connected to the pulse generator running at the sampling rate. Display voltmeters are connected to the outputs. The input DC voltage is measured by a digital voltmeter with a precision better than 1 mV. In this manner a direct and precise relationship between the input voltage and the output binary number is obtained. When the input voltage is slowly changed, the display voltmeter deflections, which relate to an average value of the two possible output states (0, 1), determine the width of the transition region from one level to another.

If the sampling switch were ideal, the circuit in Figure 18 could also be used for a pseudodynamic test, i.e., a test in which only a single command pulse is applied. However, since the memory capacitor, C, is charged through the leakage resistance R_C of the diode bridge to the input voltage

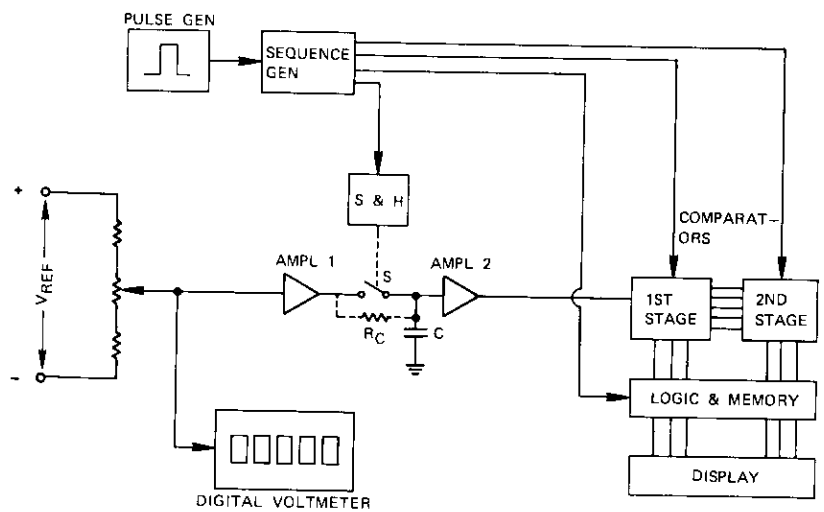


Figure 18. Static Measurement of the Analog Input-Digital Output Transfer Characteristic of the A/D Converter

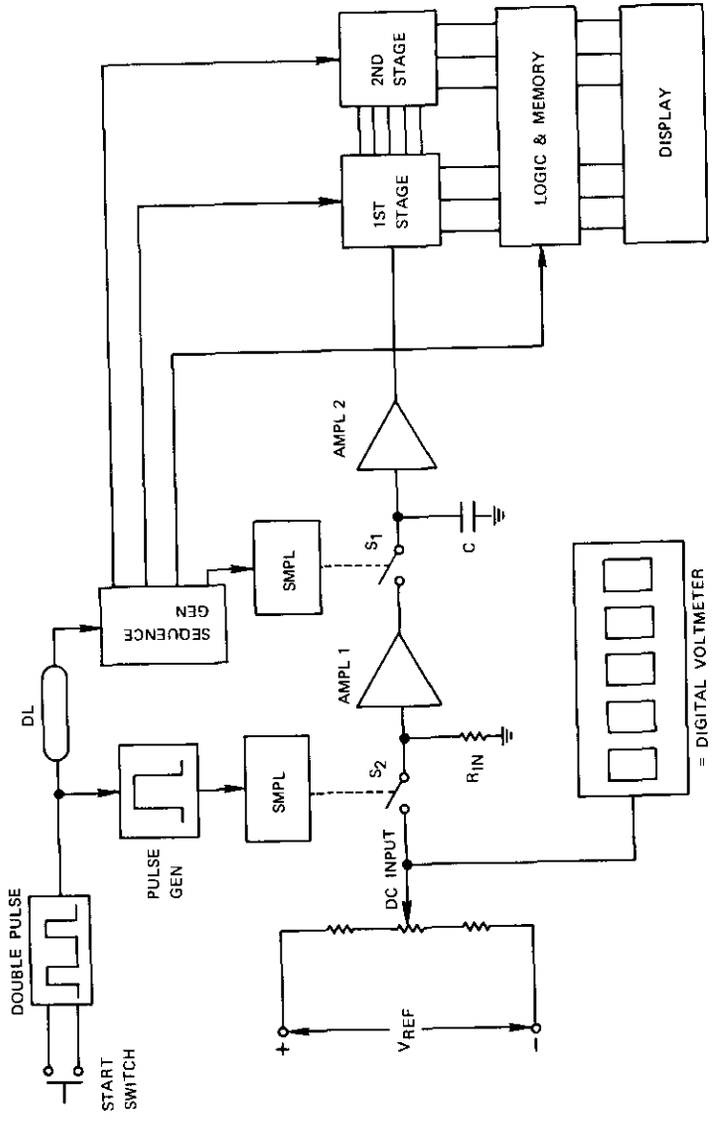


Figure 19. Pseudodynamic Method of Measuring the Transfer Characteristic

in a few milliseconds, i.e., before the single pulse can be manually applied, the circuit of Figure 18 cannot be used for pseudodynamic testing.

To measure the input/output relationship under actual dynamic conditions, i.e., under the influence of all sampling and switching transients, the circuit must be modified as shown in Figure 19. Another switch, S_2, is inserted between the input DC voltage source and the preamplifier AMP-1. The input resistance R_{IN} is chosen so that neither the leakage current of S_2 nor its internal resistance (when conducting) produces a measurable error.

Initially S_2 is open and therefore the output of AMP-1 is at zero voltage with respect to ground. The memory capacitor is charged (or discharged) to the source voltage through leakage resistor R_C. When the start switch is initiated, two pulses are generated and the measuring sequence is as follows:

a. The first pulse closes switch S_2 and holds it closed.

b. After a suitable delay provided by the delay line *DL* during which all transients die out, the first pulse initiates the sampling and the conversion cycle.

c. The second pulse then opens S_2, transfers the output digits to the memory, and displays.

In this setup, the timing of the conversion sequence is the same as under actual dynamic conditions and the voltage at C changes for each conversion cycle between zero and the specific DC input voltage. The differences between static and pseudodynamic transfer characteristics make it possible to identify the influence of frequency limits and/or transients of the different building blocks.

Quantization of the analog signal introduces an inherent error (quantization noise) in the amplitude samples. The mean signal-to-noise power ratio, S/N, depends on the number of quantization levels, 2^n. For $n \geq 3$,

$$S/N = 6n + a\,[\text{dB}] \tag{18}$$

where a is a constant whose value is given as 3 dB in Bennett's original work [14], 1.8 dB in other work by Bennett [15], and 0 according to other sources [6], [16]. If a sine-wave input signal is used for measuring S/N, its frequency, f_a, must be sufficiently low so that, during the sampling period, t_S, the instantaneous value does not change more than one quantization step. Otherwise additional noise is generated [14]. This condition leads to

$$f_a \leq \frac{1}{2^n \pi t_S}. \tag{19}$$

The amplitude of the signal must be between 90 and 100 percent of the whole range of the converter to use the maximum number of levels, but it must be carefully adjusted not to overload the device, since even one quantization level overload would cause additional distortion.

From the foregoing discussion and from Figure 2, the substantial limitations of the S/N measuring methods are apparent:

a. The input voltage changes no more than one quantization step between successive samples (as in the static measurements according to Figure 18). This test is therefore unable to detect errors caused by the finite settling times and switching transients in the system.

b. Small changes (less than one quantization step) in the amplitude of input sine wave f_a and/or a small DC offset can change the amplitudes of the second and third harmonics of f_a (which contribute most of the "noise" power) by more than 6 dB and can change the S/N ratio by more than 3 dB.

c. The measured S/N ratio also includes the effects of the D/A converter (Figure 20). If the D/A converter has a theoretical precision of one-half of the LSB, then its S/N is only 6 dB better than that of the A/D converter with the same number of levels (2^n). Therefore, the D/A conversion reduces the S/N of the whole system in Figure 20 by at least 1 dB. In this case, a realistic formula for S/N seems to be

$$S/N = 6n \text{ [dB]} \tag{20}$$

For fast converters with $n \geq 6$, it does not seem practical to use white noise as a test signal as suggested by Bennett [14], [17]. The rms amplitude of the noise must be at least 12 dB below the maximum amplitude, which equals 2^n quantization levels, and the frequency band "cutoff" from the input noise spectrum cannot be wider than approximately 10 percent of the full bandwidth of the A/D-D/A system. Therefore, the measured noise signal power is at least another 22 dB below the value given by equation (20). Hence, for an eight-bit system, the noise power measured in the "cutoff" band is more than 70 dB below the maximum signal power, and the measurements, which are influenced by different noise sources in the system, are very inaccurate and difficult to perform, especially at frequencies of several MHz [18].

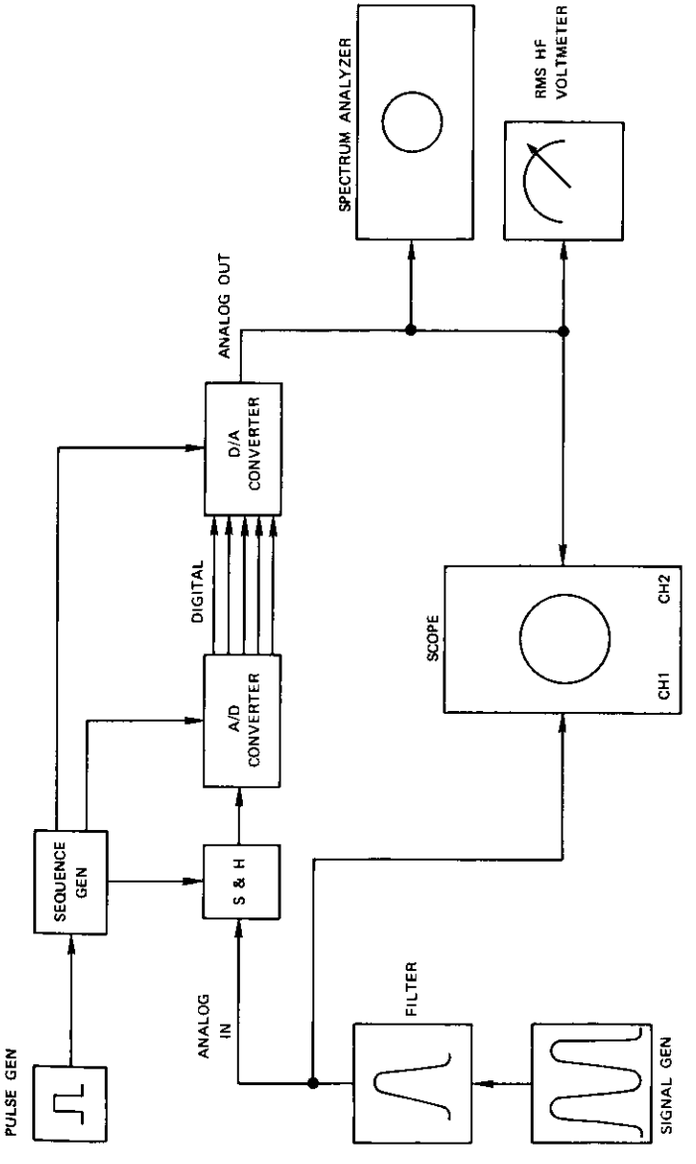


Figure 20. S/N Ratio Measurement of the A/D and D/A System. The filter between the signal generator and the analog input of the converter increases the purity of the sinusoid input signal.

In our eight-bit system, 20 mV represents one quantization step, so that the full range of the first-stage input signal is ±2560 mV. The static measurements (Figure 18) showed that the maximum deviation from linearity in the whole range was equivalent to 8 mV, mainly in the vicinity of the zero-volt input where the transition from 01 to 10 (the two MSB) occurred. When measured pseudodynamically (Figure 19), this deviation from linearity was slightly greater than 10 mV for t_S = 100 ns and approximately 12.5 mV for t_S = 60 ns. The S/N ratio measured with a sine-wave input signal, f_a = 5 kHz, and an amplitude of 5 V_{pp}, varied between 44 dB and 46 dB in accordance with small changes in amplitude and/or DC offset. The performance of a seven-bit system can be judged from Figure 21, where a direct comparison of analog input and reconverted output signal is made. The width of the trace is less than 1/128 of the maximum input amplitude.

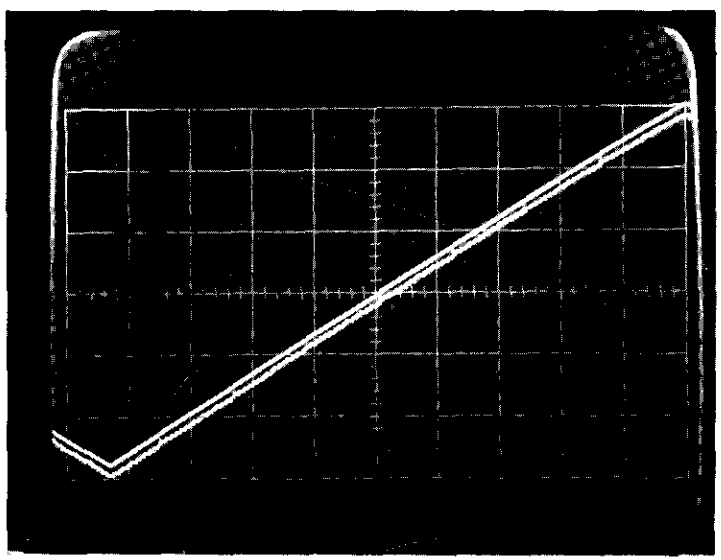


Figure 21. Triangular Waveform Test for the Seven-Bit Converter System (input top, output bottom; scales: vertical 0.2 V per division; horizontal 1 μs per division)

CONCLUSIONS

A fast and high-definition A/D and D/A conversion system with a bit rate of more than 150 Mbps was built using standard IC and double-sided printed circuit cards. This system is modular, and because it employs only five different building blocks, it is economical to use and easy to service.

A new direct method of measuring the converter properties was also developed. This method overcomes most of the limitations of indirect S/N measurements with sine-wave or white noise signals. The converters, which were tested as a part of a digital color TV transmission system, have proven their accuracy, dynamic properties, and operational stability.

REFERENCES

[1] *Notes on Analog-Digital Conversion Techniques*, edited by A. K. Susskind, Cambridge: Technology Press, M.I.T., 1957.

[2] G. A. Korn and T. M. Korn, *Electronic Analog and Hybrid Computers*, New York: McGraw-Hill, 1964.

[3] H. Schmid, "An Electronic Design Practical Guide to D/A Conversion," *Electronic Design*, Vol. 16, No. 22, October 24, 1968, pp. 49–88.

[4] H. Schmid, "Electronic Analog Switches," *Electro-Technology*, Vol. 82, No. 6, June 1968, pp. 35–50.

[5] *Switching Transistor Handbook*, second edition, Phoenix: Motorola, Inc., 1963, pp. 75–89.

[6] M. Schwartz, *Information Transmission, Modulation, and Noise*, New York: McGraw-Hill, 1959, pp. 159–194.

[7] J. S. Mayo, "Experimental 224 Mb/s PCM Terminals," *Bell System Technical Journal*, Vol. 44, No. 9, November 1965, pp. 1813–1814.

[8] "Pulse Code Modulation of Video Signals: 8-Bit Coder and Decoder," Research Dept. Report No. 1970/25, British Broadcasting Corp., July 1970.

[9] G. Hanke, "Ein Experimentalles 80 Mbit/sec-PCM-System in Integrierter Schaltungstechnik," *Technischer Bericht des Forschungsinstituts der FTZ* (West Germany), No. 442, TBR 14, February 1969.

[10] H. Wolf, "Uber den Zusammenhang Zwischen Bandbreite und Anstiegszeit," *Elektronik* (West Germany), Vol. 12, No. 10, October 1963, pp. 303–308.

[11] D. M. Tutelman, "Properties of the Sampling Theorem for Time-Division Switching," Report No. MM-64-5321-4, Bell Telephone Laboratories, Inc., Research Dept., May 1, 1964.

[12] O. A. Horna, "Fundamentals of Pulse Technique," *Sdělovací Technika* (Czechosiovakia), Vol. 14, No. 10, October 1967, pp. 364–369.

[13] O. A. Horna, "HF Transformer with Triaxial Cable Shielding Against Capacitive Currents," *IEEE Transactions on Parts, Hybrids and Packaging*, PHP-7, No. 3, September 1971, pp. 109-114.
[14] W. R. Bennett, "Spectra of Quantized Signals," *Bell System Technical Journal*, Vol. 27, No. 3, July 1948, pp. 446-472.
[15] Members of the Technical Staff, Bell Telephone Laboratories, Inc., *Transmission Systems for Communications*, fourth edition, Winston-Salem: Western Electric Co., Inc., 1970, pp. 571-574.
[16] *Reference Data for Radio Engineers* [ITT], New York: H. W. Sams & Co., 1968, pp. 21.20-21.22.
[17] W. R. Bennett, "Cross Modulation Requirements on Multichannel Amplifiers Below Overload," *Bell System Technical Journal*, Vol. 19, No. 44, October 1940, pp. 587-610.
[18] A. K. Bhushan, "Efficient Transmission and Coding of Color Pictures," thesis submitted in partial fulfillment of requirements for M.S. at M.I.T., Boston, June 1967, 92 pages.

Otakar A. Horna has worked in the field of very high speed logic circuits and analog to digital conversion techniques since he joined COMSAT Laboratories in 1969. His education includes an M.S. in electrical engineering, Ph.D. in electronics and a diploma in mathematical logic. He is presently a member of the technical staff of the Switching and Multiplexing Techniques branch.

Index: communication satellites, computerized simulation, mathematical models, parametric equations, systems engineering

A technique for modeling communications satellites

J. D. KIESLING, B. R. ELBERT, W. B. GARNER, AND W. L. MORGAN

The modeling technique presented in this paper has proved useful for planning future high-performance and cost-effective communications satellites. Dividing the satellite into subsystems makes it possible to formulate parametric relationships among satellite power, mass, and cost. These relationships reflect various classes of existing designs and reasonable extrapolations. The communications parameters permit selection of traffic capacity, earth station G/T, and types of multiple access, frequency plan, modulation, and multiplex, while the spacecraft parameters permit selection of propulsion, stabilization, and power generation methods. The parametric relationships can be programmed as computer subroutines in a complete system model to compare design alternatives and test sensitivity of results to assumptions.

INTRODUCTION

For planning communications satellite systems, a method of efficiently examining technological alternatives is highly desirable. System characteristics such as traffic capacity, earth station G/T, type of multiple access, frequency plan, modulation, and multiplex should be related to the performance and cost of the system. In this paper, a model of the space segment of a representative communications satellite system is presented.

Because of the high cost of launch vehicles, spacecraft require elaborate

quality control and testing, expensive piece parts, and large design margins. These requirements result in a corresponding increase in spacecraft cost. Since satellite mass becomes a critical factor, the projected satellite mass, determined from the input requirements of capacity and traffic flow and the constraints on system design and operating variables, is one output measure in the present study. It is also helpful to relate mass to cost; while this may appear to be an oversimplification, this approach has proved useful for system and technology planning and should continue to prove useful in the future. Major technology breakthroughs (such as greatly reduced launch costs) will obviously necessitate revision of the model.

This paper is limited to space-segment modeling. Consequently, the relative costs of the space and earth segments, overall reliability, and flexibility are not provided. Corresponding models for the earth segment would be necessary to complete a system model [1], [2]. Only satellite configurations and subsystems applicable to commercial services are considered here [3].

The satellite is divided into two major parts: the spacecraft bus and the communications payload. Hence, if the total satellite mass* is W_S, then

$$W_S = W_B + W_p \tag{1}$$

where W_B = mass of the bus

W_p = mass of the payload.

The bus contains the mechanical structure; positioning and orientation (P&O); thermal control; and tracking, telemetry, and command (TT&C) elements. The payload, devoted to the commercial communications mission, is composed of the transponder, antenna, and power subsystems.

The ratio of the payload mass to the total satellite mass, called the *utilization factor*, is

$$u = \frac{W_p}{W_S}. \tag{2}$$

Typically, u is in the range of 0.35 to 0.6. For a multipurpose satellite, W_p is apportioned among the various subpayloads; for example, one

* This is the total satellite mass (including burned out apogee motor if used) at beginning of life in the geostationary orbit.

satellite may contain separate payloads, each optimized for a particular class of earth station user, i.e., for the 4-/6-GHz and 11-/14-GHz bands. If $(W_p)_i$ is the mass of one such payload, then

$$W_S = \frac{1}{u} \sum_i (W_p)_i. \tag{3}$$

There are two basic satellite stabilization methods: spin stabilization and body stabilization. In spin stabilization, a major portion of the spacecraft mass rotates about an axis which is parallel to the earth's spin axis. In body stabilization, the spacecraft body is not spun; instead, devices such as an internal "momentum" wheel are used to maintain stability. These two types of stabilization lead to different tradeoff characteristics.

SPACECRAFT BUS

Positioning and Orientation

The positioning and orientation element may be divided into two parts. The first, called the dry mass, consists of sensors, tanks, plumbing, thrusters, control system, and a stabilizing device. The dry mass is compared for spin- and body-stabilized satellites in Figure 1 [4], [5], [6], [7]. The second part is the fuel for station acquisition, station changing, and stationkeeping. Traditional propellants for communications satellites are hydrogen peroxide and catalytic hydrazine. Recent advances make ion propulsion an attractive alternative [8]. In Figure 2, the mass of these three fuels is compared for various missions. It should be noted that experimental and aeronautical satellites and some communications satellites require only east-west (longitudinal) stationkeeping, while most operational communications satellites require east-west and north-south (inclinational) stationkeeping.

Structure

The mass allocated for structure is influenced by two factors: the presence or absence of an apogee motor and ancillary support, and the stabilization technique. The percentage of the total satellite mass devoted to the empty apogee motor case is given as a function of the total satellite mass in Figure 3. The relative mass of devices required for balancing and integrating is usually greater for spin-stabilized spacecraft than for body-

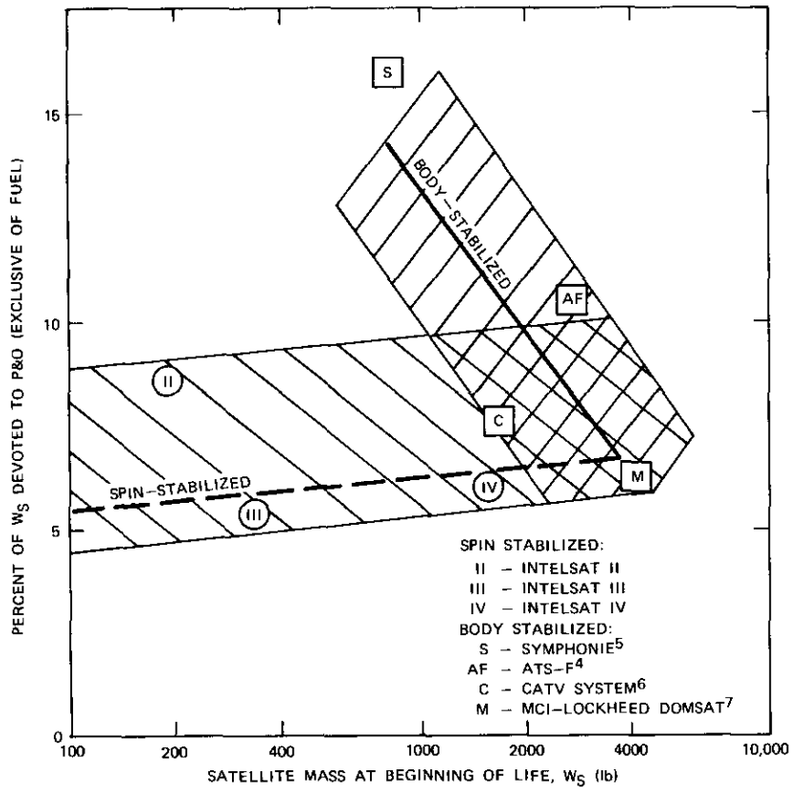


Figure 1. Percentage of the Satellite Mass at Beginning of Life Devoted to Dry Positioning and Orientation Element (equipped for full 7 years of north/south and east/west stationkeeping)

stabilized spacecraft. The structural element characteristics are compared in Figure 4.

Tracking, Telemetry, and Command

The tracking, telemetry, and command element is not considered to be part of the communications payload because its primary function is house-

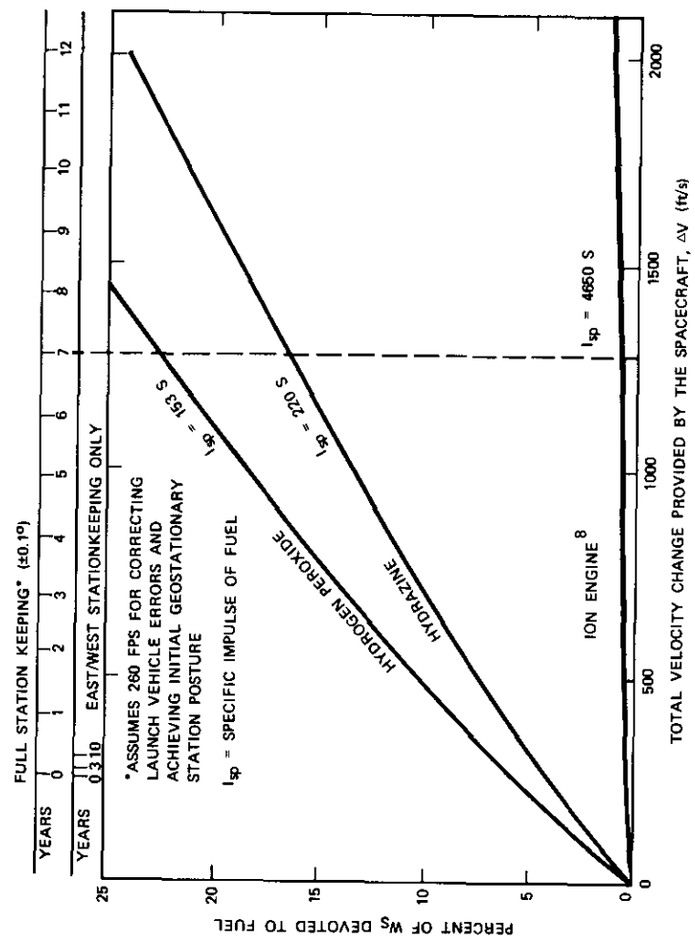


Figure 2. Fuel Mass Requirements for Various Missions

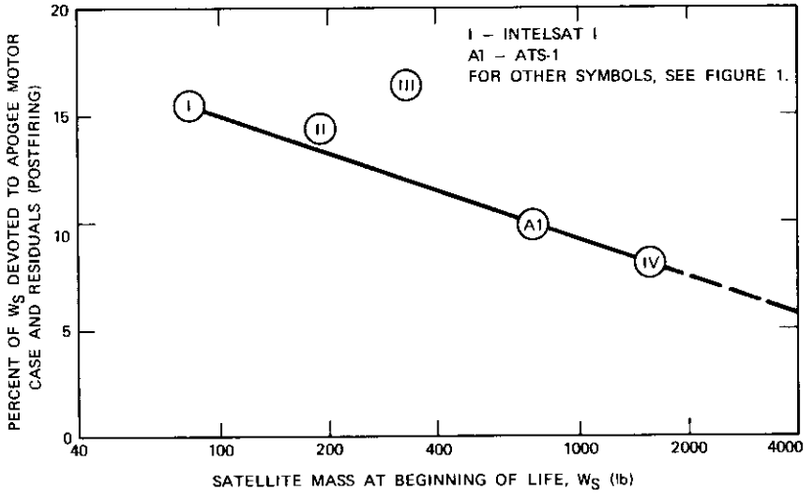


Figure 3. Percentage of the Satellite Mass at Beginning of Life Devoted to Burned Out Apogee Motor

keeping. Although the number of telemetry and command channels tends to increase with satellite mass, items with fixed weights (beacon, receiver, antenna, and digital coding and decoding) predominate. TT&C mass is typically about two to five percent of the total satellite mass.

Combined Spacecraft Bus Performance

Values for u are given in Figure 5 for a typical communications satellite with seven years of north-south stationkeeping capability. An efficient spacecraft bus is one in which the utilization factor is maximized to provide the maximum mass for the communications payload. Ideally, power is limited only by mass,* and the mechanical system is flexible enough to accommodate a wide range of missions by modifying only the communications payload.

* Present-day spin-stabilized satellites are limited in power capability by the dimensions of the launcher fairing.

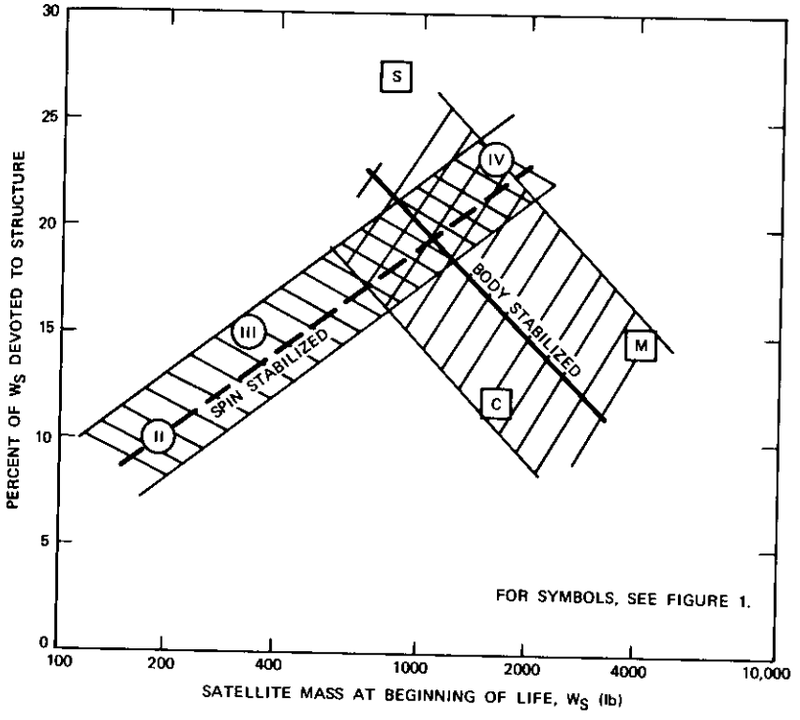


Figure 4. Percentage of the Satellite Mass at Beginning of Life Devoted to Structure

COMMUNICATIONS PAYLOAD

Figure 6 is a block diagram of the communications payload. The particular configuration and the relative distribution of mass among the subsystems are functions of the following items:

a. traffic capacity and distribution,
b. multiple-access method,
c. frequency plan,
d. modulation method,
e. multiplexing technique, and
f. earth station G/T.

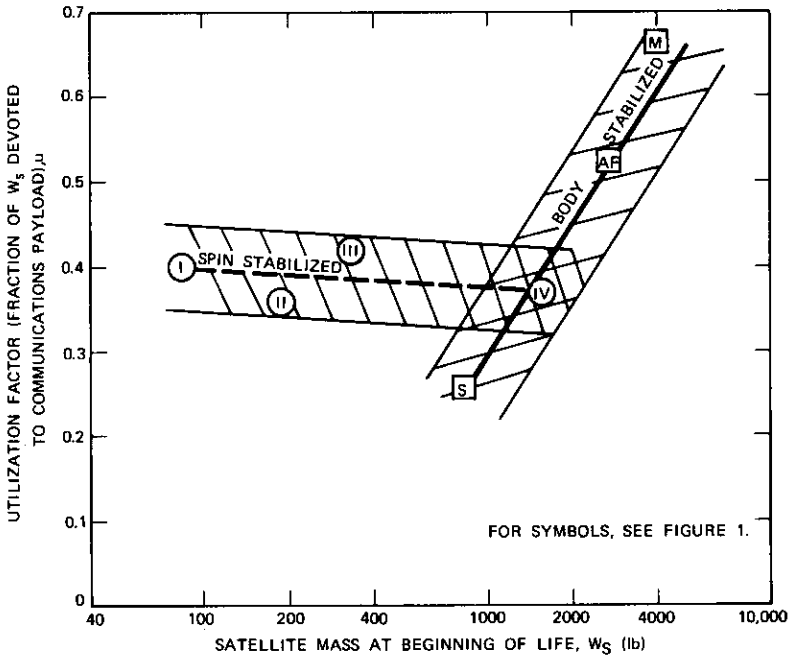


Figure 5. Utilization Factor for Geostationary Satellites

Power Subsystem

The power subsystem performs two major functions: power generation and energy storage. Figure 7 presents solar-array performance data based upon actual and proposed satellites. For oriented solar arrays, the specific power also includes the mass of the deployment and orientation mechanisms. Because of its geometry, an oriented solar panel theoretically has π times the power output of a drum; in practice, however, the improvement factor is less than π.

During eclipse, subsystems must be supplied by an energy storage device. For nickel-cadmium (Ni-Cd) batteries, the specific energy is of the order of 6 Whr/lb at rated depth of discharge. Fuel cells presently under development may achieve 20 Whr/lb.

Figure 8 shows a typical power subsystem with control and regulating electronics. The load which must be provided by the storage element

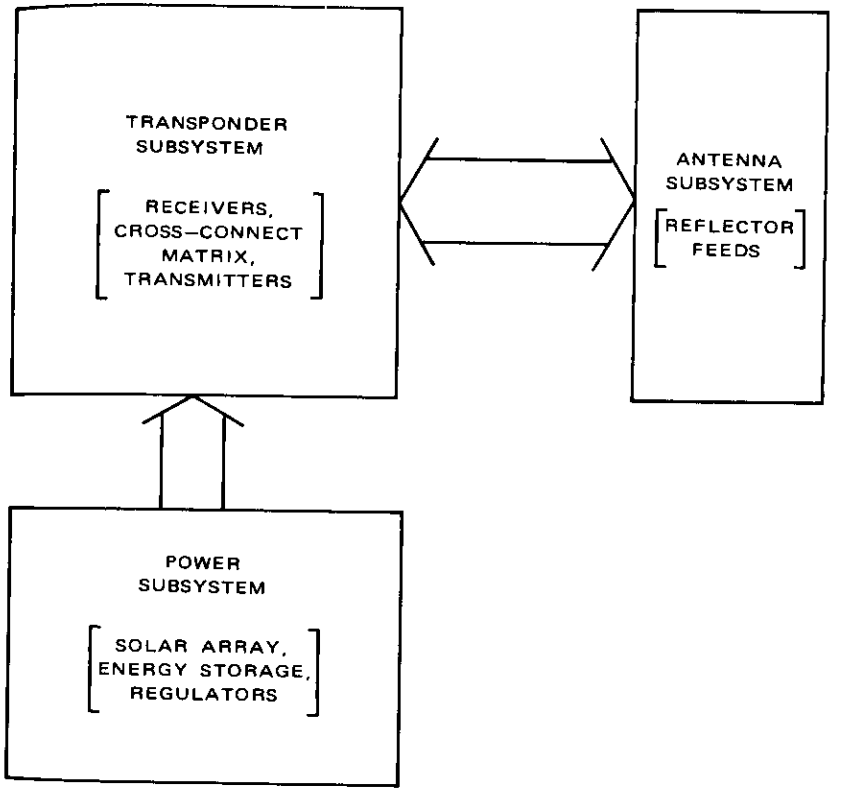


Figure 6. Communications Payload

during eclipse is

$$P_E = \frac{KP_T + P_B}{\eta_b \eta_r} \tag{4}$$

where
- K = eclipse factor, which is less than or equal to one, for partial or full loads during eclipse, respectively
- P_T = transponder subsystem power load (noneclipse)
- P_B = spacecraft bus power load (housekeeping)
- η_b = boost regulator efficiency
- η_r = storage regulator efficiency.

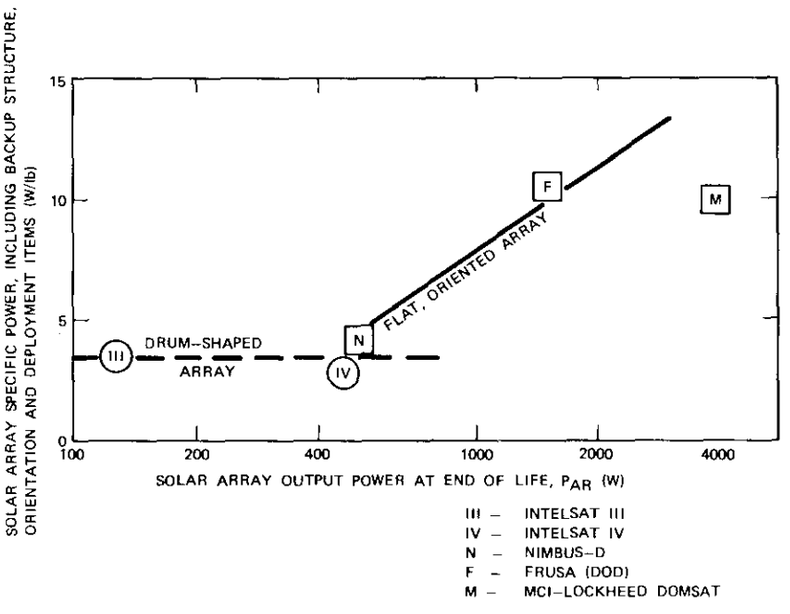


Figure 7. Solar Array Specific Power vs. Total DC Power Output (end of life)

Because P_H is a small percentage, it can be neglected. The storage element must be recharged by the solar array. For a maximum eclipse duration of approximately 1.2 hours, the storage element charging load (over a 20-hour period) is

$$P_{ES} = \frac{P_E}{20\eta_r\eta_c} \tag{5}$$

where η_r = charging efficiency
η_c = charge controller efficiency.

Combining equations (4) and (5) results in the total DC array power load,

$$P_{AR} = \frac{1}{\eta_m}\left[P_T + \frac{KP_T}{20\eta_b\eta_r\eta_c\eta_c}\right] \tag{6}$$

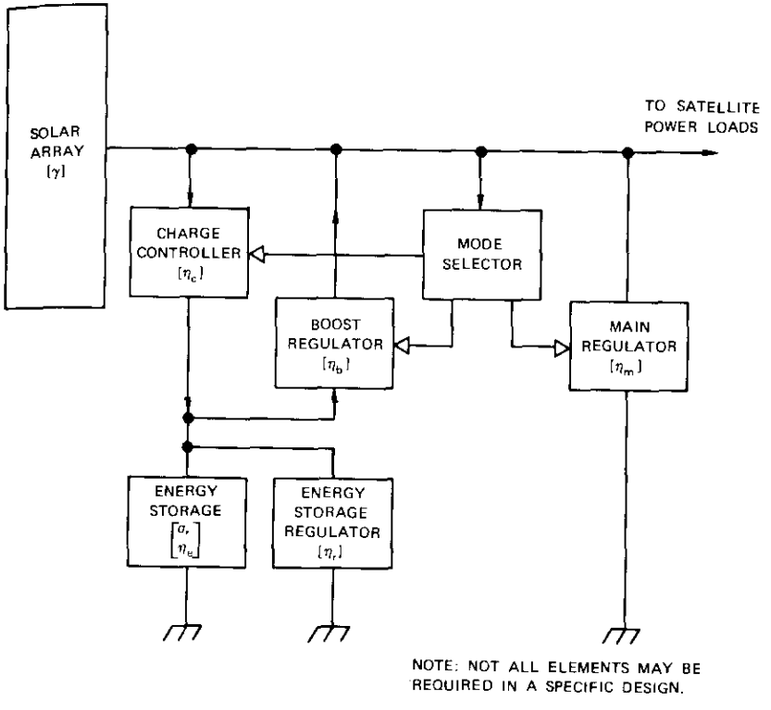


Figure 8. Power Subsystem Block Diagram

where η_m is the main regulator efficiency. The mass of the array is

$$W_{AR} = \frac{P_{AR}}{\gamma} \tag{7}$$

and the mass of the energy storage element is

$$W_{ES} = \frac{1.2\,P_E}{\sigma} \tag{8}$$

where γ = solar array specific power (Figure 7)
σ = storage element specific energy.

Transponder Subsystem

The transponder subsystem block diagram (Figure 9) assumes that frequency reuse at the satellite is provided through multiple independent spot beams and/or orthogonal polarizations [9]. Inputs to operate receivers from X different "beams" are shown. Since the full allocated bandwidth is available to each beam, the total satellite bandwidth can be very large. Links among the beams are efficiently established by a cross-connect matrix. Note that the transmitters may outnumber the receivers.

The receiver performs the functions of amplification and frequency translation over a relatively wide bandwidth. Figure 10a is a typical block diagram.

Unlike the receiver, the transmitter is highly performance-dependent; i.e., the RF output power is a direct function of the down-link require-

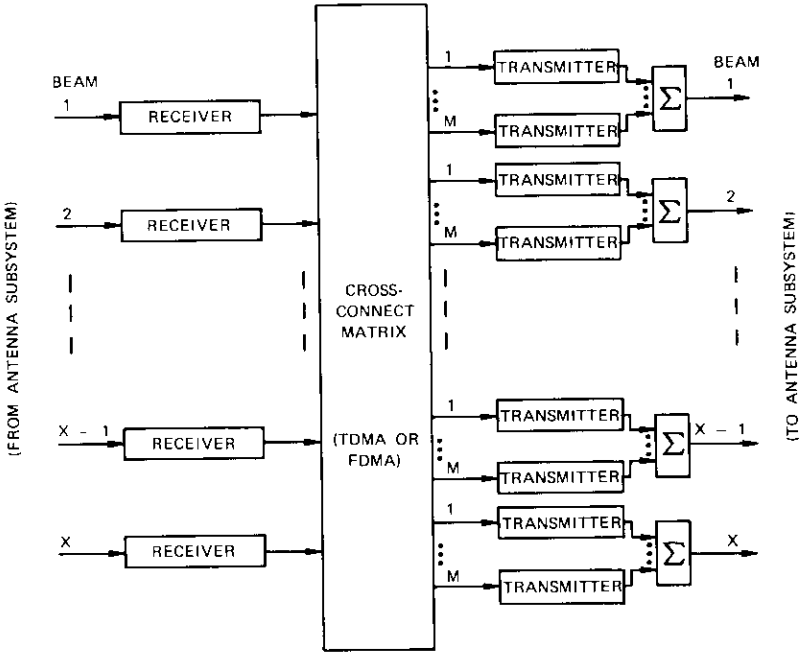


Figure 9. Transponder Subsystem Block Diagram for a Payload with X Spot Beams and M Transmitters per Beam

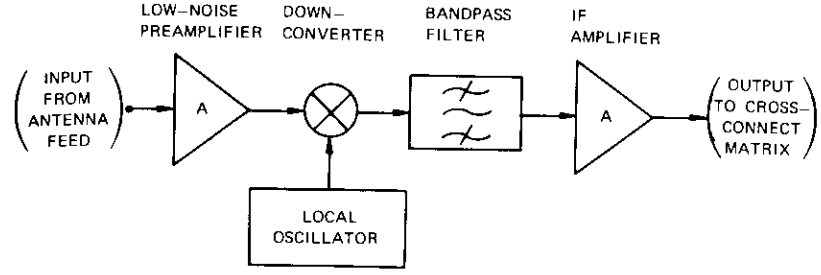


Figure 10a. Typical Receiver Block Diagram using Low-Noise Preamplifier

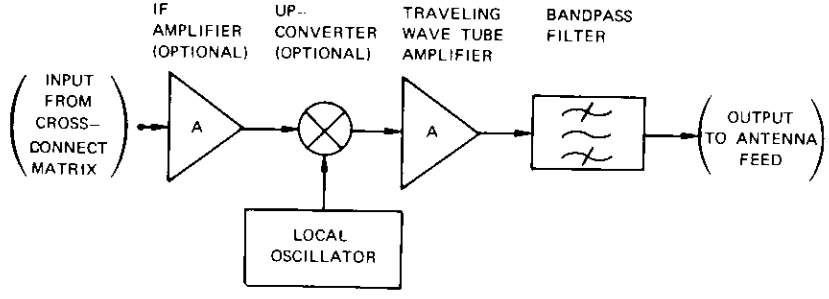


Figure 10b. Typical Transmitter Block Diagram Using Traveling Wave Tube Amplifier

ments. Because of their power, gain, and wideband characteristics, traveling wave tube output amplifiers (TWTAs), as shown in Figure 10b, are employed.

Since the TWTA has a nonlinear amplification characteristic, the requirement for low intermodulation distortion makes it necessary for the operating point to be "backed off" from saturation, significantly reducing the DC-to-RF power efficiency [10]. This can occur when the TWTA amplifies two or more carriers simultaneously. The following linear rela-

tionship can be used to obtain an approximation of the total DC input power from the RF output power:

$$P_t = P_{to} + \beta_p P_o. \tag{9}$$

Similarly, the total TWTA mass can be expressed as

$$W_t = W_{to} + \beta_w P_o \tag{10}$$

where P_{to}, β_p, W_{to}, and β_w are constants and P_o is the TWTA RF output power at saturation.

The cross-connect matrix (Figure 9) can be implemented in the time domain using time-division multiple access (TDMA) or in the frequency domain using frequency-division multiple access (FDMA). If any earth station can communicate with all other earth stations, then complete connectivity is achieved and there must be a total of X^2 paths. This configuration is simplified to model significant features; however, it can still represent the two general classes of multiple-access systems.

For FDMA, each earth station must transmit an individual carrier for each beam destination. The cross-connect matrix then contains an array of X^2 filters for separating the individual carriers and another array of X^2 filters (or power combiners) for combining the carriers at the appropriate beams. Connections among the appropriate channel filters are also required. The traffic pattern can be rearranged by making the connections with a cross-bar switch. Additional frequency converters are needed if the switching is done at a common frequency.

For TDMA, discrete time slots within a reference time frame are assigned for each beam destination [11]. The cross-connect matrix contains a matrix of X^2 switch elements which connect the appropriate receivers to the appropriate transmitters for the required time interval. The matrix may be controlled by a lightweight onboard computer so that the time intervals can be rearranged to match the instantaneous traffic demands.

The total transponder subsystem mass is

$$W_T = XW_r(1 + s) + \phi_w X^2 W_m + XM \cdot (W_{to} + \beta_w P_o)(1 + s) \tag{11}$$

where X = number of beams
W_r = receiver mass

s = spares ratio = $\dfrac{\text{number of spare elements}}{\text{number of active elements}}$
ϕ_w = 1 for TDMA
= 2 for FDMA
W_m = cross-connect matrix element mass
M = number of TWTAs per beam.

The mass of ancillary components, such as regulators, connecting cables, and switches, is added to that of the appropriate transponder element. The total transponder power is

$$P_T = XP_r + XM(P_{to} + \beta_p P_o) + \phi_p X P_m \tag{12}$$

where P_r = receiver DC input power
ϕ_p = 1 for TDMA
= 0 for FDMA
P_m = cross-connect matrix element power.

Antenna Subsystem

The antenna illustrated in Figure 11 is composed of a spherical reflector with a feed array capable of generating multiple spot beams [12]. From the geometry of a synchronous orbit satellite with global scanning capability,

$$L \simeq 1.4\ D \tag{13}$$

where L = physical diameter of the reflector
D = effective diameter of the reflector.

For a focal length-to-diameter ratio (F/D) of 0.7, the physical surface area of the reflector is

$$A = 0.837\ L^2 \simeq 1.64\ D^2. \tag{14}$$

The mass of the reflector can be obtained by multiplying A by ρ, the mass density of the reflector material.

When each beam is formed by one feed, the total antenna subsystem mass is

$$W_A \simeq 1.64\ \rho D^2 + X \cdot W_f \tag{15}$$

where W_f is the feed mass. The antenna gain is

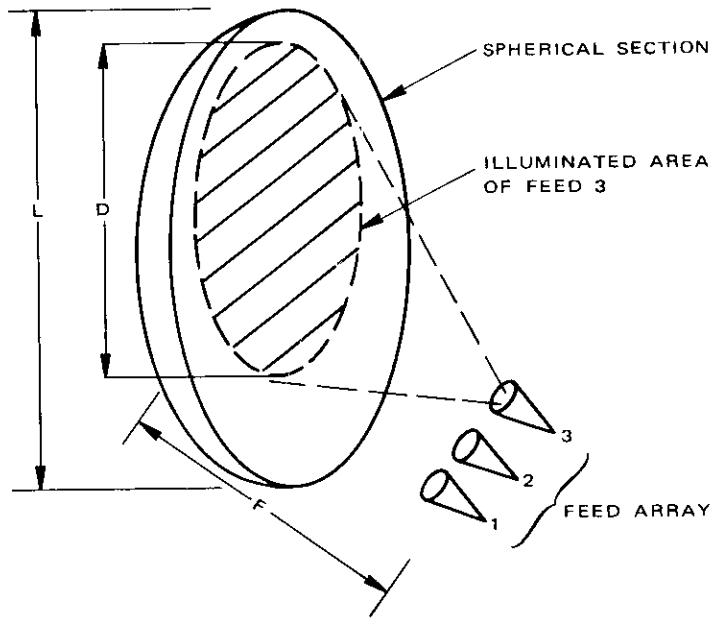


Figure 11. Multibeam Spherical Reflector Antenna

$$G_t = \frac{\eta_a \pi^2 D^2}{\lambda^2} \sim 10\eta_a f^2 D^2 \tag{16}$$

where η_a = antenna efficiency
λ = wavelength, in same units as D
f = frequency, in GHz if D is in feet

and the 3-dB beamwidth (in degrees) is

$$\theta_{3dB} \simeq \frac{69}{fD}. \tag{17}$$

Combined Payload Performance

Combining the analyses of the three preceding subsystems makes it possible to relate the total mass of the payload to its performance. In the following calculations, the satellite effective isotropic radiated power (e.i.r.p.) is assumed to be predetermined, reflecting a particular choice of modulation method, frequency plan, carrier bandwidth, and earth station G/T. Assume that the nominal e.i.r.p. per TWTA is P_w (in watts at beam center); then

$$P_o = \frac{P_w}{G_t a_t} \tag{18}$$

where a_t, which is less than or equal to one, is the transmission factor calculated from the line loss between the TWTA output and the antenna feed.

The previous derivations can be used to model the total communications payload mass (for a single service) for FDMA or TDMA. Combining equations (7), (8), (11), (12), (15), (16), and (18) results in

$$W_p = A_1 + A_2X + A_3XM + A_4X^2 + A_5XMP_w \tag{19}$$

where W_p = communications payload mass
$A_1 = W_a = 1.64\ \rho D^2$
$A_2 = W_r(1 + s) + W_f + C_1(P_r + \phi_p P_m)$
$A_3 = C_1 P_{to} + W_{to}(1 + s)$
$A_4 = \phi_w W_m$
$A_5 = C_2[\beta_p C_1 + \beta_w(1 + s)]$

and

$$C_1 = \frac{1.2K}{\sigma\eta_b\eta_r} + \frac{1}{\gamma\eta_m}\left(1 + \frac{K}{20\eta_b\eta_c\eta_e\eta_r}\right)$$

$$C_2 = \frac{1}{10\eta_a f^2 D^2 a_t}.$$

For multiple-carrier TWTA operation, equation (19) must be modified to include the effect of backoff. The e.i.r.p. at saturation becomes

$$P_w = P_a N_a B_o \tag{20}$$

where P_a = e.i.r.p. per carrier
N_a = number of carriers per TWTA
B_o = backoff factor, which is greater than one.

SYSTEM ASPECTS AND COST

Since performance versus mass relationships have been established in previous sections, cost can be related to mass. The satellite total mass is a critical parameter in terms of launch vehicle capability, payload complexity, and overall program scope. Although cost increases with W_S, experience has shown that it is not proportional to mass, but varies at a slower rate.

Figure 12 is a plot of satellite nonrecurring and recurring first costs as a function of W_S. The data points on the graph are for previous government and INTELSAT programs. It should be noted that certain assumptions were needed to generate this figure, since many programs do not segregate nonrecurring and recurring costs. Hence, these costs were apportioned on a best-fit basis. The straight lines are reasonable approximations for use in a cost model.

The nonrecurring cost can be written as

$$K_1 W_S^{n_1}$$

and the recurring cost can be written as

$$K_2 W_S^{n_1}.$$

From the data in Figure 12, $K_1 = 0.9$, $K_2 = 0.22$, and $n_1 = 1/2$. This figure is believed to be adequate for illustrating trends and comparing satellites which employ similar technology.

Figure 13 summarizes the more active launch vehicle combinations applicable to communications satellites. The Titan family is considered for simultaneously launching several satellites. A straight line of the form

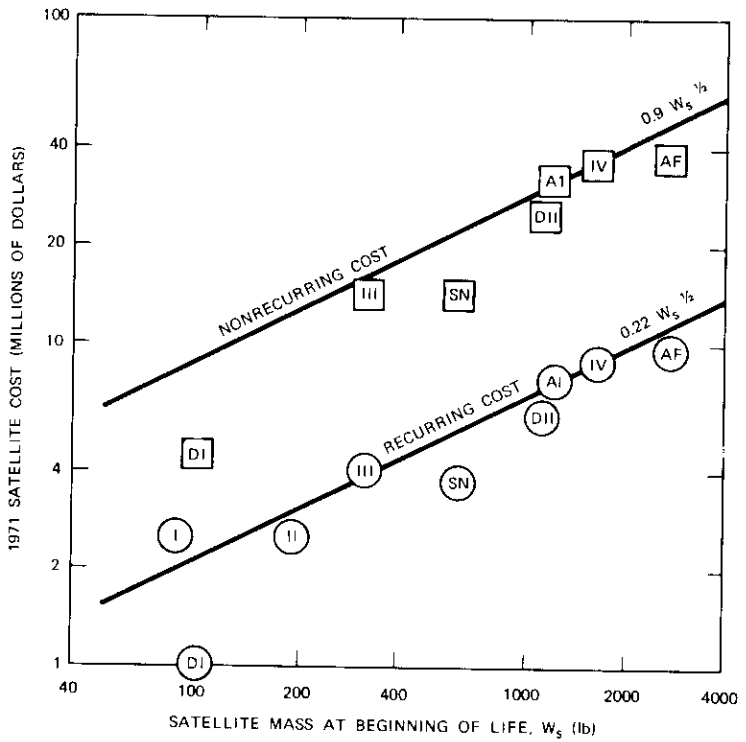
$$K_3 W_S^{n_2}$$

where $K_3 = 0.3$ and $n_2 = 1/2$ is plotted to indicate that the cost per pound of a satellite launch decreases slowly with increasing satellite mass.

It is essential to keep the following points in mind when employing a cost model such as the one just described:

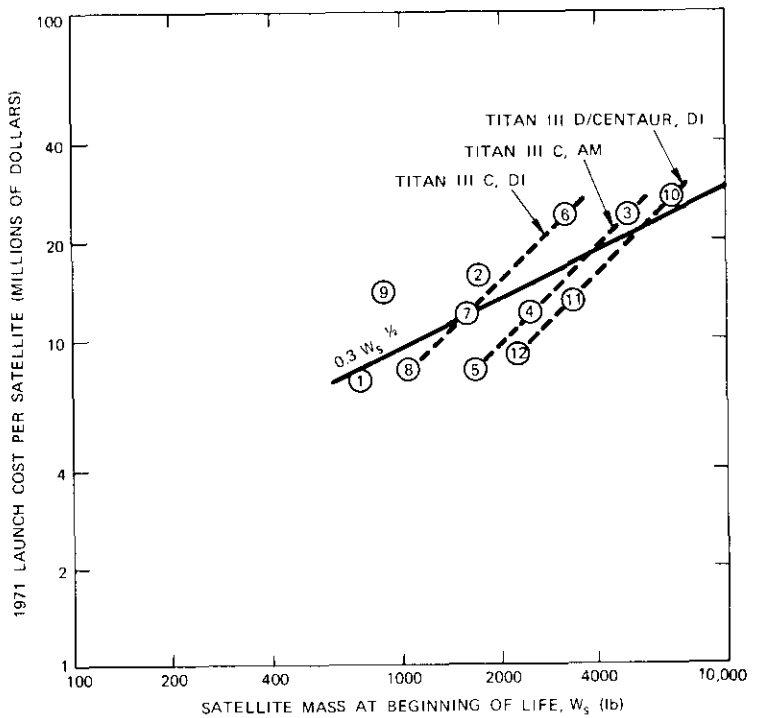
a. Cost models are no substitute for detailed cost estimates. Once the technology is sufficiently defined, the cost assumptions should be checked.

b. Communications satellite programs may have different overall objectives which must be taken into account. For example, commercial



SYMBOL		NO. OF FLIGHT MODELS
I	INTELSAT I	1
II	INTELSAT II	5
III	INTELSAT III	8
IV	INTELSAT IV	8
A1	ATS 1 THROUGH 5	5
AF	ATS F AND G	2
DI	DSCS–I (IDCSP)	32
DII	DSCS–II	6
SN	SKYNET AND NATO	4

Figure 12. Satellite Cost vs. Satellite Mass



	LAUNCH VEHICLE		NO. OF SATELLITES	STATUS
1	THOR–DELTA	AM	1	OP
2	ATLAS–CENTAUR	AM	1	OP
3	TITAN III–C	AM	1	OP
4	TITAN III–C	AM	2	OP
5	TITAN III–C	AM	3	OP
6	TITAN III–C	DI	1	OP
7	TITAN III–C	DI	2	OP
8	TITAN III–C	DI	3	OP
9	ATLAS-AGENA	DI	1	OP
10	TITAN III D/CENTAUR	DI	1	PL
11	TITAN III–D/CENTAUR	DI	2	PL
12	TITAN III–D/CENTAUR	DI	3	PL

AM APOGEE MOTOR
OP OPERATIONAL
DI – DIRECT INJECTION
PL – PLANNED

Figure 13. Launch Cost vs. Satellite Mass

communications satellites must be capable of reliable operation and long life with little or no unproven technology, whereas NASA Advanced Technology Satellites (ATS) emphasize new, advanced technology with in-orbit experimentation and demonstration.

c. Cost models do not anticipate technological breakthroughs. In the cost model presented in this paper, it is important to note that the functions of competition and the overall U.S. economic situation have not been taken into consideration.

The total space-segment cost, including the cost of R&D and all satellites and launches, is

$$C_S = R + (K_1 + ZK_2)\, W_S^{n_2} + ZK_3 W_S^{n_2} \tag{21}$$

where R = R&D cost

Z = total number of satellites launched.

To calculate Z, assume that the satellite mean lifetime, Y, can be calculated from an exponential failure model. Then, the average number of satellites required to maintain S satellites simultaneously in orbit during the system lifetime, Λ, is $S[1 + (\Lambda/Y)]$. If the probability of success of a single launch is P_l, then

$$Z \simeq \frac{S}{P_l}\left(1 + \frac{\Lambda}{Y}\right). \tag{22}$$

EXAMPLES

One suggested modeling approach is first to evaluate and optimize the communications payload(s) for the desired service(s). For a multipurpose satellite, the payload masses are summed. Then the total satellite mass is found by dividing the payload mass by the value of the utilization factor found in Figure 5 for the particular spacecraft configuration. Some iteration is necessary, since u is also sensitive to W_S. Finally, space-segment costs are calculated by using the cost model in the previous section.

Optimum Spacecraft Antenna Size

Differentiating equation (19) with respect to P_o establishes a minimum payload mass. At this optimum, the predetermined value of P_w is met by

the combination of P_o and D which minimizes total mass. Thus,

$$W_p|_{\substack{\min \\ \text{mass}}} = A_2X + A_3XM + A_4X^2 + 2C_3(XMP_w)^{1/2} \qquad (23)$$

where $C_3 = \{0.16\rho \cdot [\beta_p C_1 + \beta_w(1 + s)]a_t^{-1}\eta_a^{-1}f^{-2}\}^{1/2}$.

The mass of the optimum antenna reflector equals the partial mass of the power subsystem which supplies power to the TWTAs only. Note that this optimum reflector size establishes a beamwidth which may not provide usable coverage of the earth station network; however, it is important to the system designer to know how far a given design is from this optimum design. If there is a large disparity between the two designs, a different arrangement of antenna beams may be indicated. Fortunately W_p varies quite slowly around the optimum value for the assumptions used in Figure 14, where W_p is plotted as a function of beamwidth.

Comparison of Multiple-Access Methods

The combined payload mass relationships for the general and optimized cases [equations (19) and (23), respectively] can be used to examine different modulation and multiple-access schemes. The effects of imposing limitations on some aspect of the design, such as transponder bandwidth or bit rate per carrier, can be determined. Figure 15 shows the results of a comparison of TDMA and FDMA, both using digital four-phase coherent PSK (CPSK) modulation. It is assumed that there are eight 4° spot beams required and that the maximum capacity per beam is limited by bandwidth. For FDMA, the number of carriers per TWTA is varied and the effect of TWTA backoff included.

From Figure 15 it can be concluded that for FDMA there are circumstances in which multicarrier TWTA operation is preferable to single-carrier operation despite the need to operate with backoff. If minimum mass is the efficiency criterion, it can also be concluded that TDMA is more efficient than any of the FDMA arrangements.

Example of Satellite Systems Study

Satellite systems based on advanced and extended technologies were examined by using an assumed international communications network

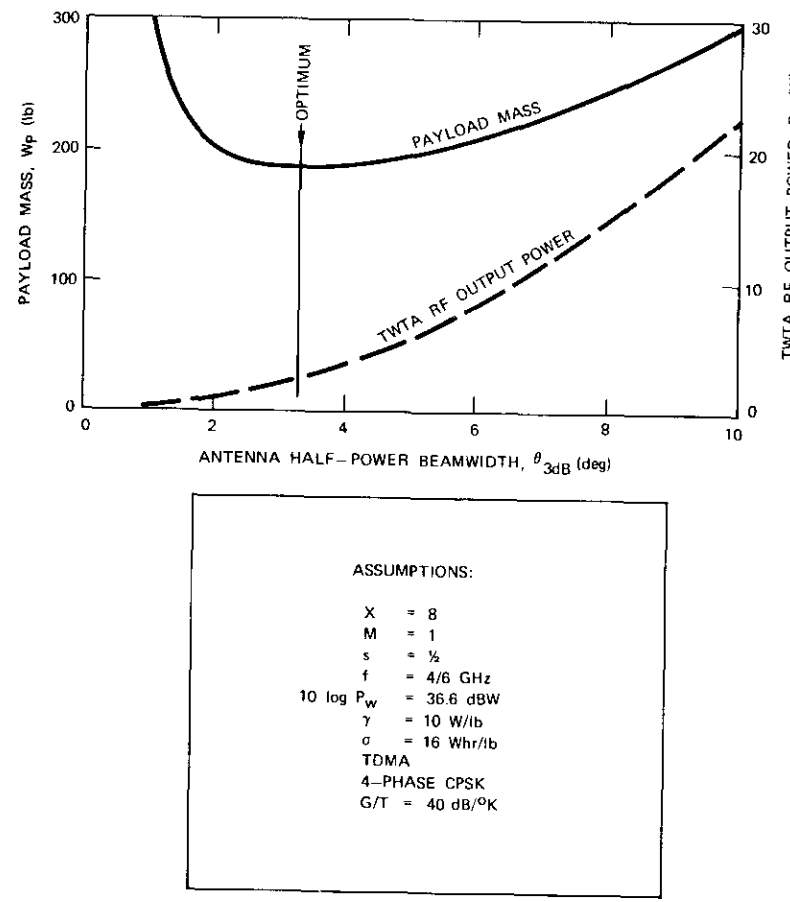


Figure 14. Example of Communications Payload Optimization

[13]. Table 1 lists assumed values of the principal parameters, and Figure 16 summarizes satellite mass and relative cost versus capacity, showing the advantages of advanced technology. To compare absolute costs, the coefficients K_1 and K_2 of equation (21) must be determined for each generic type. Additional data would be needed to compare costs of satellites based on dissimilar technological assumptions.

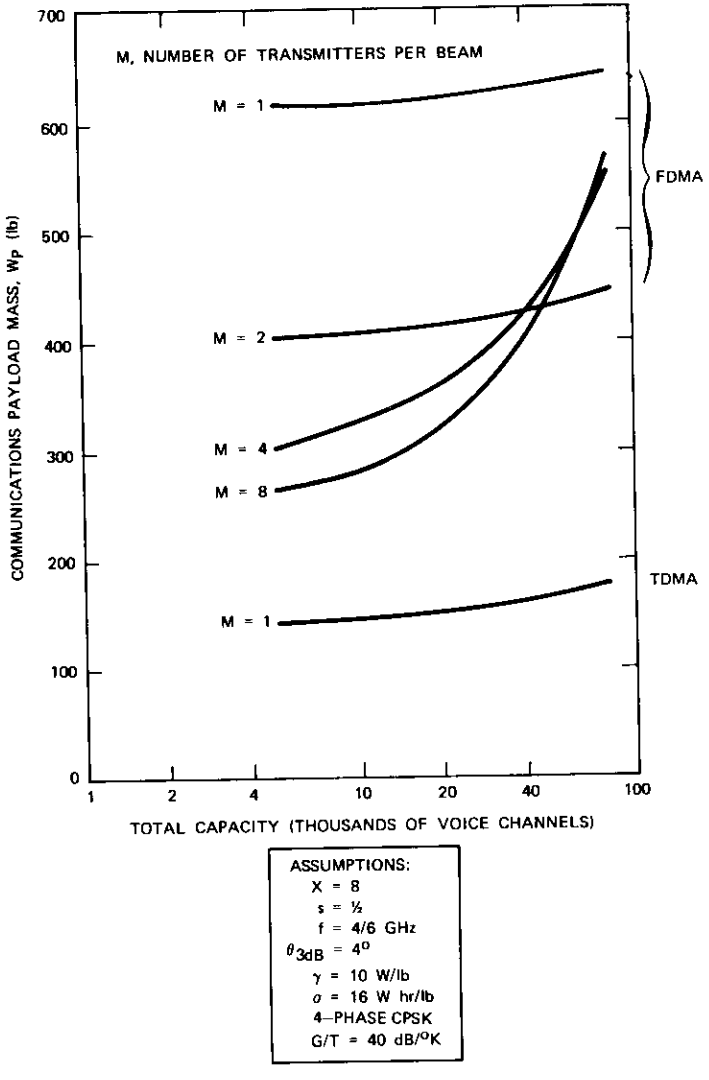


Figure 15. Multiple-Access Comparison

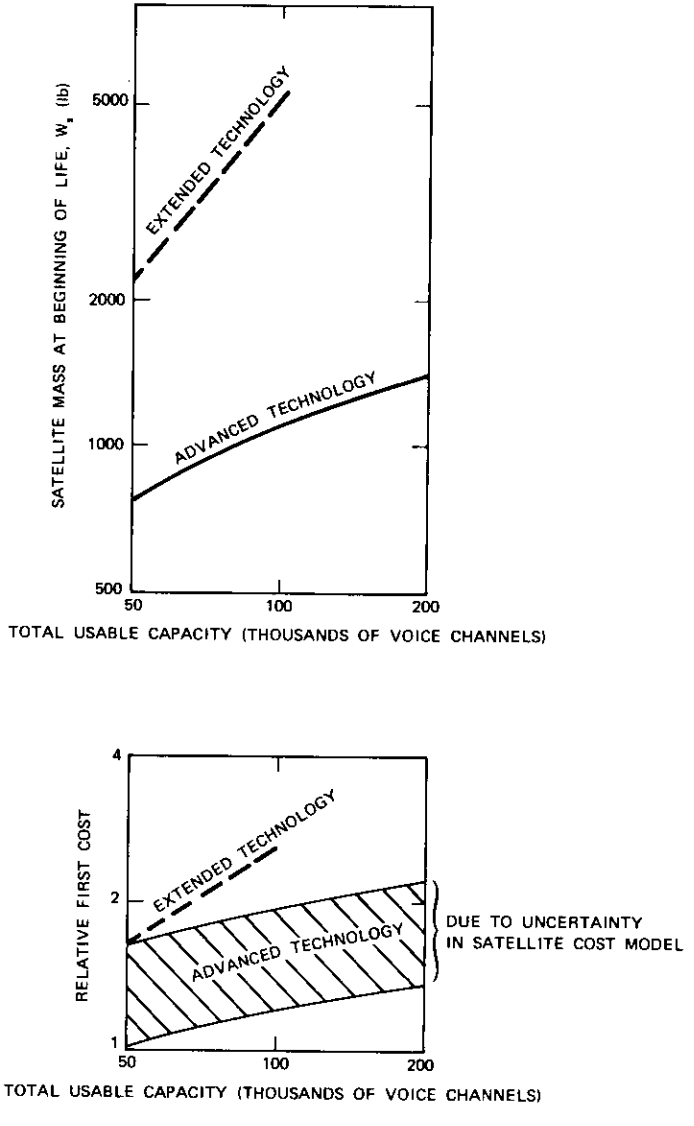


Figure 16. Satellite Mass and Relative First Cost vs. Capacity for Advanced and Extended Technology

TABLE 1. TECHNOLOGY AND PARAMETERS FOR SATELLITE SYSTEMS STUDY

Item	Advanced Technology	Extended Technology
Multiple Access	TDMA	FDMA
Modulation	Four-Phase CPSK	Four-Phase CPSK
Multiplex	TDM	TDM
Frequency Plan	4/6, 11/14, and 20/30 GHz	4/6 and 11/14 GHz
Earth Station G/T	40 dB/° K	40 dB/° K
Satellite Antenna Configuration	Multiple Beam	Multiple Beam
Stabilization Method	Body Stabilized	Body Stabilized
Solar Array	Flexible, Oriented Array γ = 10 W/lb	Rigid, Oriented Array γ = 6 W/lb
Energy Storage	Fuel Cells σ = 16 Whr/lb	Ni-Cd Batteries σ = 6 Whr/lb
Propulsion for North/South Stationkeeping	Ion Engine	Monopropellant Hydrazine

CONCLUSION

The satellite model developed in this paper is a valuable tool for the study of communications satellite systems since spacecraft, communications, and system characteristics are related to overall performance and cost. Important performance and cost factors have been introduced so that the parameter values can be altered to suit the user. By computerizing the model and maintaining an up-to-date catalog of parameters and subroutines, the user can accurately and rapidly make comparisons and measure the sensitivity of the results to the various assumptions.

LIST OF SYMBOLS

A	Reflector physical surface area
a_t	Transmission factor
B_o	Backoff factor
C_S	Total cost of satellites and launches
D	Reflector effective diameter
f	Frequency
F/D	Focal length-to-diameter ratio
G_t	Satellite antenna gain
G/T	Earth station figure of merit
K	Eclipse factor
K_1	Nonrecurring cost coefficient
K_2	Recurring cost coefficient
K_3	Launch cost coefficient
L	Reflector physical diameter
M	Number of transmitters per beam
N	Number of channels per carrier
N_a	Number of carriers per transmitter
n_1	Satellite cost exponent
n_2	Launch cost exponent
P_a	e.i.r.p. per carrier
P_{AR}	Maximum solar array DC output power
P_B	Spacecraft bus power load
P_E	Eclipse power load
P_{ES}	Energy storage subsystem charging load
P_l	Probability of launch success
P_m	Cross-connect matrix element power
P_o	TWTA RF output power
P_r	Receiver DC input power
P_T	Transponder subsystem power load
P_t	TWTA DC input power
P_{to}	Y intercept of TWTA power characteristic
P_w	e.i.r.p. per transmitter at saturation
R	R&D cost
S	Number of operational satellites
s	Spares ratio
u	Utilization factor
W_A	Antenna subsystem mass
W_a	Reflector mass

W_{AR} Solar array mass
W_B Spacecraft bus mass
W_{ES} Energy storage subsystem mass
W_f Antenna feed mass
W_m Cross-connect matrix element mass
W_p Communications payload mass
W_r Receiver mass
W_S Total satellite mass
W_T Transponder subsystem mass
W_t TWTA mass
W_{to} Y intercept of TWTA mass characteristic
X Number of spot beams (feeds)
Y Satellite mean lifetime
Z Total number of satellites procured
β_p Slope of TWTA power characteristic
β_w Slope of TWTA mass characteristic
γ Solar array specific power
η_a Antenna efficiency
η_b Boost regulator efficiency
η_c Charge controller efficiency
η_e Charging efficiency
η_m Main regulator efficiency
η_r Storage regulator efficiency
θ_{3dB} Antenna 3-dB beamwidth
Λ Satellite system lifetime
λ Wavelength
ρ Reflector mass density
σ Energy storage specific energy
ϕ_ρ Constant
ϕ_ω Constant

REFERENCES

[1] S. G. Lutz, "Small Earth Stations," Appendix C, Panel 9: Point-to-Point Communications, *Useful Applications of Earth-Oriented Satellites*, pp. 79–86, Washington, D.C.: National Academy of Sciences, 1969.

[2] L. C. Tillotson, "A Model of a Domestic Satellite Communications System," *Bell System Technical Journal*, Vol. 47, No. 10, December 1968, pp. 2111–2137.

[3] J. D. Kiesling and H. J. Meyerhoff, "TV Satellite Distribution at Frequencies Above 10 GHz," *Communication Satellites for the 70's: Systems (AIAA Progress in Aeronautics & Astronautics*, Vol. 26), edited by N. E. Feldman and C. M. Kelly, Cambridge, Mass: M.I.T. Press, 1971, pp. 171–190.

[4] *The ATS-F and -G Data Book*, Goddard Space Flight Center, Greenbelt, Maryland, October 1971, p. II–9.

[5] B. R. K. Pfeiffer and P. Viellard, "The Franco-German Telecommunication Satellite SYMPHONIE," *Communication Satellites for the 70's: Systems (AIAA Progress in Aeronautics & Astronautics*, Vol. 26), edited by N. E. Feldman and C. M. Kelly, Cambridge, Mass: M.I.T. Press, 1971, pp. 23–36.

[6] Q. B. McClannan and G. P. Heckert, "A Satellite System for CATV," *Proceedings of the IEEE* (U.S.A.), Vol. 58, No. 7, July 1970, pp. 987–1001.

[7] MCI Lockheed Satellite Corp., FCC Docket No. 16495, *Application for a Domestic Communications Satellite System*, Vol. I: Systems Application, February 1971.

[8] E. A. Richley and W. R. Kerslake, "Bombardment Thruster Investigations at the Lewis Research Center," *Journal of Spacecraft and Rockets*, Vol. 6, No. 3, March 1969, pp. 289–295.

[9] P. L. Bargellini, "Satellite Communications: Future Satellites," *Telephone Engineer and Management*, Vol. 75, No. 24, December 15, 1971, pp. 44–49.

[10] A. L. Berman and C. E. Mahle, "Nonlinear Phase Shift in Traveling-Wave Tubes as Applied to Multiple-Access Communications Satellites," *IEEE Transactions on Communications Technology*, COM–18, No. 1, February 1970, pp. 37–48.

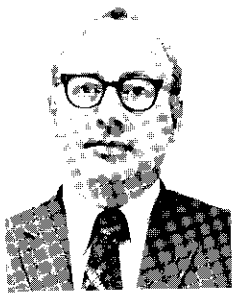
[11] J. G. Puente, W. G. Schmidt, and A. M. Werth, "Multiple-Access Techniques for Commercial Satellites," *Proceedings of the IEEE* (U.S.A.), Vol. 59, No. 2, February 1971, pp. 218–229.

[12] T. S. Chu, "A Multibeam Spherical Reflector Antenna," *1969 IEEE International Antennas and Propagation Symposium*, Austin, December 1969, *Program and Digest*, pp. 94–101.

[13] B. R. Elbert, "System Engineering Trade-offs for Advanced Communications Satellites," *AIAA Space Systems Meeting*, Denver, July 1971, AIAA Paper No. 71–843.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Mr. Emeric Podraczky for his guidance and contributions to this work.



Mr. Kiesling graduated from the Polytechnic Institute of Brooklyn (B.E.E., 1953 and M.E.E., 1958). He was employed for seven years at the Brookhaven National Laboratory performing research and development on high-energy proton linear accelerators. At RCA he was responsible for the development of the wide-band communications systems of the Relay spacecraft (a forerunner of modern communications satellites) and related programs. At COMSAT he leads a systems engineering group in studies of advanced communications satellite concepts and techniques.

Bruce R. Elbert received a B.E.E. degree from the City College of New York in 1965 and is presently a candidate for the M.S.E.E. degree from the University of Maryland. From 1965 to 1969 he served as a radio communications officer and instructor in the U. S. Army. He joined COMSAT Laboratories in 1969 where as a member of the Systems Planning Branch he performs engineering studies of communications satellite systems.

William B. Garner obtained a B.S.E.E. from Texas A&M in 1960 and has taken postgraduate courses at the University of Pennsylvania. From 1960 to 1968 he worked in the Systems Engineering department of the Astro Electronics Division of RCA. He is now a member of the technical staff in the Systems Laboratory at COMSAT Laboratories, working on modeling concepts for advanced communications satellite systems.

After graduating from Carnegie-Mellon University in 1954, Mr. Morgan joined the David Sarnoff Research Center of RCA Laboratories. With RCA he was a systems engineer for over twenty successful spacecraft (TIROS, ESSA, Lunar Orbiter, and others). Mr. Morgan joined COMSAT Laboratories in 1969, where he has been engaged in studies of future communications satellites and their applications.

Index: efficiency, electric power generation, recombination, solar cells

Theoretical and practical fill factors in solar cells

J. LINDMAYER

ABSTRACT

The fill factor determines the fraction of power available to a load with respect to the open-circuit-voltage short-circuit-current product. This paper calculates the theoretically expected values and compares them with measured values. In silicon p-n junction energy converters, fill factor values over 0.82 are expected according to diffusion theory, but actual solar cells exhibit a fill factor of about 0.72. Recombination in the space charge region is found to be the reason for the reduced fill factor in practical cells. The mechanism discussed also has implications with respect to the open-circuit voltage.

INTRODUCTION

The electric power available from solar cells may be written as

$$P_m = I_{sc} V_{oc} F \tag{1}$$

where I_{sc} is the short-circuit current, V_{oc} is the open-circuit voltage, and F is the "fill factor".* Figure 1 shows the characteristics of different generators. The "classical" generator with an ohmic internal resistance is shown as a straight line with $F = 0.25$; the "ideal" generator is repre-

* Sometimes the term "curve factor," which does not include the effect of series resistance, is used.

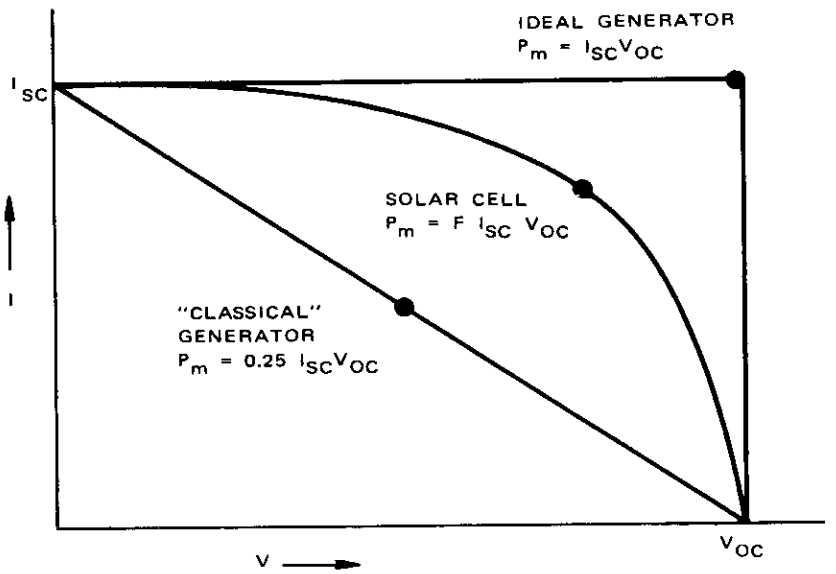


Figure 1. Current-voltage relation for different generators. The fill factor varies between 0.25 and 1.0.

sented by a rectangle with $F = 1$. A practical solar cell lies between these extreme cases.

The most efficient solar energy converters employ p-n homojunctions. The photogenerated minority carriers will be collected at the junction, adding a photocurrent in the reverse direction to the *I-V* characteristic. Figure 2 shows the corresponding displacement of the dark *I-V* curve by the photocurrent and, therefore, the fill factor should be connected with the shape of the dark *I-V* characteristic. Note that in the power-producing quadrant the junction is biased in the forward direction.

The use of the following simple diode equation is common in the solar cell literature (for the unilluminated cell):

$$I = I_0(e^{qV/AkT} - 1) \tag{2}$$

where I_0 is a constant with the dimensions of current and kT/q is the thermal voltage (26 mV at room temperature). The factor A is a dimen-

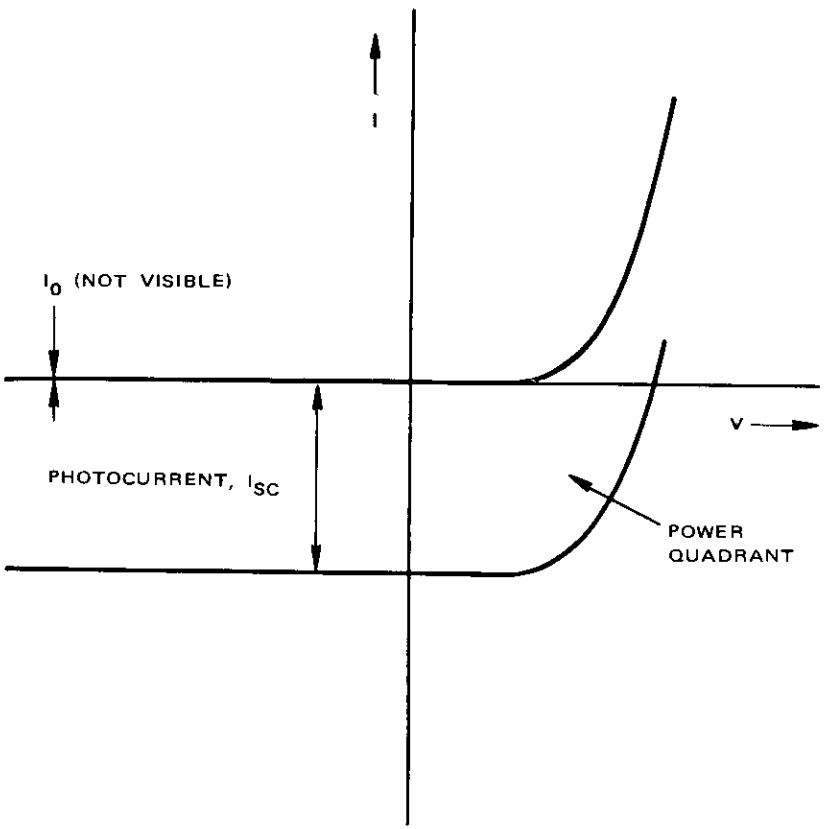


Figure 2. The photocurrent should be added to the normal I-V curve as a "reverse" current.

sionless number having a value of unity for the simple diffusion theory of diodes, but exhibiting a value greater than unity in practical silicon cells. In a recent review paper by Wolf [1] it is stated that in the absence of any accepted physical theory the magnitude of A has to be considered wholly material and process dependent. It is also stated that $A > 2$ in practical silicon solar cells. In addition to the uncertain value of A, one finds that I_0 is not the actual saturation value of the reverse current of the diode, but rather a current value applicable in the forward direction. Moreover,

the observed value of I_0 is usually different from that predictable from the diffusion theory of diodes.

In the general solid-state field it has long ago been noted that, unlike germanium, silicon p-n junctions do not follow the ideal diffusion theory. Also, silicon transistors have shown low current gains at low emitter currents, indicating that injection is shunted by some other current path. On the basis of these observations Sah, Noyce, and Shockley [2] have theorized that preferential recombination will occur in the space charge region. The origin of this effect will be discussed qualitatively with the aid of Figure 3. In this figure the depletion layer of an n-p junction is shown. The forward injection current will be shunted by a recombination current in the space charge region (SR current) which is dominant at low forward bias, but will increase with applied bias slower than the injection current. One can argue that electrons and holes injected into the space charge region can recombine via localized centers without surmounting the full barrier. Depending on the potential distribution (and therefore the impurity distribution), SR current is modulated by a fraction of the applied bias, V/n, where n is greater than unity. If recombination peaks in the center of the barrier, the value of n will be 2. The quantitative treatment of Reference [2] shows that

$$I = I_{01}(e^{qV/nkT} - 1) + I_{02}(e^{qV/kT} - 1) \tag{3}$$

where the first term represents the SR current, and the second, the normal injection. As was also shown, in silicon (and other wide bandgap crystals) $I_{01} \gg I_{02}$. This means that at low voltages the first term dominates, but above a critical applied voltage, V_{onset}, normal injection sets in. As far as the value of n is concerned, Sah *et al* [2] allow for values between 1 and 2; however, a wider range must be applicable for diffused junctions where the potential distribution can be altered significantly. While the actual value of n must be calculated for each specific impurity and state distribution (by solving the nonlinear Poisson equation), for a first approximation it may be said that the SR current peaks at half potential, yielding $n \cong 2$.

The identity between n and the constant A used in the solar cell literature is unmistakable. In later parts of this paper it will be shown that space charge recombination is indeed an important factor in actual solar cell behavior.

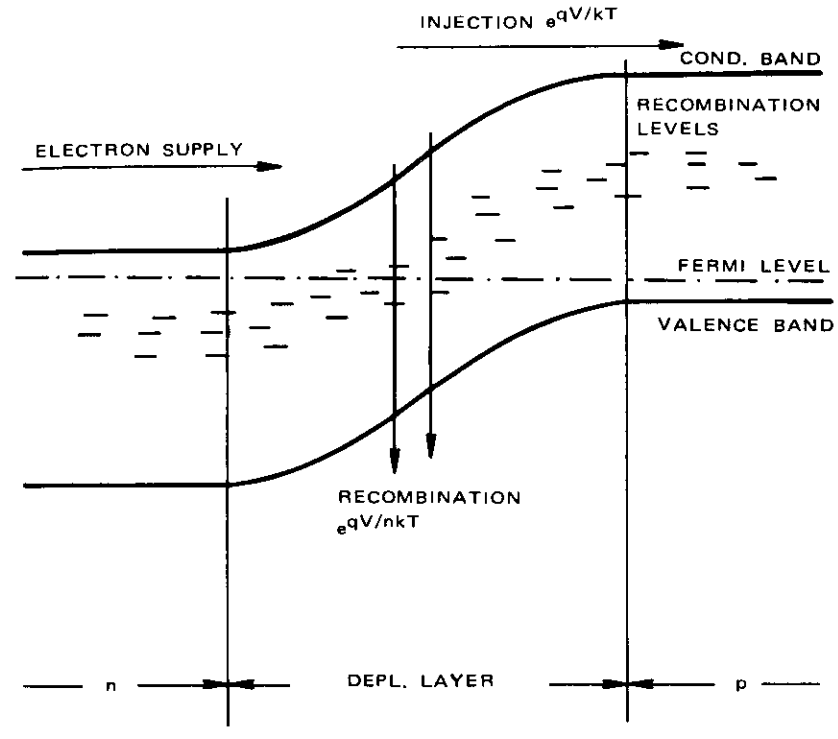


Figure 3. Energy barrier at an n-p junction. Space charge recombination current through recombination centers in the middle of the barrier is the first current component that flows as a forward bias is applied in a p-n junction. True injection will start at higher voltages.

CALCULATION OF FILL FACTOR

The current-voltage characteristics of a p-n junction solar energy converter are usually expressed in the form:*

$$I = I_{sc} - I_0(e^{V/V_0} - 1). \tag{4}$$

* Series resistance has been neglected for simplicity.

It is inherent in this relationship that the dark *I-V* curve can be described by one exponential segment and that the photocurrent is additive. (The question of a single exponential segment for the dark *I-V* curve will be discussed in more detail.) Equation (4) is applicable to heterojunctions as well, where exponential injection or tunneling occurs. For the case of injection, $V_0 = nkT/q$.

The short-circuit current, I_{sc}, and the open-circuit voltage, V_{oc}, may be interrelated by setting $I = 0$ in equation (4); thus, one obtains

$$I_{sc} = I_0(e^{V_{oc}/V_0} - 1). \tag{5}$$

The maximum power point can be calculated in the following manner. The power, P, is equal to $I \times V$. Setting dP/dV to zero defines the voltage, V_m, at which maximum power is available. The following implicit relationship is found:

$$\frac{V_m}{V_0} = \frac{V_{oc}}{V_0} - \ln\left(1 + \frac{V_m}{V_0}\right). \tag{6}$$

It follows from equation (1) that the fill factor is

$$F = \frac{I_m V_m}{I_{sc} V_{oc}}. \tag{7}$$

From equations (4) and (5), F may be written as

$$F = \frac{V_m}{V_{oc}}\left(1 - \frac{e^{V_m/V_0} - 1}{e^{V_{oc}/V_0} - 1}\right). \tag{8}$$

It may be noted that both V_{oc} and V_m are normalized with respect to V_0, in both equations (6) and (8). In fact, F may be computed as a singular function of V_{oc}/V_0, as shown in Figure 4. For an injecting junction $V_0 = nkT/q$, and therefore for room temperature F can be plotted as a function of n and V_{oc}, as shown in Figure 5. Note the monotonic decrease in F with increasing n and decreasing V_{oc}. The origin of the experimental points obtained for silicon cells will be discussed in the next section.

APPLICATION OF RESULTS TO SILICON CELLS

The open-circuit voltage for n^+-p silicon solar cells is typically 550 mV

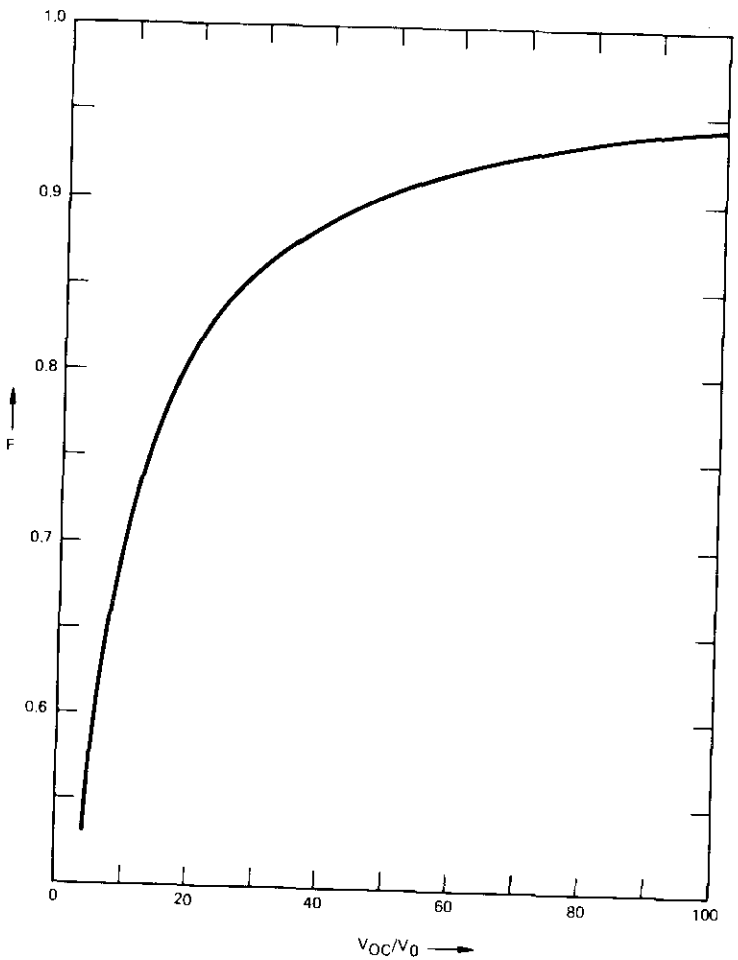


Figure 4. Fill factor as a function of V_{oc}/V_0, applicable to all exponential junction characteristics.

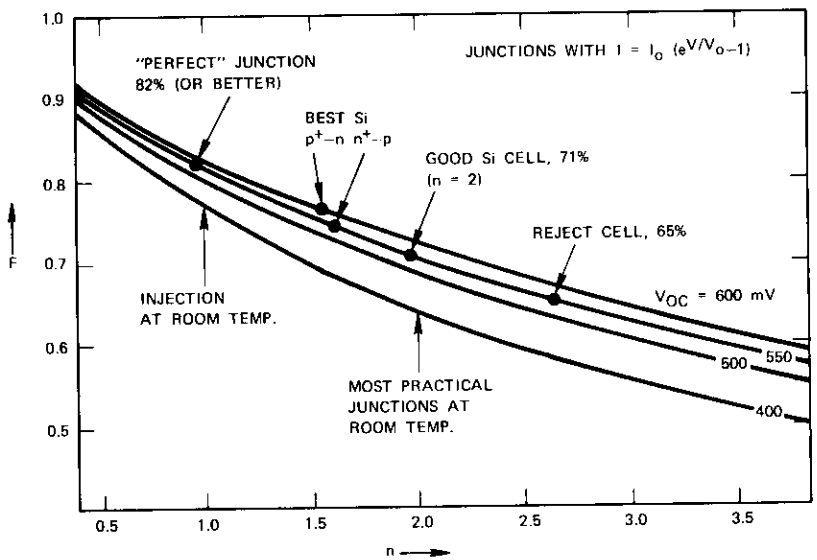


Figure 5. Calculated fill factor for injecting junctions as a function of V_0 and the open-circuit voltages. The points plotted correspond to samples discussed in the text.

at 1 sun intensity and room temperature. For the ideal diode, the theoretical fill factor would be [3]

$$F = 0.82 \ (n = 1, \text{room T}, V_{oc} = 550 \text{ mV}). \tag{9}$$

If the cell were to operate in the SR current region, $n \cong 2$, the expected fill factor would be (from Figure 5)

$$F = 0.71 \ (n = 2, \text{room T}, V_{oc} = 550 \text{ mV}). \tag{10}$$

Equipped with a theoretical interrelationship between F and n, one can determine whether the F value measured under a solar simulator is the same as that predicted from the n value measured in the dark.

The value of n can be determined from the slope of the DC characteristics (log I versus V). The dark I-V curve, however, is quite complicated. As is well known from transistor studies, at low bias voltages in the forward direction the onset of current flow occurs at the edges; at higher

biases the current spreads over the whole junction with a uniform current density. This nonuniformity is not present at biases corresponding to 1 sun illumination where uniformity is further ensured by the uniform illumination. There are four regimes observable in the I-V characteristic as the forward bias is progressively increased:

a. enhanced current flowing at the edges at low applied forward bias (dominant in the dark);

b. space charge recombination current with $n > 1$;

c. normal injection with $n = 1$, but with V lowered by the drop across the series resistance; and

d. predominantly ohmic behavior at higher currents as the series resistance assumes control.

Examination of numerous I-V characteristics has shown that normal injection ($n = 1$) will not be reached at the operating voltage V_m at which F is measured. Since point-by-point DC measurements are very time consuming and require evaluation, a simpler method employing an automatic capacitance-conductance bridge* was used. If an exponential segment is characterized by $I = I_0 \exp(V/V_0)$, the conductance is $G = dI/dV = I/V_0$. For the ideal diode $V_0 = kT/q$, but for the actual diode $V_0 = nkT/q$. Therefore, we find that the ratio of ideal to measured conductance supplies the constant n:

$$n = \frac{\text{calculated conductance with } n = 1}{\text{measured conductance}} = \frac{qI/kT}{G_m}. \tag{11}$$

This method is simple and accurate and makes it possible to plot n directly as a function of forward current.

The bridge measurement also determines the capacitance of the diode. The capacitance of a p-n junction in the forward direction is the sum of the depletion layer capacitance and the so-called diffusion capacitance. The latter is proportional to the forward current (in the ideal case) and represents charge storage in the form of injected minority carriers. The stored charge is $I \times \tau$, where τ is the lifetime of the injected minority carriers in the bulk region. For a wide-base diode the measured capacitance is [4] $C_m = G_m \tau/2 + C_J$, where C_J is the junction capacitance. Thus, the quantity τ is determined from

* The conductance and capacitance are measured automatically at 1 kHz on a GR 1673-A bridge locked in range 6, which limits the measuring signal to 10 mV.

$$\tau = \frac{2(C_m - C_J)}{G_m} . \qquad (12)$$

Figure 6 shows the *I-V* curves of three solar cells. The "typical" n^+-p cell exhibits a value of $n = 2$ in the low current region for both AC and DC measurements. According to Figure 5, the fill factor should be 0.71;

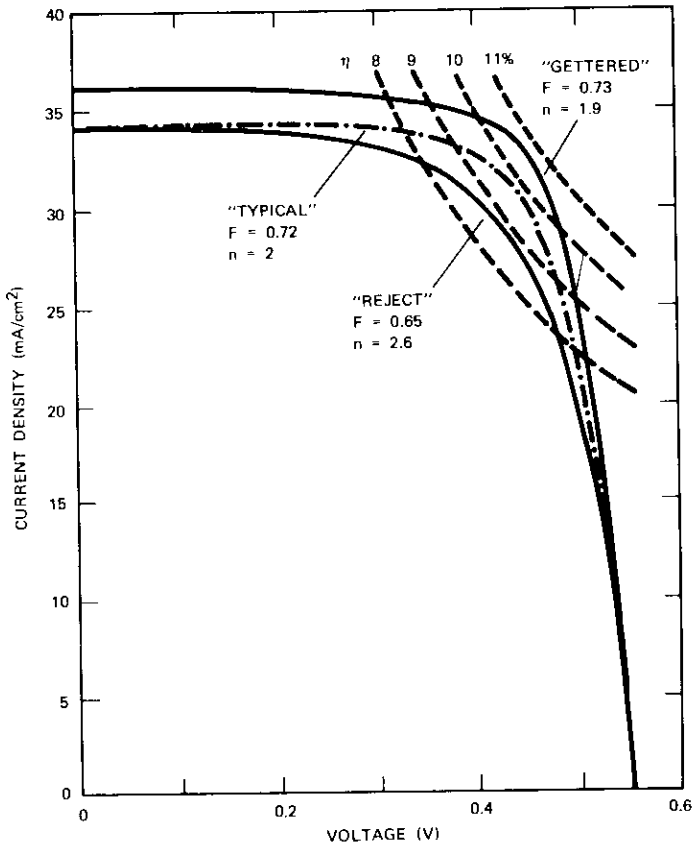


Figure 6. I-V characteristics under 1 sun illumination for "reject" and "gettered" cells. The dashed line represents the typical cell.

the measurement under illumination yields a value of 0.72. The lifetime of electrons in this cell was measured to be 6 μs.

A number of cells having higher recombination state densities were also measured. The increased state density was manifested by a reduced lifetime, in the range of 1-3 μs. The lower curve in Figure 6, marked "reject," is the *I-V* curve for one such cell. It may be noted that, although the open-circuit voltage is the same and the short-circuit current is slightly less than for the typical cell, the fill factor is only 0.65. With reference to Figure 5, this is only possible if $n = 2.6$, or if a series resistance is involved. Figure 7 shows the measured lifetime and n as evaluated from admittance measurements. At very low values of forward current, the edge effects

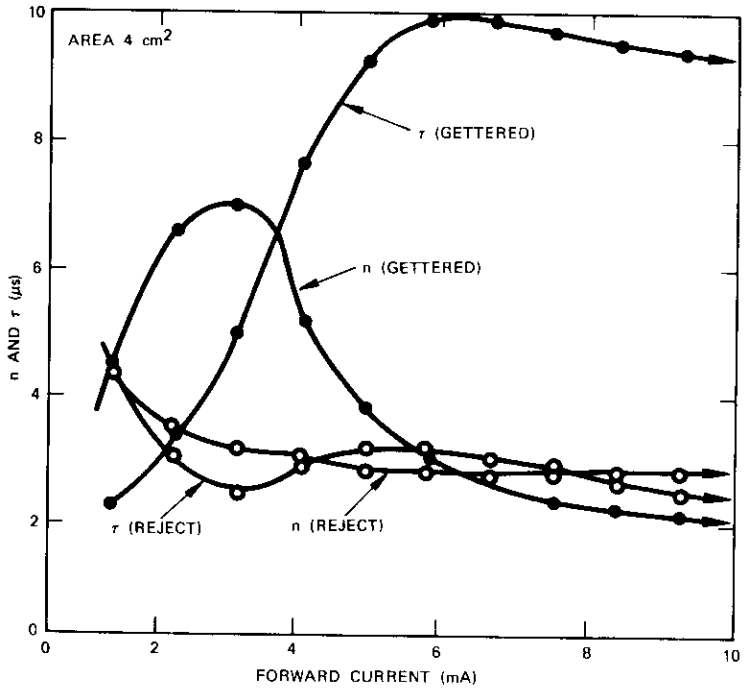


Figure 7. Lifetime and factor n as a function of forward current determined from the junction admittance. The values level off before the effects of a series resistance can be detected.

disturb the measurement. For forward currents as low as the photocurrent produced by an illumination of 0.2 sun, the parameters saturate before effects of any series resistance are noted, with $\tau_n = 2.5$ μs and $n = 2.6$, in perfect agreement with values implied by the F value.

Recent silicon cells employing improved gettering [5] (to reduce recombination state density) appear to be better than the typical cell. For these cells a fill factor as high as 0.74 has been observed. The upper curve of Figure 6 shows the I-V curve for a thin (150-micron) gettered cell. All numbers (I_{sc}, V_{oc} and F) are slightly better than those of the typical cell. The lifetime and n were also measured independently with the results shown in Figure 7. One finds an improved lifetime ($\tau > 9$ μs) and an n value of about 1.9. The observed fill factor for this cell is 0.73, in good agreement with the value of 0.72 expected for $n = 1.9$.

A number of p^+-n cells were also examined. It is well known that this type of cell is more efficient than the n^+-p version, but shows a greater sensitivity to radiation. The form of the I-V curve is of course similar, but higher F values were found. The highest value seen was $F = 0.76$, which, according to the calculated results in Figure 5, corresponds to $n = 1.55$. It is significant that the admittance measurements at low currents also show a saturated n value in agreement with that expected from the fill factor, namely 1.55, as seen in Figure 8. It appears that p^+-n cells have a lower state density than n^+-p cells, reducing the value of n and vastly increasing the lifetime.

Repeated measurements on a variety of silicon cells show that a direct correlation exists between the fill factor, as measured by the I-V characteristic under 1 sun illumination, and n, as measured in the dark by the admittance. In all cases, a lower n (higher F) is always associated with an increased lifetime, indicating a lower recombination state density.

DISCUSSION

The experiments, when compared with the calculations, indicate that space charge recombination is responsible for the reduced fill factor in actual silicon cells. In the better cells n is less than 2, but still much higher than unity; accordingly, in the n^+-p cell F is greater than 0.71, but still much lower than the expected 0.82.

Space charge recombination and recombination in general are controlled by recombination sites in the crystal structure. It is generally believed that this site density is controlled by traces of metals and by defects introduced during processing. Variations in the recombination state

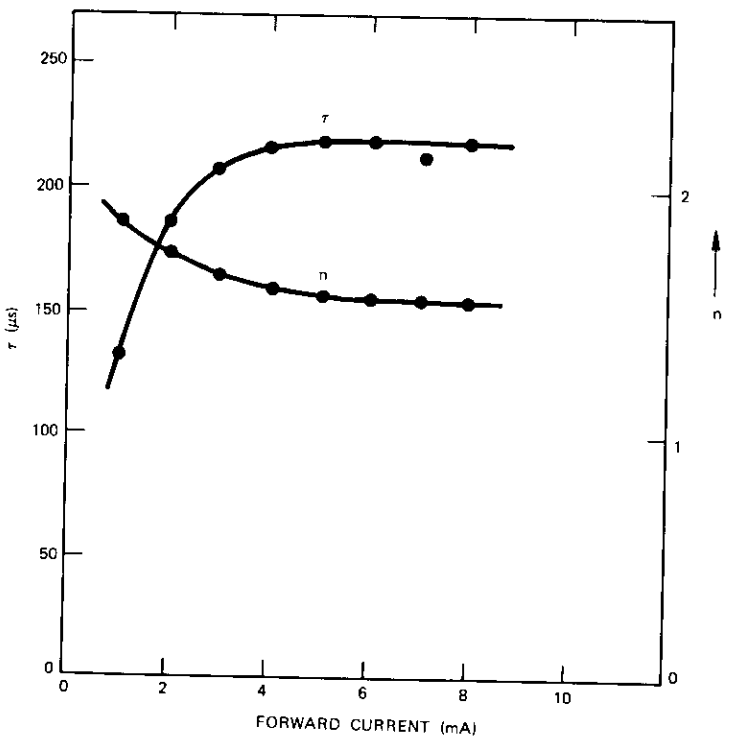


Figure 8. Lifetime and n in an extremely good p^+-n cell, as determined from diode admittance measurements.

density have a direct effect on the bulk lifetime (and therefore on the short-circuit current), but a modest effect* on n and F. When the state density is increased, F will not follow directly because V_{onset}, the boundary be-

* Consideration must be given to the fact that in diffused cells the impurity concentration at the surface reaches the solid solubility limit. It has been widely observed in the semiconductor field that stress-induced dislocations will advance ahead of the diffused junction. It is therefore expected that the recombination state density will be higher in the depletion layer than in the bulk and will also have a different energy distribution.

tween SR and injection currents, is the quantity that increases, leaving n relatively unchanged.

The present study indicates that presently available silicon solar cells operate in the space charge recombination regime, e.g., at maximum power $V_m < V_{onset}$.

The presence of space charge recombination has the following consequences:

a. The efficiency is reduced by space charge recombination through the reduced fill factor by approximately 16 percent in the n^+-p cell.

b. Space charge recombination, while negligible in the short-circuit current, is a factor in the open-circuit voltage. This recombination represents an internal, parallel load on the voltage generator. Generalizing equation (4) with the use of the current components in equation (3), we find for the case $I = 0$ that

$$V_{oc}\ (\text{without SR}) = \frac{kT}{q} \ln\left(1 + \frac{I_{sc}}{I_{02}}\right) \tag{13}$$

$$V_{oc}\ (\text{with SR}) = \frac{nkT}{q} \ln\left(1 + \frac{I_{sc}}{I_{01}}\right) \tag{14}$$

If SR could be eliminated, n would go from 2 to 1 and I_{01} would go to I_{02}. While n decreases, the logarithm increases with a net improvement in photovoltage when the differences between I_{01} and I_{02} are sufficiently great.

c. Generation of new states by penetrating, ionizing radiation will not change n or the fill factor significantly. While V_{onset} will increase, n will change little if the cell is already in the SR condition. The value of I_{01} will increase, of course, but because it resides in a logarithmic term, it will not result in a significant decrease in F. Irradiation by 1-MeV electrons to the point where τ decreases by an order of magnitude actually results in very little change in F. This shows again that the typical cell is in the SR condition. Proton damage is a different matter, since in this case the state density can be altered locally. For example, when defects are introduced just beyond the junction, injection will be enhanced and n will approach 1, in spite of the fact that the cell performance will deteriorate.

d. As Figure 9 shows, the efficiency is quite constant for illumination levels between 0.1 and 1 sun, the cell being locked in the SR condition. However, use of higher-intensity radiation, such as focused sunlight,

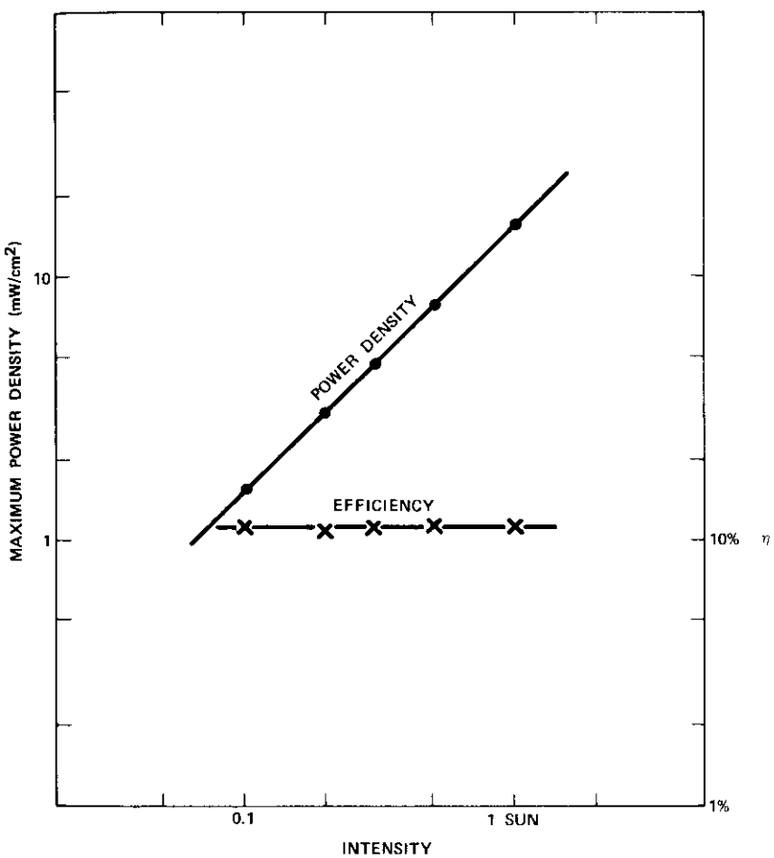


Figure 9. Efficiency and maximum power density vs. intensity of illumination. The efficiency of Si cells is quite constant between 0.1 and 1 sun. When the intensity is decreased V_{oc} drops very slightly, but I_{sc}/unit intensity improves somewhat.

should improve the efficiency if true injection conditions could be achieved. This would improve the fill factor and the open-circuit voltage as well.

e. Space charge recombination (and a correspondingly higher value of n) is expected to be increasingly dominant as the bandgap of the material used is increased.

SUMMARY

This study shows that the fill factor of silicon cells can be related to the low current n value in the diode equation. Specifically, a measured set of V_{oc} and F values supplies an n value through the calculated curves in Figure 5; the conductance measurement in the forward diode direction at moderate currents supplies another n value which is independent of series resistance. These values coincide in the range of $n = 1.55$ to $n = 2.6$, indicating that space charge recombination is an important factor in presently available silicon cells.

REFERENCES

[1] M. Wolf, "A New Look at Silicon Solar Cell Performance," *IEEE Eighth Photovoltaic Specialists Conference*, Seattle, August 1970, *Conference Record*, pp. 360–371.

[2] C. T. Sah, R. N. Noyce, and W. Shockley, "Carrier Generation and Recombination in P-N Junctions and P-N Junction Characteristics," *Proceedings of the IRE* (U.S.), Vol. 45, No. 9, September 1957, pp. 1228–1243.

[3] M. Wolf, "Limitations and Possibilities for Improvement of Photovoltaic Solar Energy Converters," *Proceedings of the IRE* (U.S.), Vol. 48, No. 7, July 1960, pp. 1246–1263.

[4] J. Lindmayer and C. Y. Wrigley, *Fundamentals of Semiconductor Devices*, Princeton, N.J.: D. Van Nostrand Co., Inc., 1965; p. 47.

[5] H. Fischer, E. Link, and W. Pschunder, "Influence of Controlled Lifetime Doping on Ultimate Technological Performance of Silicon Solar Cells," *IEEE Eighth Photovoltaic Specialists Conference*, Seattle, August 1970, *Conference Record*, pp. 70–77.

ACKNOWLEDGMENT

The author is thankful to E. S. Rittner for helpful discussions and criticism of the manuscript; to R. Rostron, A. Meulenberg and D. Curtin for making radiation data available; and to J. Haynos for help in the solar simulator measurements.

Dr. Joseph Lindmayer was educated as an E.E. in Hungary. He received his M.S. degree in Physics in 1963 at Williams College and his Ph.D. in Aachen, Germany, in 1968. Since 1968 he has been Manager of the Solid State Physics Branch at COMSAT Laboratories. He has contributed to the general field of solid-state electronics and is the author or co-author of numerous scientific articles in this field, as well as a textbook, "Fundamentals of Semiconductor Devices."

Index: communication satellites, ion propulsion, positioning, spacecraft propulsion

Chemical and electric propulsion tradeoffs for communications satellites

B. A. FREE

ABSTRACT

This paper analyzes an electrostatic ion thruster system to demonstrate the potential benefits which accrue from the low propellant mass requirements associated with electric propulsion. The propulsion tasks selected for this analysis are north-south stationkeeping and large longitude changes.

The use of electric propulsion for north-south stationkeeping appears especially advantageous in terms of propulsion system weight reduction. For example, if electric propulsion is used instead of a hydrazine monopropellant system, the predicted weight saving is about 13 percent of the entire satellite mass for a 10-year mission. For large longitude changes, such as repositioning across 120° of longitude, a lower weight saving is predicted, i.e., about 3 percent of the entire satellite mass. These weight savings can be converted directly into additional communications capacity since no additional structural requirements are introduced.

INTRODUCTION

The chief advantage of electric propulsion over chemical propulsion is the high propellant exhaust velocity attainable, which greatly reduces the

---

This paper is based upon work performed at COMSAT Laboratories under Corporate sponsorship and under the sponsorship of the International Telecommunications Satellite Consortium (INTELSAT). Any views expressed in this paper are not necessarily those of INTELSAT.

propellant mass requirement for any given propulsion task. Although this reduction is partly neutralized by the higher fixed mass of the electric propulsion system hardware, for many propulsion tasks there is a substantial net reduction in the mass of the propulsion system. This paper attempts to estimate this mass reduction for two selected propulsion tasks.

Selected Propulsion Tasks

The comparisons and conclusions given in this paper are for two propulsion tasks: north-south stationkeeping and longitudinal repositioning of a 700-kg synchronous satellite. Propulsion tasks calling for much larger thrust levels, such as large altitude changes [1], and propulsion tasks for which much smaller thrust levels are sufficient, such as attitude control [2]-[4], require different methods of evaluation.

North-south stationkeeping and longitudinal repositioning were chosen for contrast. North-south stationkeeping can be accomplished without incurring any significant penalties by using low thrust and power over a long total thrusting time. These characteristics define a propulsion task which is ideally suited to electric propulsion. On the other hand, longitudinal repositioning must usually be accomplished in a relatively short maneuver time (weeks to months), which requires high thrust over a short total thrusting time. In addition, when electric propulsion is used to accomplish this task, the velocity increment is twice that required for chemical propulsion. Therefore, repositioning is a propulsion task which severely tests the limitations of electric propulsion.

Propulsion Systems Characteristics

Hydrazine monopropellant is used throughout this paper to represent chemical propulsion. The simplicity and reliability of the monopropellant system, amply demonstrated in space, more than make up for its exhaust velocity, which is lower than that of other chemical propellants. Hydrazine propulsion systems onboard satellites generally operate at thrust levels of 1 to 25 N (0.2 to 5 lb).

The electrostatic ion thruster has been chosen to exemplify electric propulsion, partly to simplify the presentation and partly because ion thrusters are eminently suitable for the propulsion tasks discussed. This form of propulsion requires external power to vaporize, ionize, accelerate, and neutralize the propellant. Of all of these processes, only the acceleration step produces thrust. In order to maintain a low power supply mass,

ion thrusters operate at very low thrust levels. For north-south stationkeeping, a thrust level of 5 to 10 mN suffices, and for repositioning 10 to 100 mN may be required. A consequence of the low thrust level is that electric propulsion systems must be operated over much longer thrust periods than chemical propulsion systems.

Propulsion System Component Masses

For hydrazine monopropellant, the bulk of the propulsion system mass is propellant, while for electric propulsion most of the mass is the system hardware. Let us look first at the method of estimating propellant mass.

In terms of required velocity increment, u, propellant exhaust velocity,* v, and total initial mass, M_o, the required propellant mass, M_p, is given by the basic propulsion relationship:

$$M_p = M_o(1 - e^{-u/v}). \tag{1}$$

In equation (1) it is important to note that the ratio of velocity increment to propellant exhaust velocity determines the propellant mass. The velocity increment, which is defined as the integral over time of the absolute value of the thrust-to-mass ratio, is a convenient measure of the amount of propulsion required. It is numerically equal to the change in satellite velocity which is produced in the absence of all other forces.

Because hydrazine monopropellant has a fixed amount of chemical energy per unit mass, the exhaust jet has a fixed velocity of about 2.2 km/s. Therefore, in this case the propellant mass is a function only of the initial satellite mass and the velocity increment. The solution of equation (1) for hydrazine and a 700-kg initial mass is illustrated in Figure 1. The small amount of mass allocated for hardware in Figure 1 was estimated from the known component masses of the INTELSAT IV propulsion system.

For electric propulsion, it is more difficult to calculate the propellant mass since the propellant exhaust velocity is not fixed. The variation of propulsion system component masses with increasing exhaust velocity is illustrated in Figure 2 for a velocity increment of about 420 m/s (10 years of north-south stationkeeping). The propellant mass decreases exponen-

* Specific impulse, I_{sp}, which is the ratio of thrust to propellant mass flow rate when the thrust is expressed in units of weight, is an alternate measure of exhaust velocity. Exhaust velocity is equal to the product of the specific impulse and the earth's gravitational acceleration.

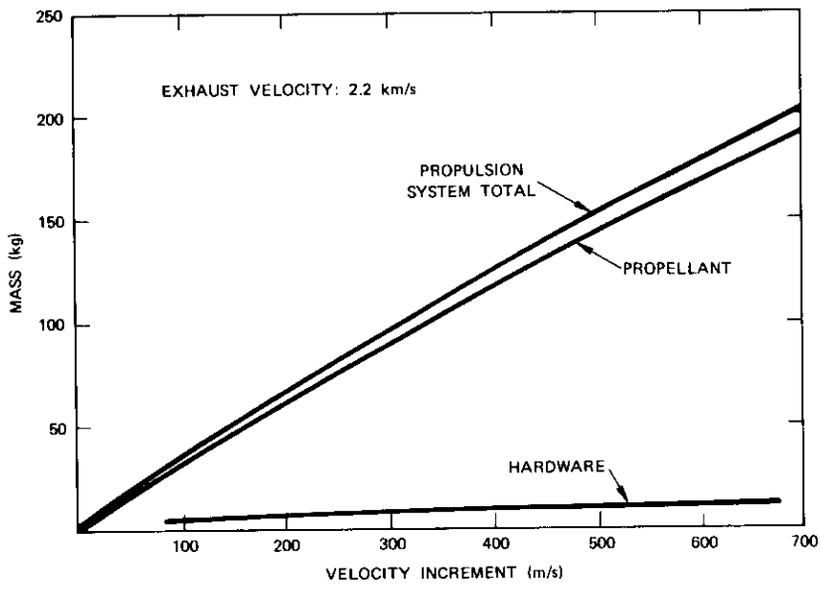


Figure 1. Mass of Hydrazine Propulsion System Components

tially as the exhaust velocity increases, whereas the mass of the power supply increases with exhaust velocity. The slope of the curve for the power supply mass can be gradual (as illustrated) for low-powered propulsion tasks, or much steeper for high-powered rapid maneuvers.

Power Level vs. Exhaust Velocity

For electric propulsion systems, it is desirable to select an exhaust velocity which will not only minimize the propulsion system mass, as shown in Figure 2, but also keep the power level low. The variation of the power/thrust ratio with propellant exhaust velocity is illustrated in Figure 3. The lowest curve (straight line) represents the power/thrust ratio for an ideal electric thruster, i.e., one in which the input power is entirely converted into thrust. The lower shaded area represents the range of performance which has been demonstrated for large (>20-mN thrust) ion thrusters, and the upper shaded area represents the range of performance for ion thrusters which operate at 2 to 10 mN of thrust.

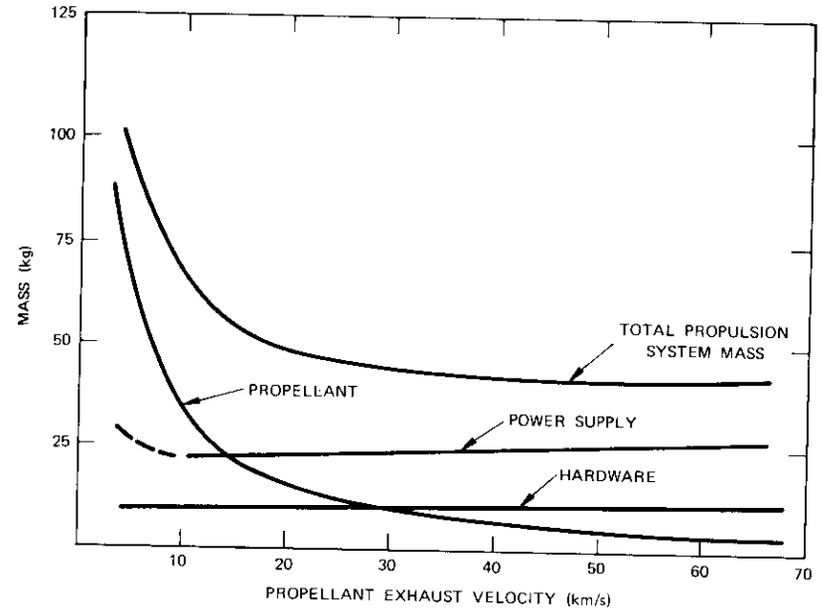


Figure 2. Mass of Electric Propulsion System Components

In the cases shown, there is a power penalty associated with the high propellant exhaust velocity which is intrinsic to the process of converting electric power to thrust. Real thrusters have additional power losses (associated with heating, ionization, electron emission, etc.) which do not produce thrust. At high exhaust velocity these losses are small relative to the power used to accelerate the propellant, and the operating curves tend toward the ideal case. At low exhaust velocity these losses are dominant; hence the slope of the curves is reversed.

It is apparent from Figure 3 that small ion thrusters are less efficient in converting input power to thrust than large ion thrusters; consequently, small ion thrusters will generally operate at a higher propellant exhaust velocity, as shown by the relative positions of the minima.

NORTH-SOUTH STATIONKEEPING

Because of the gravitational disturbances of the sun and moon, a satellite

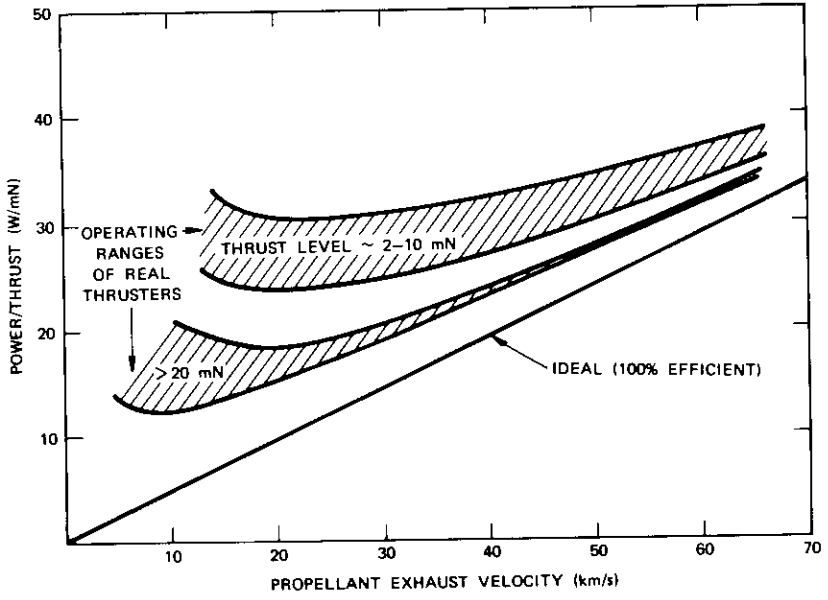


Figure 3. Power/Thrust Ratios for Ion Thrusters

orbit will not remain exactly in the equatorial plane. At synchronous altitude, the orbit inclination with respect to the equator accumulates at an average rate of about 0.8° per year if allowed to go uncorrected. The required correction is equivalent to a velocity increment of about 42 m/s for each year in orbit. At present, the use of hydrazine monopropellant is the accepted way of removing unwanted inclination, a propulsion task commonly referred to as north-south stationkeeping. The technical feasibility of using electric propulsion for north-south stationkeeping is scheduled for testing in the near future on the ATS-F satellite [5] and the Canadian experimental satellite [6]. This section will examine the tradeoffs between hydrazine monopropellant and a mercury ion thruster as applied to north-south stationkeeping of a 700-kg synchronous communications satellite. The most important comparisons involve propulsion system power and mass requirements.

Thrust Program

High- and low-thrust programs for north-south stationkeeping (exaggerated inclination) are shown in Figure 4. When a high-thrust propulsion device, such as hydrazine monopropellant, is used for north-south stationkeeping, the orbit inclination is usually allowed to increase to some prespecified limit, for example, 0.1°. When this limit is reached, one or more short, high-thrust pulses are activated at one of the points (or nodes) where the inclined orbit crosses the equatorial plane. The direction of effective

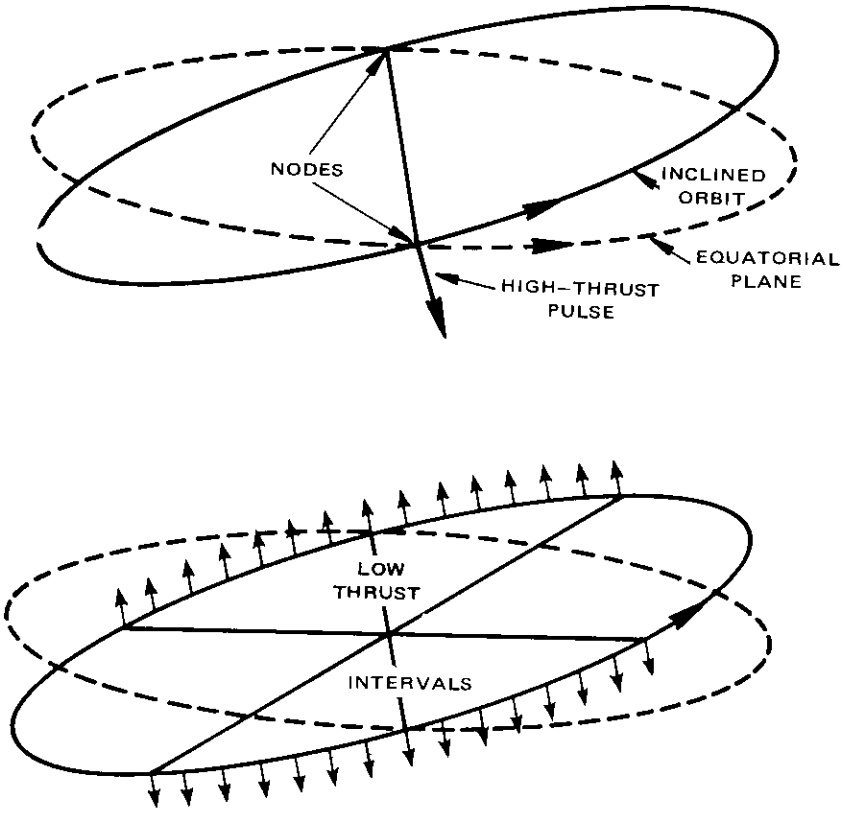


Figure 4. Thrust Programs for North-South Stationkeeping: a. High Thrust, b. Low Thrust

thrust is normal to the orbit plane, pointing north if the descending node is chosen, or south if the ascending node is chosen. The total applied thrust is sufficient to eliminate the unwanted inclination, or even to produce an equal inclination in the opposite direction.

When an electric thruster is used for north-south stationkeeping, it is desirable to minimize both the size of the device and the required power by employing low thrust over a long period of time. This can be accomplished by two long periods of thrust centered on the nodes, as illustrated in Figure 4. It is not desirable to apply the thrust too far away from the nodes because thrust is ineffective (for removing inclination) near the antinodes. For simplicity it is assumed in this paper that an ion thruster is on every day for two periods of about six hours each, except during eclipse. In practice, substantial variations from this schedule are allowable provided that the drift limits are not exceeded.

Thrust Level

For chemical propulsion, the thrust level should be kept low to avoid large torques caused by misalignment of the thrust vector with the center of gravity. Except for such general guidelines, the thrust level is immaterial. Conversely, electric propulsion requires that the thrust level be kept very low to avoid excessive power consumption. If the low thrust associated with the ion thruster were completely effective in removing unwanted inclination, then the thrust program described previously, operating for 275 days each year, would require only 2.5 mN (0.6 mlb) of thrust for north-south stationkeeping of a 700-kg satellite.

In practice, the total generated thrust is never completely effective. First, to avoid interaction between the exhaust plume and other spacecraft components, electric thrusters are often mounted so that the thrust axis passes through the center of gravity, but at a considerable angle relative to the north-south axis. Only the component of thrust normal to the orbit plane,

$$T_N = T \cos \alpha \tag{2}$$

where T is the total thrust and α is the angle which the thrust axis makes with the north-south axis, is at all effective in eliminating unwanted inclination. In this paper, α will be assumed to be 45° so that T is about 40 percent greater than T_N.

Second, as previously mentioned, the process of removing orbit inclination with thrust is most effective at the orbit nodes and not at all

effective at the antinodes. Between the nodes and antinodes, the effective thrust, T_e, is equal to the normal component of thrust, T_N, multiplied by the cosine of the angle, θ, between the line of nodes and the satellite position vector. The mean effective thrust is simply

$$T_e = T_N \frac{\int_0^\theta \cos \theta \, d\theta}{\int_0^\theta d\theta} . \tag{3}$$

For the thrust program outlined in the previous section, T_e is equal to $0.9\, T_N$.

For north-south stationkeeping of a 700-kg satellite, application of these two corrections results in a minimum required total thrust of 4 mN for 12 hours of thruster on-time per day. To allow some flexibility in the duration of the thrust period, we will add a margin of 25 percent and specify a total thrust level of 5 mN.

Selection of Exhaust Velocity

For an electric propulsion task, a suitable exhaust velocity is nearly always chosen on the basis of a tradeoff between power and propulsion system mass. For a thrust level of 5 mN, the power part of the tradeoff is supplied by the upper shaded area of Figure 3. The lower boundary of this area represents the best performance demonstrated to date. In terms of power it is obviously preferable to choose an exhaust velocity near the shallow minimum of this boundary, i.e., about 20 km/s. Note, however, that the use of exhaust velocities up to about 50 km/s entails only moderate increases in power.

To obtain the second part of the tradeoff, i.e., propulsion system mass, curves similar to those in Figure 2 must be constructed. The propellant mass can be calculated by using equation (1) and adding 60 percent to compensate for the thrust inefficiencies discussed in the previous section. A generous allowance for the mass of a propellant tank and feed system is 10 percent of the propellant mass. The mass of the power supply is estimated by applying a factor of 10 W/kg* to the power requirements indi-

* This value is a compromise. For near-term cylindrical spinning satellites, a value of about 6 W/kg should be used. For oriented arrays, 15 to 40 W/kg is an appropriate range, depending on the size and design of the array.

cated in Figure 3. The thruster mass and the power conditioning and control system mass are taken from Reference [5]. Since two thrusters are needed (one at each node) for north-south stationkeeping with electric propulsion, four thrusters are included in the redundant propulsion system. These estimates are summarized in Table 1 and in Figure 5, which shows only a slight variation in propulsion system mass for the range of exhaust velocity from 30 to 60 km/s.

When the power curves of Figure 3 are used in combination with the mass curves of Figure 5, it can be seen that the preferred range of exhaust velocity is 30 to 50 km/s. Power-limited satellites would tend to operate near 30 km/s, whereas weight-limited satellites would tend toward 50 km/s. For this study an exhaust velocity of 40 km/s is used for north-south stationkeeping with electric propulsion.

Mass Savings and Increased Capacity

Once the exhaust velocity for the electric propulsion system has been

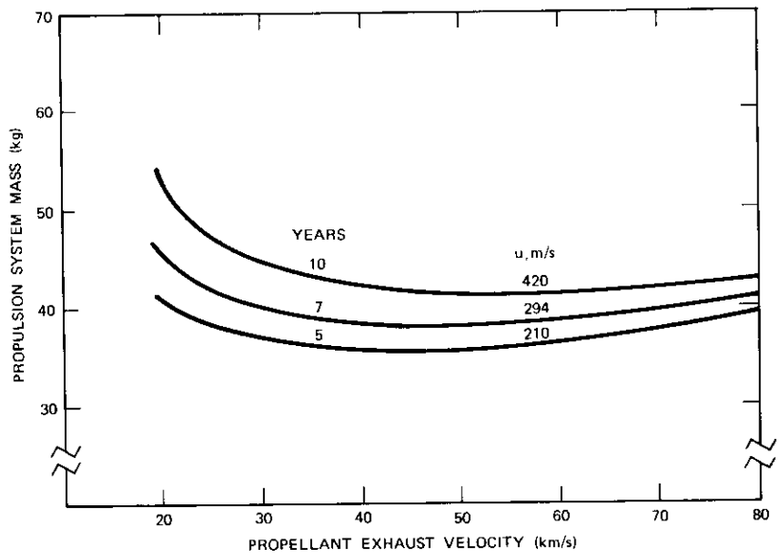


Figure 5. Electric Propulsion System Mass for North-South Stationkeeping

TABLE 1. ELECTRIC PROPULSION SYSTEM MASS

Exhaust Velocity (km/s)	Years in Orbit	Velocity Increment (m/s)	Power (W)	Propulsion System Component Mass (kg)					
				Power Supply at 10 W/kg	Propellant [a]	Tank and Feed Lines	Power Conditioning and Controls [b]	Electric Thrusters [b] (4)	Total
20	5	210	120	12.0	11.8	1.2	8	8	41.0
	7	294	120	12.0	16.4	1.6	8	8	46.0
	10	420	120	12.0	23.5	2.4	8	8	54.0
40	5	210	135	13.5	5.8	0.6	8	8	35.9
	7	294	135	13.5	8.3	0.8	8	8	38.6
	10	420	135	13.5	11.8	1.2	8	8	42.5
60	5	210	170	17.0	3.9	0.4	8	8	37.3
	7	294	170	17.0	5.5	0.6	8	8	39.1
	10	420	170	17.0	7.8	0.8	8	8	41.6
80	5	210	204	20.4	2.9	0.3	8	8	39.6
	7	294	204	20.4	4.1	0.4	8	8	40.9
	10	420	204	20.4	5.9	0.6	8	8	42.9

[a] Includes corrections for thrust inefficiency.
[b] References [5] and [6] supply the background for these estimates.

assigned, it is an easy matter to compare the electric propulsion system mass with that of hydrazine by using Figures 1 and 5. The results are given in Table 2 and Figure 6, which show significant savings in propulsion system mass for long-term stationkeeping.

TABLE 2. MASS REDUCTION ACHIEVED BY USING ELECTRIC PROPULSION FOR NORTH-SOUTH STATIONKEEPING

Years in Orbit	Propulsion System Mass (kg)		Reduction Achieved by Using Electric Propulsion	
	Hydrazine	Electric	Mass (kg)	Percent of Total Satellite
3	43	34	9	1
5	70	36	34	5
7	95	38	57	8
10	130	42	88	13

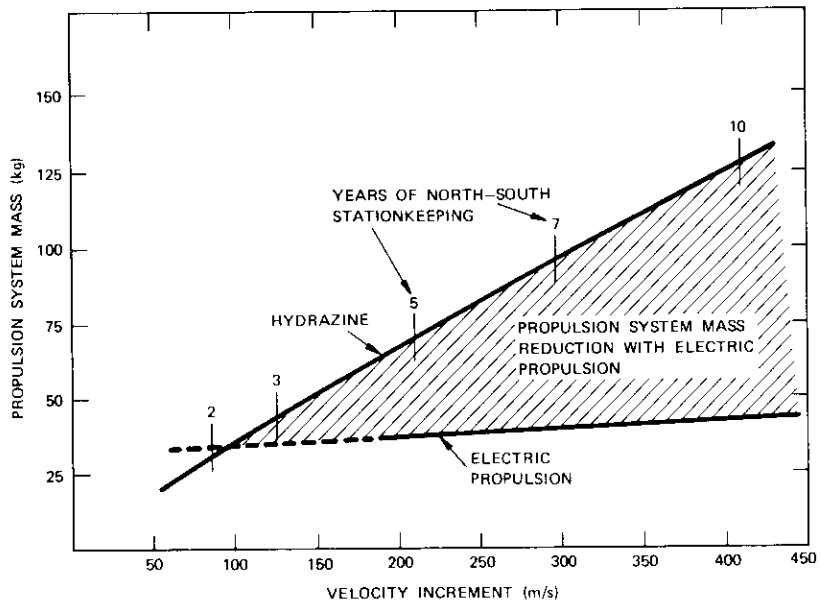


Figure 6. Comparison of Hydrazine and Electric Propulsion System Masses

Since the use of electric propulsion does not entail any additional structural mass, these propulsion system mass reductions can be applied directly to increasing the capacity of the satellite. The mass allocated to power and communications on a typical satellite of this size (e.g., INTELSAT IV) is about 40 percent of the total satellite mass. In those cases where the useful satellite capacity is a linear function of the mass of the power and communications equipment, the mass savings predicted for 5, 7, and 10 years of north-south stationkeeping can be used to increase satellite capacity by 12, 20, and 32 percent, respectively.

Development of a flight-qualified ion thruster system for this mission would require a minimum of two years at a cost of about $2.5 M. The recurring hardware cost per satellite would be an order of magnitude less than this value. If the development cost were spread over about 10 satellites, the prorated cost of the electric propulsion system would be $500 K per satellite.

REPOSITIONING

During the operating lifetime of a satellite, growth or shift in communications traffic occasionally calls for relocation of the satellite. For a three-satellite global system, this relocation, or repositioning, is usually across 120° of longitude. This maneuver can be accomplished with either hydrazine or electric propulsion by using a single thruster aligned with the roll axis. If electric propulsion is used, the thruster should be large enough to minimize the maneuver time by using all of the available satellite power which is not used for traffic during the maneuver. For both propulsion systems, the amount of propellant required rises sharply as the maneuver time is decreased.

For chemical propulsion with hydrazine, electric power is required to perform only minor functions; hence, it is ignored in these calculations. The main relationship of interest is the tradeoff between maneuver time and propellant consumption. For electric propulsion, the propellant consumption is less important, and the pertinent tradeoff is between power and maneuver time.

Large ion thrusters suitable for repositioning are considerably more efficient than those used for north-south stationkeeping; the effect of this increased efficiency is to lower the power/thrust ratio at any given propellant exhaust velocity. This effect can be seen in Figure 3, which also shows that for large thrusters the minima of the curves are shifted to lower exhaust velocities. For this type of propulsion, the range of exhaust velocity

which gives the best balance among propellant mass, power, and maneuver time for repositioning is 20 to 25 km/s. A value of 22 km/s has been chosen for the following comparison with hydrazine, which has an exhaust velocity of about 2.2 km/s. As mentioned previously, the propulsion task chosen to illustrate the tradeoffs for this kind of maneuver is repositioning a 700-kg satellite over 120° of longitude.

Thrust Program

It is assumed that the thrust vector for both impulsive high thrust and continuous low thrust is tangent to the orbit. The thrust program for the hydrazine system consists of two high-thrust pulses; one injects the satellite into an elliptical orbit with a slightly changed period, and the second recircularizes the orbit after the required longitudinal drift has occurred. The thrust program for electric propulsion consists of a small tangential thrust during the entire maneuver, with the thrust direction reversed at the half-way mark.

The two thrust programs are qualitatively illustrated in Figure 7. The solid lines show when and how the velocity increment accumulates as a function of maneuver time. Note that the total velocity increment required for continuous electric propulsion is twice that required for the impulsive high-thrust program. The dashed lines, which are mirror images of the velocity increment curves after the point of thrust reversal, show that the rate of drift is affected by thrust reversal, while the velocity increment is not. The longitudinal drift is represented by the shaded areas.

Repositioning with Hydrazine

The period, P, of a satellite is given by

$$P = \frac{2\pi a^{3/2}}{\mu} \tag{4}$$

where μ is the product of the gravitational constant and the mass of the earth ($3.986 \times 10^{14}\ m^3/s^2$), and a is the semi-major axis of the orbit. For a circular orbit at synchronous altitude, the linear velocity is 3072 m/s; if an impulsive thrust is applied in the direction of motion, the instantaneous velocity at the point of application is $3072 + u$ (the velocity increment). This point in the orbit is the perigee of a perturbed, elliptical orbit, and the radius at perigee, r_p, is equal to the synchronous radius. For this new

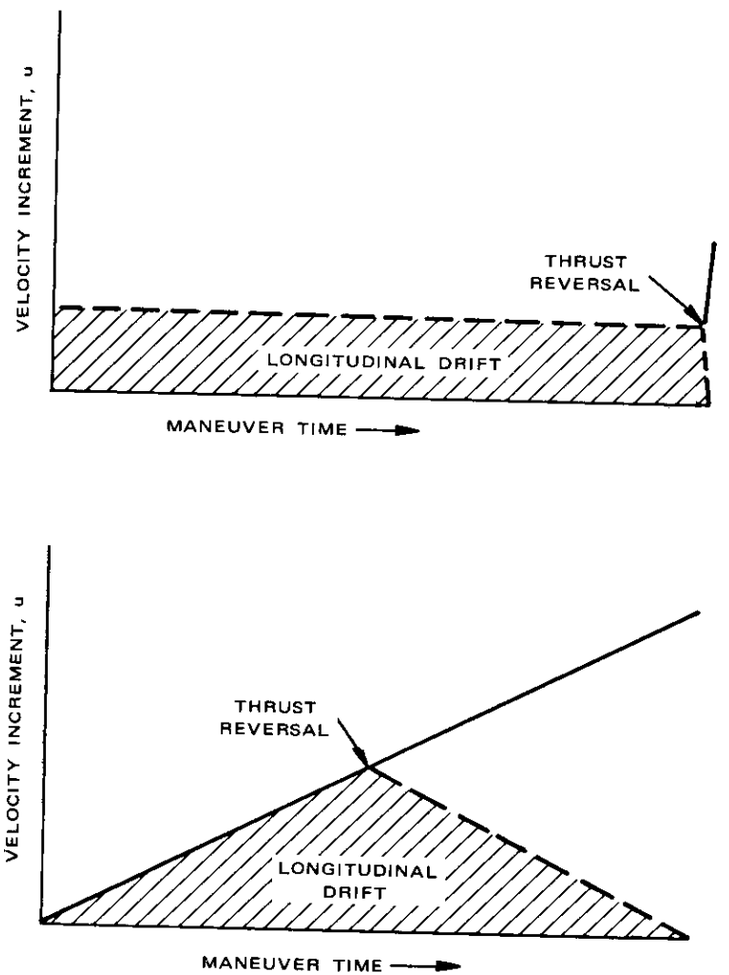


Figure 7. Velocity Increment and Longitudinal Drift for High- and Low-Thrust Propulsion: a. High Thrust, b. Low Thrust

orbit, the semi-major axis is a function of the perigee radius and the linear velocity at perigee, V_p; i.e.,

$$a = \frac{\mu r_p}{(2\mu - r_p V_p)} . \tag{5}$$

If the periods are given in days, then the average angular velocity of the satellite in the perturbed orbit is

$$\phi = -360\left(\frac{1}{P_s} - \frac{1}{P_e}\right) \tag{6}$$

in degrees per day relative to a point on the earth. The subscripts s and e refer to synchronous and elliptical, respectively. The first negative sign in equation (6) indicates retrograde slippage, which occurs because the perturbed orbit period is longer than the synchronous orbit period when the velocity increment is added to the synchronous orbit velocity.

Substitution of equations (4) and (5) into equation (6) yields

$$\phi = -360\left\{1 - \left[2 - \left(1 + \frac{u}{V_s}\right)^2\right]^{3/2}\right\} \tag{7}$$

in degrees per day. In this equation, the velocity increment, u, is for the initial pulse only, i.e., for injection into the drifting orbit. An equal but opposite pulse is required at the end of the maneuver to terminate the drift. Solutions of this equation and equation (1) for selected values of total velocity increment are given in Table 3.

TABLE 3. LONGITUDE REPOSITIONING OF A 700-KG SATELLITE WITH HYDRAZINE MONOPROPELLANT

u (m/s)	ϕ (deg/day)	Time to Drift 120° (days)	Propellant Mass (kg)
5	0.88	136	1.5
10	1.76	68	3.1
20	3.52	34	6.2
30	5.28	23	9
40	7.04	17	12
50	8.80	14	16
100	17.6	7	31
150	26.4	4.5	46

Repositioning with Electric Propulsion

If the repositioning maneuver is accomplished with a small, continuous, tangential thrust, the satellite orbit may be assumed to be circular at all times without introducing any appreciable error. The linear tangential velocity is given by

$$V_t = V_o - \frac{F}{m} t \tag{8}$$

where V_o is the initial linear velocity in synchronous orbit and F/m is the thrust-to-mass ratio, or acceleration. The acceleration may be assumed to be constant, since the satellite mass will change little during the maneuver. The instantaneous orbit radius is a function of the square of the tangential velocity; i.e.,

$$R = \frac{\mu}{V_t^2} \tag{9}$$

and, in terms of equations (8) and (9), the angular velocity, ω, is

$$\omega = \frac{d\theta}{dt} = \frac{1}{\mu}\left(V_o - \frac{F}{m} t\right)^3 . \tag{10}$$

The total angular motion, θ_t, is obtained by integrating equation (10) over time:

$$\theta_t = \frac{1}{\mu}\int_0^t \left(V_o - \frac{F}{m} t\right)^3 dt . \tag{11}$$

On the other hand, the angular motion of a synchronous satellite is merely $\theta_s = \omega_s t$. The difference between this motion and the motion described by equation (11) is equal to the displacement of a satellite relative to a point on earth after a period of time, t; i.e.,

$$\theta_s - \theta_t = \frac{V_s}{R_s} t - \frac{1}{\mu}\int_0^t \left(V_o - \frac{F}{m} t\right)^3 dt. \tag{12}$$

Integration of equation (12) for a displacement of 60° gives

$$\frac{\pi}{3} = \frac{V_s}{R_s} t + \frac{1}{\mu}\left\{\frac{[V_o - (F/m)t]^4}{4F/m} - V_o^4\right\} . \tag{13}$$

If Ft/m is much smaller than V_s, which is the case in all but unrealistically

rapid maneuvers, then since $V_o = V_s$ and $\mu = R_s V_s$, equation (13) can be expanded and simplified as follows:

$$\frac{2\pi}{9} = R_s \left(\frac{F}{m} t\right) t . \tag{14}$$

In this equation, t is the time required to complete half of the maneuver, and the product Ft/m is equal to the velocity increment u for half of the maneuver.

Compatible values of u, t, and propellant mass for the whole maneuver, calculated from equations (1) and (14), are given in the first three columns of Table 4. Values for the electric power, shown in the fourth column of Table 4, are calculated as follows. Dividing the velocity increment by the thrusting time results in the thrust-to-mass ratio. Since the satellite mass is known, the thrust is easily calculated. The power/thrust ratio for thrust levels exceeding 45 mN (10 mlb) is about 16 W/mN (71 W/mlb) for advanced ion thrusters operating at a specific impulse of 2200 seconds [7]. At a thrust level of 4 to 5 mN (1 mlb), about 140 W of power are required [5], and for 2 mN of thrust about 75 W of power are required [6]. For intermediate thrust levels, values for the required power are obtained by interpolating between these known values.

Results from Tables 3 and 4 are presented graphically in Figures 8 and 9. Figure 8 illustrates the propellant consumption for both forms of propulsion versus the maneuver time. Were this the only consideration, electric propulsion would be clearly superior over the entire range of

FIGURE 4. LONGITUDE REPOSITIONING OF A 700-KG SATELLITE WITH LOW-THRUST ELECTRIC PROPULSION

u Total (m/s)	Time to Drift 120° (days)	Mass of Mercury Propellant (kg)	Power (W)	Thrust (mN)
20	68.6	0.6	75	2
40	34.3	1.2	210	10
60	23.0	1.8	380	21
80	17.2	2.4	600	36
100	13.8	3	910	57
200	7.0	6	3,600	223
400	3.6	12	14,000	875

maneuver time. However, Figure 9 shows that the use of electric propulsion when the maneuver time is short results in excessive power requirements.

Obviously, short maneuver times are detrimental to both forms of propulsion. For chemical propulsion, maneuver times of less than about 20 days result in increasingly heavy consumption of onboard propellant. Nevertheless, the primary advantage of chemical propulsion for this maneuver is that emergency maneuvers can be performed rapidly. For electric propulsion the propellant mass requirements are modest even for rapid maneuver times, but the maneuver time is restricted by the on-board power. Lower limits on the maneuver time are indicated in Figure 9.

Mass of Propulsion Systems

Since the magnitude of the maneuver discussed previously (and other similar maneuvers) depends on the time available to perform the mission,

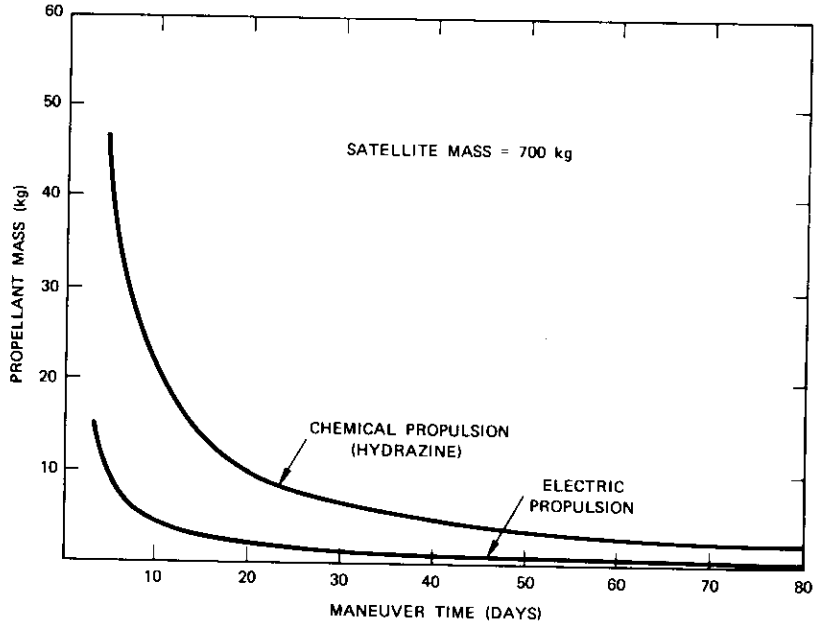


Figure 8. Propellant Consumption of Hydrazine and Electric Propulsion for a 120° Longitude Change

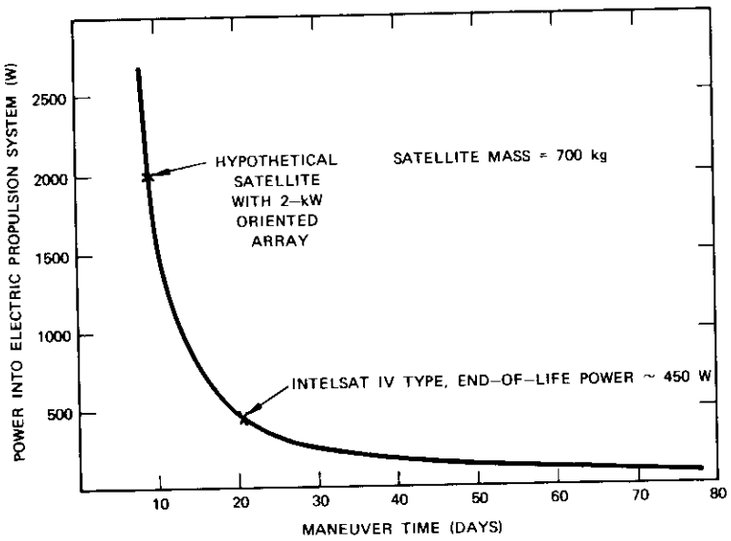


Figure 9. Electric Power vs. Maneuver Time for a 120° Longitude Change with Electric Propulsion

it is difficult to assign propulsion requirements. For comparison purposes, it is assumed that, when hydrazine is used, a total velocity increment of 125 m/s, which is roughly equivalent to two 10-day longitude changes of 120°, will be provided. For low-thrust electric propulsion, a velocity increment of 250 m/s will be needed for the same task (see Figure 7). These values can be used to calculate propellant mass from equation (1) for both hydrazine and mercury. The results are given in Table 5, which also includes estimates of the mass of propulsion system hardware.

The 23-kg reduction in propulsion system mass achieved by using the mercury ion thruster is equal to 3.3 percent of the entire satellite mass. By the same reasoning which was discussed in the section on north-south stationkeeping, this represents a potential increase in satellite capacity of about 8 percent. Despite the use of a larger thruster, the prorated cost of the electric thruster system for this maneuver would probably be less than the cost of a north-south stationkeeping thruster system, since repositioning requires only one thruster per satellite.

TABLE 5. PROPULSION SYSTEM COMPONENT WEIGHTS FOR SATELLITE REPOSITIONING

Component	Weight (kg)	
	Hydrazine Monopropellant	Mercury Ion Thruster
Propellant	38.7	8.5
Storage Tank and Feed System	4.6	1.3
Thruster (1)	0.5	3.0
Power Conditioning and Controls	—	8.0
TOTAL	43.8	20.8

FEASIBILITY OF USING ELECTRIC PROPULSION

In terms of efficiency, i.e., in converting electric power to thrust, current ion thruster technology is more than adequate for both north-south stationkeeping and repositioning. Furthermore, laboratory endurance tests have established a lifetime of over 10,000 hours (continuous) and a capability of at least several hundred restarts. The longest flight test of an ion thruster to date is the SERT-2 (space electric rocket test 2), which demonstrated several months of cumulative operation and several cold restarts in 1970 [8].

Repositioning maneuvers require only a few restarts and a total operating time not exceeding a few months. Since these characteristics have already been demonstrated on the SERT-2 spacecraft, the application of electric propulsion to repositioning does not require new technology.

North-south stationkeeping is a more difficult job in terms of propulsion system reliability and lifetime. About 3000 restarts and a cumulative on-time of 18,000 hours are required for 10 years of north-south stationkeeping. These combined capabilities have not yet been demonstrated either in space or on the ground. However, in the past few years major improvements [7], [9], [10] have been made in the state of the art, and ion thrusters based on the best technology available could be designed to surpass these requirements. Brief north-south stationkeeping experiments using ion thrusters [5], [6] are already scheduled for the ATS-F and the Canadian experimental satellite, but more extensive ground and space endurance tests must be completed before electric propulsion can be used for long-term applications.

CONCLUSIONS

The use of electric propulsion for long-term north-south stationkeeping and satellite repositioning appears promising, since it substantially reduces the propulsion system mass. Based on presently available ion thruster performance, combined weight savings equal to 16 percent of the entire satellite mass are predicted for a 10-year satellite. These weight savings could produce a 40-percent increase in communications capacity. The major task remaining before electric propulsion can be applied commercially is demonstration of the long-term reliability needed for north-south stationkeeping.

REFERENCES

[1] B. Free, "Satellite Raising to Synchronous Orbit, an Evaluation of the Hydrogen Resistance Jet and the Mercury Bombardment Ion Thruster," *Journal of the British Interplanetary Society*, Vol. 23, No. 10, October 1970, pp. 669–690.

[2] K. I. Duck, R. O. Bartlett, and R. J. Sullivan, "Evaluation of an Ion Propulsion System for a Synchronous Spacecraft Mission," *AIAA Electric Propulsion and Plasmadynamics Conference*, Colorado Springs, September 1967, AIAA Paper No. 67–720.

[3] W. Guman and D. Nathanson, "Pulsed Plasma Microthruster Propulsion System for Synchronous Orbit Satellite," *AIAA 7th Electric Propulsion Conference*, Williamsburg, March 1969, AIAA Paper No. 69–298.

[4] W. Cutler, W. Owens, and W. Coleal, "Electric Propulsion Application to Large Long-Life Communication Satellites," *AIAA 3rd Communications Satellite Systems Conference*, Los Angeles, April 1970, AIAA Paper No. 70–478.

[5] E. James *et al.*, "A One Millipound Cesium Ion Thruster System," *AIAA 8th Electric Propulsion Conference*, Stanford, August-September 1970, AIAA Paper No. 70–1149.

[6] P. D. Reader *et al.*, "A Submillipound Mercury Electron Bombardment Thruster," *Journal of Spacecraft and Rockets*, Vol. 7, No. 11, November 1970, pp. 1287–1292.

[7] W. Knauer, R. Poeschel, and J. Ward, "The Radial Field Kaufman Thruster," *AIAA 7th Electric Propulsion Conference*, Williamsburg, March 1969, AIAA Paper No. 69–259.

[8] W. R. Kerslake *et al.*, "Flight and Ground Performance of the SERT II Thruster," *AIAA 8th Electric Propulsion Conference*, Stanford, August-September 1970, AIAA Paper No. 70–1125.

[9] R. Moore, "Magneto-Electrostatically Contained Plasma Ion Thruster," *AAIA 7th Electric Propulsion Conference*, Williamsburg, March 1969, AIAA Paper No. 69–260.

[10] W. D. Ramsey, "12 Centimeter Magneto-Electrostatic Containment Mercury Ion Thruster Development," *AIAA/SAE 7th Propulsion Joint Specialist Conference*, Salt Lake City, June 1971, AIAA Paper No. 71–692.

ACKNOWLEDGMENT

The author would like to thank Dr. Gary Gordon, Dr. Edmund Rittner, and Mr. Joachim Kaiser for their helpful suggestons.

Bernard A. Free received a B.S. degree in chemistry from Seton Hall University in 1950 and an M.S. degree in theoretical chemistry from the University of Cincinnati in 1965. He has worked in the areas of analytical chemistry, high-energy chemical propellants, and several forms of electric propulsion. He joined COMSAT Laboratories in 1969 to identify and promote the application of electric propulsion in communications satellite technology.

Index: communication satellites, demodulation, frequency modulation, intermodulation

Baseband distortion caused by intermodulation in multicarrier FM systems

N. K. M. Chitre and J. C. Fuenzalida

ABSTRACT

In communications satellite systems a common nonlinear amplifier is used by a number of angle-modulated carriers. The RF intermodulation products generated by the amplifier have the characteristics of angle-modulated carriers and appear as baseband distortion (noise) at the output of the phase demodulator after each carrier is demodulated. It has been customary to assume that these products are statistically independent of the carriers when computing the baseband distortion. This paper negates that assumption and considers the statistical dependence of the signals modulating the intermodulation product and the carriers. The relative phase of the intermodulation products with respect to the demodulated carrier has also been taken into account. The levels and spectral distributions of the baseband distortion obtained by using this approach are shown to be quite different from those obtained by using the conventional approach.

A heuristic explanation using the modulation or envelope gain approach and based only on third-order RF intermodulation products is given. A set of experiments which confirm the validity of this approach is described. Finally, the impact on present and future satellite utilization is discussed.

This paper is based upon work performed at COMSAT Laboratories under the sponsorship of the International Telecommunications Satellite Consortium (INTELSAT). Any views expressed in this paper are not necessarily those of INTELSAT.

INTRODUCTION

A widely used multiple-access method in satellite communications is frequency-division multiple access (FDMA). In this method, a common nonlinear amplifier is used simultaneously by a number of independently modulated carriers. When the carriers are phase demodulated, the odd-order intermodulation products generated by the amplifier create signal impairment, which is usually called baseband distortion or noise. To control this impairment, it is customary to operate the amplifiers at less than the rated power with a consequent reduction of system efficiency. This is particularly significant in the case of a satellite, since the available power has a great impact on the maximum traffic capacity.

The present generation of communications satellites (INTELSAT IV) is equipped with a larger number of transponders, each having a smaller bandwidth than transponders on previous satellites. The average number of carriers per transponder is therefore considerably smaller, and the techniques which have been used for estimating the distortion caused by intermodulation are no longer sufficiently accurate.

The study of the impairments resulting from the use of nonlinear amplifiers in the operation of multicarrier angle-modulated systems can be divided into three stages:

a. characterization of the device or amplifier nonlinearity,

b. computation of the magnitude and phase of the RF intermodulation products generated by the preceding characterization, and

c. prediction of the baseband distortion caused by the presence of RF intermodulation products.

The nonlinear voltage transfer characteristics of devices have been expressed as ideal limiters [1], as polynomial expansions [2], or as Fourier series [3]. These voltage transfer characterizations have been used to estimate nonlinear amplitude distortion. However, the output phase of the signal in a nonlinear bandpass device such as a TWTA may also be a nonlinear function of the input envelope power. This nonlinear phase transfer has been characterized as a polynomial expansion [1] and also as a particular function requiring only three coefficients [2]. A general and mathematically rigorous method, in which the complex output envelope is expressed as a function of the input envelope, has been presented by Shimbo [4].

The nonlinear phase characterization may be visualized as a nonlinear phase modulation of the input signal, in which the amount of phase modulation is a nonlinear function of the instantaneous input envelope power.

Similarly, the combined effects of the nonlinear amplitude and phase characteristics of a bandpass device can be visualized as a complex nonlinear modulation process in which the input signals are modulated in both amplitude and phase. This modulation or envelope gain approach will be presented in the next section.

As an extension of the device characterization, Standing [5] and Shimbo [4] have reported methods to reduce the RF level of intermodulation products relative to the carrier level.

Device characterization reduces the continuous nonlinear transfer characteristics to a finite set of coefficients. A set of equations is then developed to express the center frequency and the output level of each RF intermodulation product as functions of these coefficients and the input levels of the carriers. Such sets are available for the referenced characterizations and can be used to predict the RF output level. Except in Shimbo's paper [4], the effects of amplitude and phase nonlinearities are treated separately, causing cross-product terms to be neglected.

For the sake of simplicity and ease of understanding, a heuristic explanation using the modulation or envelope gain approach and based only on third-order intermodulation products is presented. In this approach all intermodulation products are visualized as modulation sidebands of input carriers which are modulated by all possible instantaneous difference frequencies. Each individual intermodulation product can be visualized as a carrier angle modulated by an algebraic combination of the modulating signals of the carriers which generated it.

The last vital link in the prediction of communications system performance is the calculation of baseband distortion caused by the presence of in-band RF intermodulation products. Baseband distortion predictions have conventionally been based on two assumptions:

a. that the RF intermodulation products are statistically independent of the generating carriers and that they can therefore be replaced by RF noise of equivalent spectral density when calculating the baseband noise, and

b. that the equivalent RF power spectral density has a Gaussian shape. These assumptions make it possible to combine the intermodulation products to generate an equivalent "colored" noise spectral density at RF.

The purpose of this paper is to present a different view of intermodulation distortion as it appears at baseband. Hence, the method used in this paper to calculate baseband distortion introduces several important modifications to conventional procedure:

a. the intermodulation products are treated as interfering carriers,

b. the statistical dependence of the signals modulating the intermodulation products and the carrier being demodulated is taken into account, and

c. the relative phase between the intermodulation products which appear as symmetrical modulation sidebands on the carrier being demodulated is taken into account.

It is shown that the level and spectral shape of the baseband distortion cannot generally be predicted from the RF level and spectral density of an intermodulation product. The dependence of the level and spectral shape of the baseband distortion on the statistical relationship between the modulating signals of the intermodulation products and the carrier being demodulated, and on the relative phase between the sidebands and the carrier has been experimentally confirmed. A more accurate representation of the power spectral densities of both the RF intermodulation products and the baseband distortion caused by these products is also presented.

The approach presented in this paper can sometimes lead to a prediction of the baseband distortion which is entirely different from the predictions obtained by using approaches based on conventional assumptions even though the starting point, i.e., the level and spectrum of the RF intermodulation products, is the same in both cases. A mathematical model which includes the concepts developed here has been generated to predict the baseband distortion in the operation of a multicarrier FM system. The predictions of this model lead to a closer agreement with measured values of baseband distortion. Finally, the impact of this new approach and the improved model on present and future satellite utilization is discussed.

CHARACTERISTICS OF INTERMODULATION PRODUCTS

Nonlinear amplification can be visualized as a complex nonlinear modulation process, as shown in Figure 1. Let the input and output bandpass signals be given by

$$e_1(t) = Re[V_1(t) \exp(j\omega_0 t)] \tag{1}$$

$$e_2(t) = Re[V_2(t) \exp(j\omega_0 t)] \tag{2}$$

where $V_1(t)$ and $V_2(t)$ are the complex envelopes related by the complex envelope power-dependent gain function, $G(P)$. That is,

$$V_2(t) = G[P(t)]V_1(t) \tag{3}$$

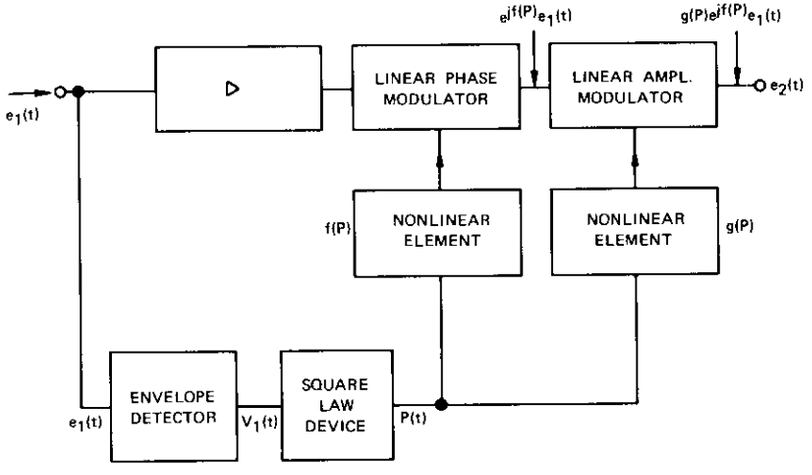


Figure 1. Equivalent Circuit of a Nonlinear Amplifier

where $$G(P) = g(P) \exp[jf(P)] \tag{4}$$

and $$P(t) = \tfrac{1}{2}V_1(t) \cdot V_1^*(t). \tag{5}$$

When the input consists of N angle-modulated carriers,

$$V_1(t) = \sum_{l=1}^{N} A_l \exp(j\theta_l) \tag{6a}$$

$$\theta_l = \omega_l t + \psi_l(t) + \mu_l \tag{6b}$$

and $$P(t) = P_1 + \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} A_j A_k \cos(\theta_j - \theta_k) \tag{7a}$$

$$P_1 = \tfrac{1}{2} \sum_{l=1}^{N} A_l^2 \tag{7b}$$

where P_1 is the total input power.

Although $G(P)$ is a constant for a linear device, it is a function of the envelope power, P, for a nonlinear device. When the envelope power is

varied with time, the gain function will also be a function of time and, for a nonlinear device, it will contain not only the frequency components in $P(t)$, but also higher order difference frequencies. $G[P(t)]$ may therefore be expanded as a series whose first two terms are

$$G[P(t)] \simeq z_0 + \sum_{j=1}^{N-1} \sum_{k=1 \neq j}^{N} z_{jk} \exp [j(\theta_j - \theta_k)]. \quad (8)$$

Even though equation (8) has the same format as equation (7a), it cannot be assumed that there is a linear relationship between $G[P(t)]$ and $P(t)$, since the coefficients (z's) in equation (8) are derived by using all of the terms in the particular expansion of $G(P)$ as a function of P.

The third-order intermodulation products (i.e., those products whose frequencies are an algebraic combination of three carrier frequencies) are by far the largest; therefore, they are the only significant source of distortion below saturation. The higher order terms neglected in equation (8) contribute to the generation of the fifth- and higher-order intermodulation products whose effects become more significant near saturation.

The coefficients (z's) of equation (8) will generally be complex quantities and functions of the amplitudes of the input carriers. Substituting equations (6) and (8) into equation (3) yields

$$V_2(t) \simeq \sum_{i=1}^{N} \left\{ z_0 + \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} z_{jk} \exp [j(\theta_j - \theta_k)] \right\} A_i \exp (j\theta_i). \quad (9)$$

This representation of the output voltage may seem to indicate that all carriers are compressed equally and that their levels remain unchanged. However, this is not the case, since the actual signal voltage at a carrier frequency ω_l will have additional components caused by sidebands of the type $\omega_k + (\omega_l - \omega_k)$, which are collocated and coherent with the carrier l. Similarly, products of the type $\omega_i + \omega_j - \omega_k$ can be generated by modulating the i-th and j-th carriers.

Figure 2, which is a symbolic representation of operation with three carriers per transponder, illustrates the way in which the output carriers and intermodulation products are formed. For simplicity, equal carriers are used and the phase effects are not shown. Each carrier is modulated by the three difference frequencies (as shown in the three middle rows) and the resultant output of each carrier includes some coherent sidebands. This diagram also helps to explain why intermodulation products of the type $\omega_1 + \omega_2 - \omega_3$ have twice the voltage of those of the type $2\omega_1 - \omega_2$.

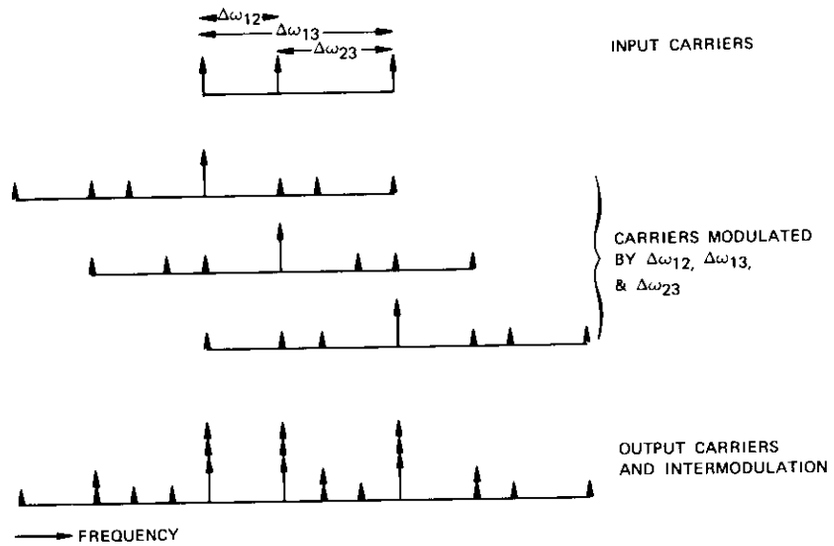


Figure 2. Output Expressed as a Sum of Modulated Carriers

Equation (9) can be rewritten as a sum of carrier frequency terms and third-order intermodulation product terms:

$$V_2(t) \simeq \sum_{l=1}^{N} H_l \exp (j\theta_l) + \sum_{j=1}^{N} \sum_{k=1 \neq j}^{N} H_{jk} \exp [j(2\theta_j - \theta_k)] + \sum_{i=1}^{N-1} \sum_{j>i}^{N} \sum_{k=1 \neq i \neq j}^{N} H_{ijk} \exp [j(\theta_i + \theta_j - \theta_k)]. \quad (10)$$

Here, H_l, H_{jk}, and H_{ijk} are the complex amplitudes at the output and are related to the input amplitudes through the coefficients (z's); i.e.,

$$H_l = z_0 A_l + \sum_{m=1 \neq l}^{N} z_{lm} A_m \quad (11a)$$

$$H_{jk} = z_{jk} A_j \quad (11b)$$

$$H_{ijk} = z_{jk} A_i + z_{ik} A_j. \quad (11c)$$

By substituting equation (6b) into equation (10), we can see that the general intermodulation product is an angle-modulated carrier given by

$$I_{ijk}(t) = |H_{ijk}| \cos [\omega_0 t + \omega_{ijk} t + \psi_{ijk}(t) + \mu_{ijk}] \tag{12}$$

where

$$\omega_{ijk} = \omega_i + \omega_j - \omega_k \tag{13a}$$

$$\psi_{ijk}(t) = \psi_i(t) + \psi_j(t) - \psi_k(t) \tag{13b}$$

$$\mu_{ijk} = \mu_i + \mu_j - \mu_k + \tan^{-1} \left\{ \frac{Im(H_{ijk})}{Re(H_{ijk})} \right\} \tag{13c}$$

The computation of the magnitude and phase of H_{ijk} is closely related to the particular characterization used for the nonlinear transfer function of the device. This problem has been tackled by Westcott [2] and Berman and Mahle [3] for amplitude and phase nonlinearities, and by Shimbo [4] for the general case using an envelope distortion approach.

BASEBAND DISTORTION PRODUCED BY ANGLE-MODULATED INTERFERENCE

Having shown that the general intermodulation product has the characteristics of an angle-modulated carrier, we use the following results derived from interference theory [6]. As in equations (6a) and (6b), let the wanted and interfering carriers be represented by

$$V_l(t) = A_l \exp (j\theta_l) \tag{14a}$$

$$V_u(t) = A_u \exp (j\theta_u). \tag{14b}$$

It has been shown [6] that the baseband distortion produced by interference which is detected by an ideal phase demodulator is given by

$$\lambda(t) = \sum_{m=1}^{\infty} \lambda_m(t) \tag{15a}$$

where

$$\lambda_m(t) = (-1)^{m+1} \left(\frac{A_u}{A_l} \right)^m \sin \{m[(\omega_u - \omega_l)t + \psi_u(t) - \psi_l(t) + \mu_u - \mu_l]\}. \tag{15b}$$

The baseband distortion therefore consists of phase-modulated subcarriers at harmonics of $\omega_u - \omega_l$ whose amplitudes decrease rapidly with the harmonic number m. Usually the first term is sufficient to describe the distortion produced by a small interfering signal. That is,

$$\lambda(t) \simeq \frac{A_u}{A_l} \sin [(\omega_u - \omega_l)t + \psi_u(t) - \psi_l(t) + \mu_u - \mu_l] \tag{16}$$

which has a power spectrum given by

$$S_\lambda(f) = \frac{1}{4} \left(\frac{A_u}{A_l} \right)^2 [T_{\Delta\psi}(f - \Delta f) + T_{\Delta\psi}(f + \Delta f)] \tag{17}$$

where

$$T_X(f) = \text{power spectrum of a signal sin } x(t) \tag{18a}$$

$$\Delta\psi(t) = \psi_u(t) - \psi_l(t) \tag{18b}$$

$$\Delta f = \frac{1}{2\pi} (\omega_u - \omega_l). \tag{18c}$$

In equation (17), the power spectrum of the distortion is related to the power spectrum of a zero-frequency carrier modulated by the signal $\Delta\psi$. The important conclusion to be derived from this equation is that the spectral shape of the baseband distortion is determined by the difference between the two modulating signals. For Gaussian modulating signals [7],

$$T_X(f) = \int_{-\infty}^{\infty} \exp [R_X(\tau) - R_X(0)] \exp (-j\omega\tau)\, d\tau. \tag{19}$$

For independent Gaussian modulating signals, if

$$x(t) = \sum_{i=1}^{N} a_i x_i(t)$$

then

$$T_X(f) = T_{a_1 x_1}(f) \otimes T_{a_2 x_2}(f) \otimes T_{a_3 x_3}(f) \ldots \otimes T_{a_N x_N}(f) \tag{20}$$

where $T_{a_i x_i}(f) \otimes T_{a_j x_j}(f)$ indicates the convolution of $T_{a_i x_i}$ with $T_{a_j x_j}$. If x_i is wideband Gaussian noise of rms voltage σ_i so that the autocorrelation approximates a delta function and/or the modulation index is high, $T_{a_i x_i}(f)$ can be approximated by a Gaussian spectral shape:

$$T_{a_i x_i}(f) = \frac{1}{\sqrt{2\pi}\, a_i \sigma_i} \exp\left[\frac{-f^2}{2a_i^2\sigma_i^2}\right] \tag{21a}$$

Then

$$T_x(f) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[\frac{-f^2}{2\sigma^2}\right] \tag{21b}$$

where

$$\sigma^2 = \sum_{i=1}^{N} a_i^2 \sigma_i^2 \,. \tag{22}$$

The terms "effective deviation" and "effective power spectral density" will be used to describe σ and $S_\lambda(f)$.

DEMODULATION OF INTERMODULATION PRODUCTS

It has been noted that intermodulation products are generated by modulating the incoming carriers with all possible difference frequencies. Those intermodulation products which can be regarded as sidebands of each carrier and which can therefore appear symmetrically about each carrier are hereafter termed correlated intermodulation products. They have two important properties: their phase relative to the demodulated carrier becomes a significant factor and their detected output after demodulation does not contain the modulating signal of the demodulated carrier.

According to Reference [8], they could be called coherent, but we prefer to reserve that term for the case in which only the carrier of interest is modulated and the intermodulation products are tracking this carrier. If these symmetrical "twins" are in phase with the carrier, they represent amplitude modulation sidebands and hence will not be detected by a phase demodulator. Conversely, if they are in phase quadrature with the carrier, they represent phase modulation sidebands and will add in voltage, not power, at the output of the phase demodulator. To a good approximation, i.e., barring cross product terms, amplitude modulation is generated by nonlinearities in the amplitude transfer characteristics, $g(P)$, and phase modulation is generated by nonlinearities in $f(P)$. Hence, in this special case, the output distortion will contain only terms generated by phase nonlinearity.

In general, if the l-th carrier and an intermodulation product labeled ijk are fed to an ideal phase demodulator, the baseband distortion term will be given by

$$\lambda(t) = \Delta H \sin\left[\Delta\omega t + \Delta\psi(t) + \Delta\mu\right] \tag{23}$$

where

$$\Delta H = \frac{|H_{ijk}|}{|H_l|} \tag{24a}$$

$$\Delta\omega = \omega_i + \omega_j - \omega_k - \omega_l \tag{24b}$$

$$\Delta\psi(t) = \psi_i(t) + \psi_j(t) - \psi_k(t) - \psi_l(t) \tag{24c}$$

$$\Delta\mu = \mu_i + \mu_j - \mu_k - \mu_l + \tan^{-1}\left\{\frac{Im(H_{ijk})}{Re(H_{ijk})}\right\} \tag{24d}$$

On the basis of equation (24) and the concepts developed in the previous paragraphs, the intermodulation products can be classified with respect to a wanted carrier l as shown in Table 1. This table also gives the effective power spectral density shape and the rms deviation of its Gaussian approximation, σ.

In the conventional treatment of intermodulation products, the baseband distortion spectra are represented by a convolution of the RF spectra of the carrier and the intermodulation product or, in our terminology,

$$T_{\Delta\psi}(f) = T_{\psi_i}(f) \otimes T_{\psi_j}(f) \otimes T_{\psi_k}(f) \otimes T_{\psi_l}(f) \tag{25a}$$

and

$$\sigma^2_{\Delta\psi} = \sigma_i^2 + \sigma_j^2 + \sigma_k^2 + \sigma_l^2 \,. \tag{25b}$$

A comparison of the previous equation with Table 1 clearly shows that only the products of the first category fit into the conventional treatment and can be considered truly independent. The products in the second category can be presumed to be modulated by the same signal as on carrier l, but in the opposite sense, so that the instantaneous frequency is moved in the opposite direction. This has the effect of spreading the baseband spectrum, as indicated by the $T_{2\psi_l}(f)$ and $4\sigma_l^2$ terms in the effective power spectrum.

The most interesting products are those of category 3. Since they are correlated (i.e., the absence of the modulation signal, ψ_l, of the carrier l from the baseband distortion term leads to a decreased effective deviation in the baseband) and symmetrical, only the phase quadrature component appears at the baseband output.

In the first two categories, the baseband distortion power is directly related to the intermodulation product level, but only in the first category can the baseband power spectrum be predicted from the RF power spectra of the carrier and the intermodulation product. In the second and

TABLE 1. CLASSIFICATION OF INTERMODULATION PRODUCTS
(wanted carrier, l)

Intermodulation Product			Baseband Distortion			
Class	Description	Generic Type	Location in BB, $\Delta\omega$	BB Modulating Signal, $\Delta\psi(t)$	BB Spectrum, $T\Delta\psi(f)$	BB Effective Deviation, $\sigma^2_{\Delta\psi}$
1a	Asymmetrical and Independent	$i = j \neq k \neq l$	$\omega_i + \omega_j - \omega_k - \omega_l$	$\psi_i + \psi_j - \psi_k - \psi_l$	$T_{\psi_i} \otimes T_{\psi_j} \otimes T_{\psi_k} \otimes T_{\psi_l}$	$\sigma_i^2 + \sigma_j^2 + \sigma_k^2 + \sigma_l^2$
1b	Asymmetrical and Independent	$i = j \neq k \neq l$	$2\omega_j - \omega_k - \omega_l$	$2\psi_j - \psi_k - \psi_l$	$T_{2\psi_i} \otimes T_{\psi_k} \otimes T_{\psi_l}$	$4\sigma_j^2 + \sigma_k^2 + \sigma_l^2$
2a	Asymmetrical	$k = l$	$\omega_i + \omega_j - 2\omega_l$	$\psi_i + \psi_j - 2\psi_l$	$T_{\psi_i} \otimes T_{\psi_j} \otimes T_{2\psi_l}$	$\sigma_i^2 + \sigma_j^2 + 4\sigma_l^2$
2b	Asymmetrical	$i = j,\ k = l$	$2\omega_i - 2\omega_l$	$2\psi_i - 2\psi_l$	$T_{2\psi_i} \otimes T_{2\psi_l}$	$4\sigma_i^2 + 4\sigma_l^2$
3a	Symmetrical and Correlated	$i = l = j \neq k$	$\omega_j - \omega_k$	$\psi_j - \psi_k$	$T_{\psi_j} \otimes T_{\psi_k}$	$\sigma_j^2 + \sigma_k^2$
3b	Symmetrical and Correlated	$i = k = l \neq j$ $i = j = l \neq k$	$\omega_j - \omega_l$ $\omega_l - \omega_k$	$\psi_j - \psi_l$ $\psi_l - \psi_k$	$T_{\psi_j} \otimes T_{\psi_l}$ $T_{\psi_l} \otimes T_{\psi_k}$	$\sigma_j^2 + \sigma_l^2$ $\sigma_l^2 + \sigma_k^2$

third categories, the baseband power spectrum is not related to the RF power spectra. Furthermore, in the third category, the baseband distortion power is not related to the RF level of the intermodulation product. This is especially true near saturation, where a large third category product may appear at RF, but will appear at a much reduced level in the baseband.

CALCULATION OF BASEBAND DISTORTION PRODUCED BY INTERMODULATION PRODUCTS

It is assumed that the constants (H's) of equation (11) are computed by using one of the various reported device characterizations, i.e., that the output level of each carrier term and the frequency, magnitude and phase, and composition of the carriers which generate each intermodulation product are known. There are then two basic questions regarding the computation of the baseband distortion caused by the intermodulation products:

a. Should the intermodulation products be added at RF or should the equivalent distortion be added at baseband?

b. Should the intermodulation products generated by the amplitude and phase nonlinearities be added together on a power basis or be treated separately?

Some of the approaches used recently can be summarized as follows. In the first approach, the contributions of phase and amplitude nonlinearities are added on a power basis and a composite RF intermodulation noise spectrum, using the Gaussian spectral approximation, is generated. Since the composite noise spectrum is not uniform, the term "colored noise" is often used. This colored noise density is replaced by an "equivalent" uniform noise density over the Carson's Rule bandwidth of the wanted carrier, and the FM equation for thermal noise is used to calculate the baseband distortion.

A second approach is the convolution of the colored noise spectral density with the power spectral density of the desired carrier to compute the baseband distortion.

In the third approach, the contributions of phase and amplitude nonlinearities are added on a power basis and a composite RF power level for each intermodulation product is generated. The intermodulation products are then treated as incoherent interference entries and the baseband distortion is calculated by using interference formulas. Although the method is different, this approach should give results identical to those obtained by using the previous approach.

In the last of the four approaches, that described here, the information regarding the relative phase of the intermodulation product is retained. Each product is then treated as an interference entry and its baseband distortion is calculated by using the effective power spectral density or its Gaussian approximation, as given in the appropriate category of Table 1. Thus the effects of correlation, effective deviation, and relative phase are included.

One advantage of the last two approaches is that the intermodulation products making major contributions to the baseband distortion can be readily identified. Also, for FDM/FM systems, the use of the Gaussian spectral approximation in these approaches makes it possible to replace the generation and convolution of power spectra with a simple analytic formula leading to more efficient use of computer time and memory.

The noise power ratio caused by the general intermodulation product *ijk* at a particular telephone channel of the *l*-th carrier is [9]

$$(NPR)_{ijk} = \frac{p(x_r)\sqrt{2\pi}\, m_l^2 m_{eff}}{R(1-\epsilon)x_r^2}\left[\operatorname{sech}\left(\frac{x_r x_s}{m_{eff}^2}\right)\exp\left(\frac{x_s^2+x_r^2}{2m_{eff}^2}\right)\right] \qquad (26)$$

where $p(x_r)$ = CCIR pre-emphasis improvement factor

m_l = rms modulation index of the carrier of interest (l)
$= \sigma_l/\omega_{ml}$

ϵ = ratio of minimum to maximum baseband frequency
$= \omega_1/\omega_{ml}$

x_r = normalized frequency (to top baseband frequency) of the telephone channel of interest
$= \omega/\omega_{ml}$

x_s = normalized frequency separation between the carrier and the average frequency of interference
$= \Delta\omega/\omega_{ml}$

R = ratio of undesired signal power to carrier power
$= |H_{ijk}|^2/|H_l|^2$

$m_{eff} = \sigma_{\Delta\psi}/\omega_{ml}$.

The total *NPR* is the sum of the contributions of all intermodulation products.

If the statistics of the intermodulation products are neglected (as in the third approach), the ratio R is simply the ratio of the RF power of the intermodulation product to the desired carrier power, and $\sigma_{\Delta\psi}$ is calculated according to the expression given in Table 1 for products of category 1.

The proposed approach differs from the third (or conventional) approach in two respects:

a. $\sigma_{\Delta\psi}$ is calculated according to the expressions given in Table 1 for the category of interest.

b. If the intermodulation product is in category 3, its phase relative to the demodulated carrier is included in the calculations by grouping the symmetric products and by letting $R = 4\Delta H^2 \sin^2\alpha$, where α is the phase angle relative to the demodulated carrier. If this product is in phase with the carrier so that $\alpha = 0$, the resulting *NPR* will go to infinity and no distortion will occur.

RELEVANCE AND EXPERIMENTAL CONFIRMATION

The concepts developed in this paper and the method of calculating baseband distortion derived from these concepts lead to results which cannot be predicted from conventional intermodulation theory. Some of the experimental and field test results which confirmed the concepts developed here and were directly relevant to INTELSAT system transmission planning are described in this section.

Operation with Two FM Carriers per Transponder

Operation with two FM carriers per transponder is the classic case in which the third-order intermodulation product levels are usually measured. For two equal carriers with equal rms deviation, σ, on each, it can be shown that the RF power spectral density of the $2\omega_j - \omega_l$ products corresponds to an rms deviation of $\sqrt{5}\sigma$, causing an overlap between the RF power spectral densities of the carriers and intermodulations. In conventional approaches, it is assumed that this overlap within the Carson's Rule bandwidth will cause baseband distortion.

In the proposed approach, it can be seen that the product $2\omega_l - \omega_j$ has a symmetrical twin located at $\omega_l + \omega_j - \omega_l$ (i.e., at ω_j) and is in category 3. Therefore, only the component generated by phase nonlinearity will appear at the demodulator output. Also, when the carrier at ω_l is demodulated, the effective deviations produced by the $\omega_l \pm (\omega_l - \omega_j)$ product and the carrier at ω_j are identical, although the distortion produced by the latter is an order of magnitude larger. Heuristically, the carrier and intermodulation track in frequency to maintain a difference in instantaneous frequency which is the same as the difference between the two carriers. Hence, if the system is designed to achieve adequate protection against the

adjacent carrier, it is automatically protected against the intermodulation product, and little baseband distortion results.

FM/Video Transmission on INTELSAT IV

For FM/video transmission on INTELSAT IV, one FM/video carrier and two program channels share the same transponder. The carrier locations and the intermodulation products generated at near-saturation operation of the output TWTA are shown in Figure 3. The symmetric category 3a products appearing about the video carrier *l* were the subject of this investigation.

To show the reduced deviation predicted for these products, carriers *j* and *k* were left unmodulated while carrier *l* was modulated, first with a 25-kHz tone and then with white noise band-limited between 12 kHz and 1975 kHz. The RF spectra of the carrier and the intermodulation products, shown in Figures 4a and 4c, indicate considerable spectral overlap. The output of an FM demodulator was displayed on a baseband spectrum

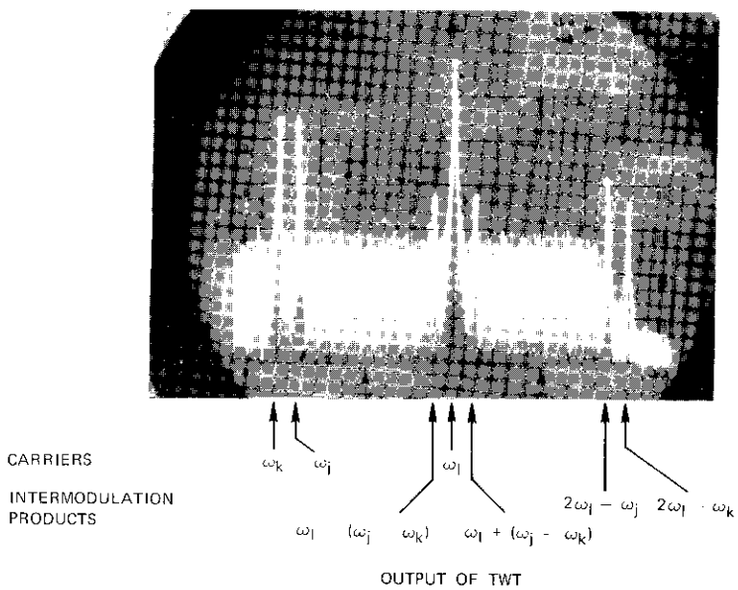


Figure 3. Carriers and Intermodulation Products

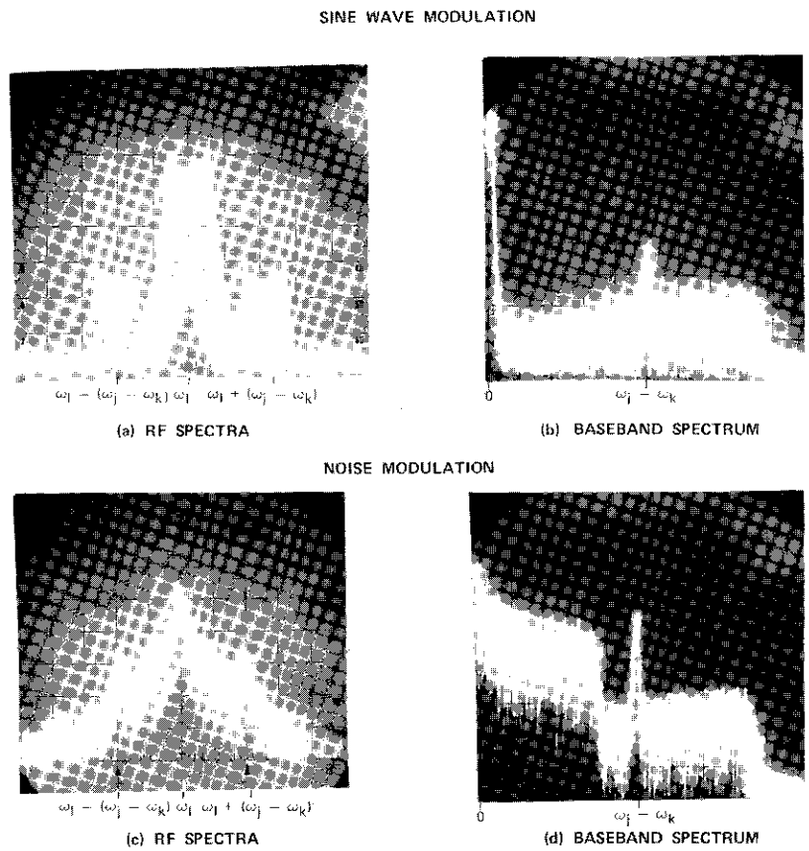


Figure 4. Correlated Intermodulation, Carrier l Modulated

analyzer as shown in Figures 4b and 4d. As predicted, the baseband distortion term contained only a single frequency tone at the difference frequency of $\omega_j - \omega_k$.

On the other hand, when only carrier *j* was modulated by a 25-kHz tone, this modulation was transferred to the intermodulation products (Figures 5a and 5c) and, as expected, the same spectrum shape reappeared at the baseband output (Figures 5b and 5d).

To show the effect of in-phase and phase quadrature modulation, the

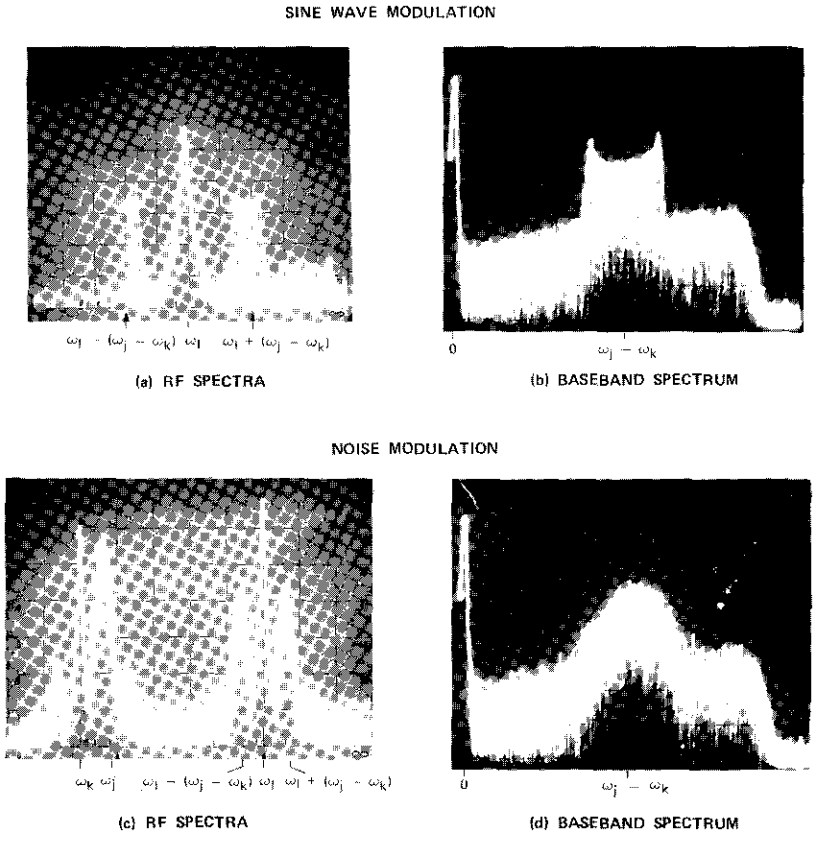


Figure 5. Uncorrelated Intermodulation, Carrier j Modulated

level of the intermodulation product was measured at RF and at baseband. The latter measurement was performed with both an envelope detector to measure the in-phase component and an FM demodulator to measure the component in phase quadrature. The measured carrier-to-intermodulation ratios are plotted in Figure 6. Note that near TWTA saturation, where most of the intermodulation power is in phase with the carrier, the phase demodulator output is considerably smaller than either the RF or the envelope detector output. As a result, very little baseband distortion is

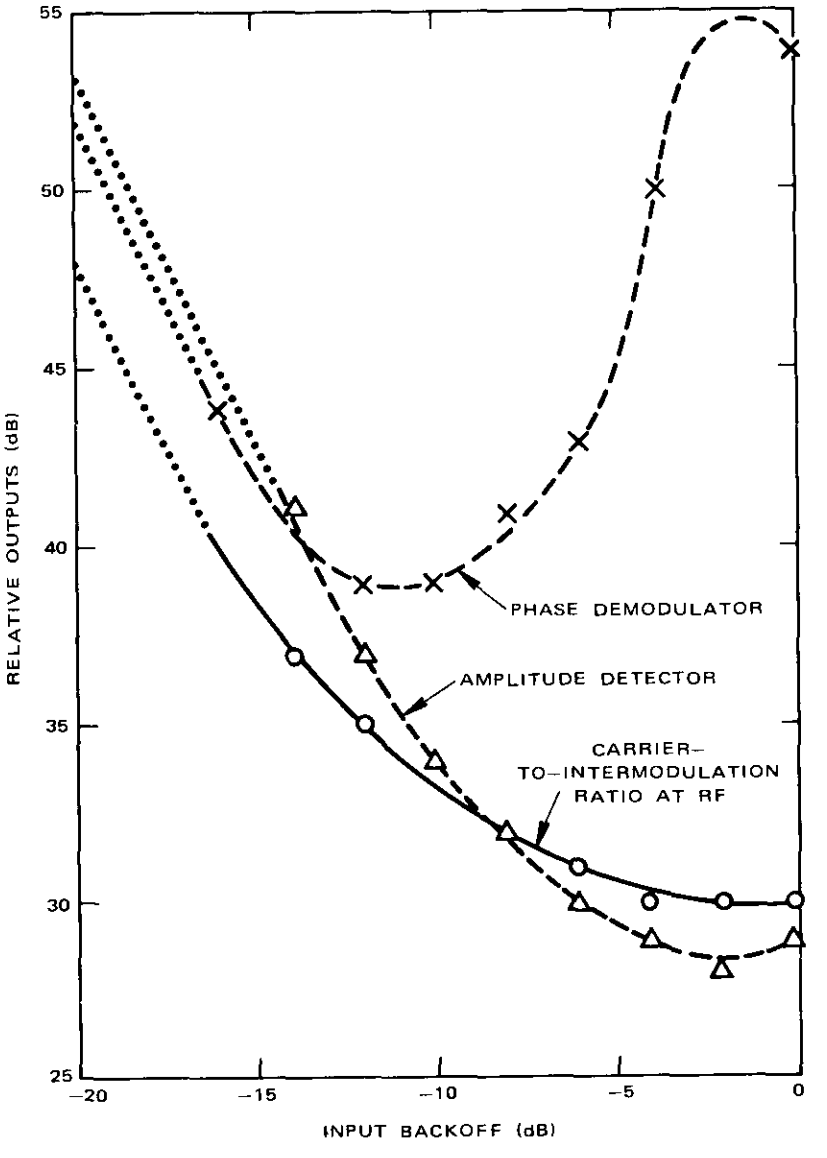


Figure 6. Relative Outputs as a Function of Input Backoff

produced even though the RF carrier-to-intermodulation level is 30 dB.

In terms of video quality, the peak-to-peak picture-to-weighted-noise ratio (S/N_w) predicted from conventional intermodulation theory was 59.1 dB, that predicted from the present approach was better than 66 dB, and the measured value was 68.6 ± 2 dB.

Multicarrier FDM/FM Transmission on INTELSAT IV

Since multicarrier FDM/FM transmission is the major mode of operation within the INTELSAT system, an extensive program of laboratory simulations [10] and field tests was undertaken to test the transmission plan. As an example, the five-carrier frequency plan shown in Figure 7 was used in both the laboratory simulations and the field tests; the baseband frequencies, multichannel and test-tone deviations, and relative carrier levels were in accordance with the INTELSAT IV transmission standards.

The noise power ratio (NPR) in the top baseband slot was calculated by using the last two of the four approaches described previously. The starting point of both computations was the RF level of the intermodulation products generated by amplitude and phase nonlinearities of the TWTA. As an example, Table 2 lists the frequency, composition, and RF level of the intermodulation products falling in the vicinity of carrier A in the spot-beam configuration for a TWTA input backoff of 5 dB. The level of the intermodulation products was computed by using a program based on Berman and Mahle's study [3]. The products were classified in accordance with Table 1.

In the conventional approach (corresponding to the third approach described previously), all intermodulation components are assumed to be independent, whereas in the proposed approach their classification is taken into account. The NPR caused by each individual intermodulation product was calculated by using both approaches. Table 2 includes all intermodulation products for which an NPR of 60 or less was calculated. It should be noted that the contribution of the class 1 products is the same for both cases, that there are no class 2 products, and that the contribution of the class 3 products is substantially different because the proposed technique takes into account only the phase quadrature component of the intermodulation product and also assumes a lower effective deviation. Since the class 3 products generate the major component of baseband distortion for this particular carrier configuration, the total NPR is substantially different for the two calculations.

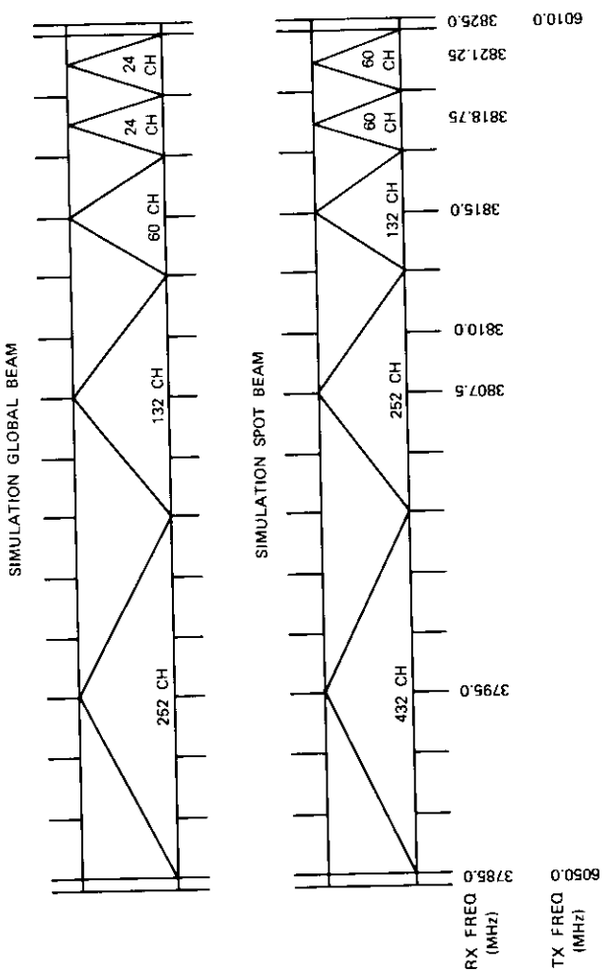


Figure 7. Frequency Plans Used in Simulations

TABLE 2. BASEBAND DISTORTION (NPR) CAUSED BY INTERMODULATION PRODUCTS IN THE VICINITY OF CARRIER A, (frequency = 3795.0 MHz, output level = −5.0 dB, multicarrier input backoff = 5 dB)

Frequency (MHz)	Class	Composition of Intermodulation Product	RF Level	NPR (dB) Conventional Method In Phase	Conventional Method In Quadrature	Proposed Method In Phase	Proposed Method In Quadrature
3787.50	3a	A + B − C	−24.5	43.4	46.1	—	82.0
3788.75	3a	A + C − E	−29.3	44.5	46.8	—	—
3791.25	3a	A + C − D	−29.3	37.0	39.2	—	52.0
3792.50	3a	A + D − E	−32.2	38.0	40.0	—	37.1
3793.75	1b	2B − E	−38.7	44.5	44.6	44.5	44.6
3796.25	1b	2B − D	−38.7	44.5	44.6	44.5	44.6
3797.50	3a	A + E − D	−32.2	38.0	40.0	—	37.1
3798.75	3a	A + D − C	−29.3	37.0	39.1	—	52.4
3800.00	1b	2B − C	−35.9	46.5	46.5	46.5	46.5
3801.25	1a	B + C − E	−34.0	52.0	52.7	52.0	52.7
3801.25	3a	A + E − C	−29.3	44.5	46.8	—	—
3802.50	3a	A + C − B	−24.5	43.4	46.1	—	82.0
		SUBTOTAL		30.1	32.0	40.0	33.0
Total NPR Caused by In-Phase and In-Quadrature Components				27.9		32.2	

The two techniques were also used to calculate the NPR for varying TWTA backoffs. The results are plotted in Figure 8, as are the values measured on the carrier. The results obtained by using the proposed approach are in good agreement with the measurements. The estimated error incurred by excluding the fifth-order products is between 0.5 and 1.0 dB at saturation and decreases with increasing backoff.

The difference in the levels of baseband distortion predicted by using the two approaches depends on the mixture of the three categories of products.

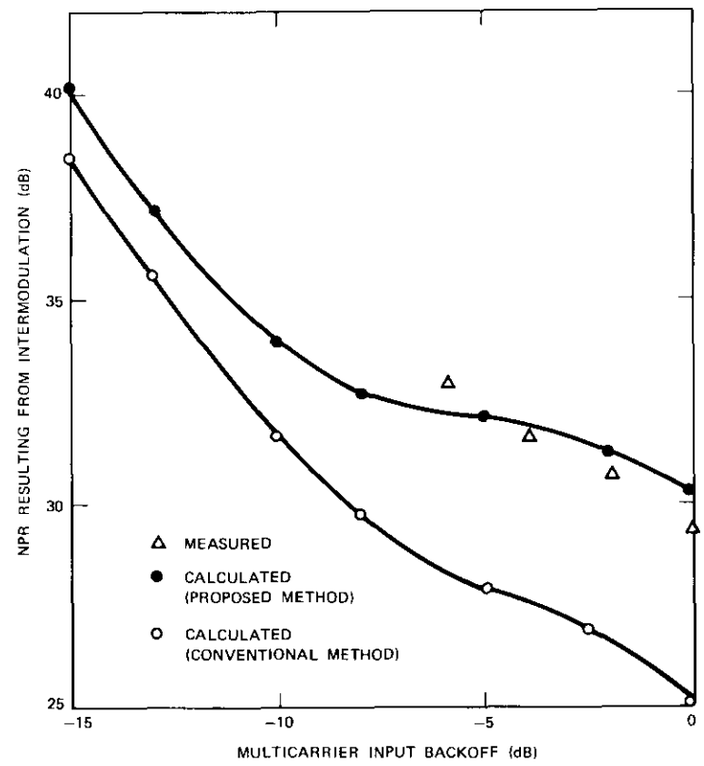


Figure 8. Measured and Calculated NPR (caused by intermodulation) vs. Multicarrier Input Backoff for Carrier A, Spot-Beam Configuration

If the third category products predominate, a large difference is to be expected. However, this is not always the case, as illustrated in Figure 9 for carrier B. (Although global-beam parameters are used in Figure 9, the frequency plan is the same as that shown in Figure 7.)

Work is in progress on a general intermodulation analysis program based on Shimbo's [4] unified treatment of amplitude and phase nonlinearities and including fifth-order intermodulation product levels and their classifications as an extension of the approach discussed here.

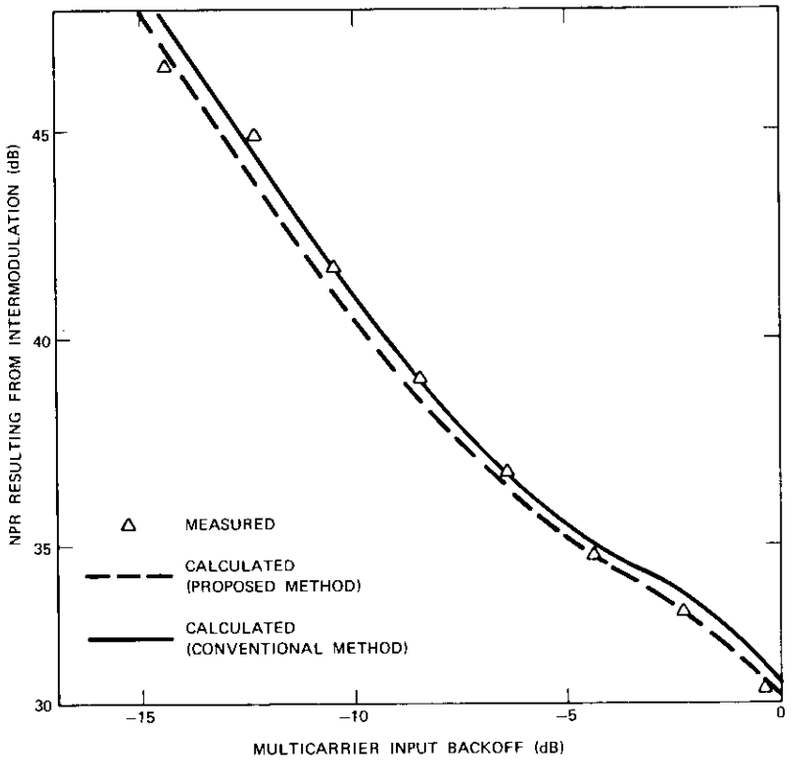


Figure 9. Measured and Calculated NPR (caused by intermodulation) vs. Multicarrier Input Backoff for Carrier B, Global-Beam Configuration

CONCLUSIONS

In this paper we have shown that the intermodulation products generated by a nonlinear amplifier in multicarrier operation should be treated as angle-modulated carriers rather than noise. The baseband distortion produced by these products has therefore been calculated by using the results of angle modulation interference theory rather than "colored" noise theory, which was used previously.

The concepts of correlated intermodulation products, effective deviation, and effective power spectral density have been introduced and the

significant intermodulation products have been divided into three categories. The consequence of these concepts is that the baseband distortion level and spectral shape are different from predictions obtained by using the conventional approach and considering the RF power spectral shape of the intermodulation products as additive colored noise.

A closer agreement between the measured results and predictions obtained by using the concepts developed here was demonstrated for certain cases. Finally, improvements in system performance were shown for certain carrier combinations and the implications for system design were discussed.

REFERENCES

[1] E. D. Sunde, "Intermodulation Distortion in Multicarrier FM Systems," *IEEE International Convention*, New York, March 1965, *Convention Record*, Vol. 13, Part 2, pp. 130–146.

[2] R. J. Westcott, "Investigation of Multiple FM/FDM Carriers Through a Satellite TWT Operating Near to Saturation," *Proceedings of the IEE* [U.K.], Vol. 114, No. 6, June 1967, pp. 726–750.

[3] A. L. Berman and C. E. Mahle, "Nonlinear Phase Shift in Traveling-Wave Tubes as Applied to Multiple-Access Communications Satellites," *IEEE Transactions on Communications Technology*, COM-18, No. 1, February 1970, pp. 37–48.

[4] O. Shimbo, "Effects of Intermodulation, AM-PM Conversion, and Additive Noise in Multi-carrier TWT Systems," *Proceedings of the IEEE* [U.S.A.], Vol. 59, No. 2, February 1971, pp. 230–238.

[5] A. F. Standing, "An Active-Phase and Amplitude-Correction Device for Reducing the Intermodulation Produced by TWT's and Klystrons," *IEE Conference on Earth Station Technology*, London, October 1970, IEE Conference Publication 72, pp. 274–279.

[6] V. K. Prabhu and L. H. Enloe, "Interchannel Interference Considerations in Angle-Modulated Systems," *Bell System Technical Journal*, Vol. 48, No. 7, September 1969, pp. 2333–2358.

[7] M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw-Hill, 1966.

[8] S. Goldman, *Frequency Analysis Modulation and Noise*, New York: Dover Publications, Inc., 1967, p. 334.

[9] R. Hamer, "Radio-Frequency Interference in Multi-channel Telephony FM Radio Systems," *Proceedings of the Institution of Electrical Engineers* [U.K.], Part B: Electronics and Communication Engineering, Vol. 108, January 1961, pp. 75–89.

[10] N. K. M. Chitre, "INTELSAT IV Communications System Simulation," *Proceedings of the IEEE International Conference on Communications*, Montreal, June 1971, pp. 35–24 to 35–29.

Dr. Nandkishore M. Chitre was born in Calcutta, India. He received the B.Sc. and M.Sc. degrees in physics from the Banaras Hindu University in 1953 and 1955 and the M. Tech. degree in Electronics from the Indian Institute of Technology, Kharagpur, in 1956. He studied at the electrical engineering department of the Imperial College of Science and Technology, London, and was awarded the Ph.D. degree in physics from the University of London in 1962. Dr. Chitre joined COMSAT in December 1968. He is Manager of the Systems Simulation Branch in the Systems Laboratory of COMSAT Laboratories.

Jorge C. Fuenzalida was born in Santiago, Chile. After finishing engineering school, he emigrated to the U. S. A. in 1965. He received the M.S.E.E. degree from Columbia University in 1966, and subsequently was a Ph.D. candidate in systems engineering at the University of Pennsylvania. Mr. Fuenzalida is a member of the technical staff in the Systems Laboratory at COMSAT Laboratories, working on interference and nonlinear intermodulation in FM systems, on satellite capacity optimization problems, on frequency sharing between satellites and radio-relay systems and on orbit and spectrum sharing. He has been active in the Study Group 4 of the CCIR.

Index: communication satellites, modulation, radio relay systems, voice communications

A comparison of voice communication techniques for aeronautical and marine applications

S. J. CAMPANELLA AND J. A. SCIULLI

ABSTRACT

The intent of this paper is to make a comparative analysis of several voice modulation techniques which are candidates for aeronautical and marine satellite applications. The techniques considered are FM (either with discriminator or PLL demodulation), PCM/PSK, DM/PSK, and PDM/PSK. The objective is to determine the relative performance of each system under the power and bandwidth constraints expected for the aeronautical and marine applications. For each modulation system, a test-tone-to-noise ratio vs. C/N_o (carrier-to-noise density ratio) characteristic is first determined. This is then converted to an articulation index vs. C/N_o characteristic by using an analytic method based on noise masking of frequency bands with equal articulation index weight in the speech spectrum.

INTRODUCTION

Economic considerations in aeronautical and marine radio communications via satellite place great importance on methods for realizing acceptably intelligible voice communications with a minimum commitment of available satellite repeater power per channel. Essentially, because of the low-gain antennas used on ships and aircraft, the satellite repeater

must deliver a considerably greater down-link power per voice channel than is required for a typical voice channel in the international communications satellite system. A typical earth station in the INTELSAT system has an antenna-gain-to-system-noise-temperature ratio (G/T) of 41 dB/°K.

In the case of aircraft and ships, it is expected that antennas may have gains ranging from 0 to 15 dB. Greater gains are difficult to achieve because of the pointing problem that arises on a mobile platform. For an L-band (1600-MHz) system noise temperature of 500°K, the G/T value for the mobile antenna would lie in the range from -27 to -12 dB/°K. When a space loss difference of 8 dB between C- and L-bands is taken into account, it is seen that the required power per voice channel would be 45 to 60 dB greater than that used in the international system for the same modulation techniques and performance requirements. In view of this, attention is currently being focused on various voice communication techniques which deliver acceptable voice quality at relatively low values of C/N_o.

The present paper is specifically intended to evaluate the relative performance of several methods of speech communication to mobile platforms. The results do not include the effects of fading or multipath. The voice communication systems considered are frequency modulation (FM), pulse-duration modulation (PDM), delta modulation (DM), and pulse-code modulation (PCM). Relationships which express the performance of each method in terms of test-tone-to-noise ratio (T/N) versus carrier-to-noise density ratio (C/N_o) are derived.

Unfortunately T/N is not a sufficient performance characteristic for measuring the relative performance of different voice modulation systems. Accordingly, the so-called articulation index (AI) has been chosen as an absolute figure of merit for all systems considered. The AI is a weighted fraction representing, for a given speech channel and noise condition, the effective proportion of the normal speech channel which is available to the listener for conveying intelligibility [1].

A preassigned ratio of test-tone power to speech power, combined with assumptions pertaining to the shape of the received noise spectrum, is employed to develop curves relating AI to T/N. These results are then used to obtain plots of AI versus C/N_o for the various speech processing methods for channel bandwidths of 10, 20, and 40 kHz. It is assumed that the speech is bandpassed through a filter with a high-pass of 200 Hz and a low-pass of 2500 Hz.

Each of the methods analyzed possesses a different AI versus C/N_o characteristic. To some extent, those techniques possessing the greatest

value of AI for the lowest value of C/N_o are the most desirable, since, under conditions of very low received signal strength, they will provide the most satisfactory performance. This is true even if the method results in some sacrifice of AI at higher values of C/N_o. From this point of view, FM and DM provide comparable performance capabilities for the 10-kHz bandwidth. For the 20-kHz bandwidth, DM and PDM perform better than FM. In the 40-kHz case, the relative performance of PCM improves significantly, DM provides good performance at high and intermediate values of C/N_o, and PDM appears to give the best performance at very low values of C/N_o. It will be seen that some enhancement of the performance of DM and PCM at intermediate values of C/N_o is obtained by using rate one-half coding combined with four-phase PSK modulation; however, performance at low C/N_o deteriorates.

RELATIONSHIP BETWEEN ARTICULATION INDEX AND TEST-TONE-TO-NOISE RATIO

Articulation Index Determination

For each modulation technique considered, a plot of T/N versus C/N_o will be determined. Since the desired result is a plot of AI versus C/N_o for each method, a relationship between AI and T/N must be determined for each case. A graphical method for determining AI is outlined in Reference [2], but it is cumbersome and time consuming and was not used here. Instead, a computational method based on certain assumptions regarding the spectral densities of the speech signal and the perturbing noise was used. A knowledge of the noise power spectrum, $N(\omega)$, and the speech power spectrum, $F(\omega)$, makes it possible to compute the articulation index, AI, as a function of the test-tone-to-noise ratio, T/N.

Power Spectrum of Speech

The power spectral distribution corresponding to a speech signal that occupies the available amplitude range is given by

$$S(\omega) = k_o F(\omega) \tag{1}$$

where k_o is a gain constant which is adjusted so that the peaks of the speech signal just attain the maximum amplitude limit. The total speech power, P_S, is

$$P_S = k_o \int_{\omega_0}^{\omega_m} F(\omega)\, d\omega \tag{2}$$

where ω_m and ω_0 are the upper and lower audio cutoff frequencies, respectively.

It is also possible to relate P_S to the power of a test tone, P_T, that occupies the same peak-to-peak amplitude range. That is,

$$P_S = cP_T \tag{3}$$

where c, expressed in dB, is the difference between the rms level of a sinusoidal tone whose peak is adjusted to attain just the upper amplitude limit and the rms level of a speech signal adjusted so that its peak values just attain the same upper limit. From equations (2) and (3),

$$k_o = \frac{cP_T}{\int_{\omega_0}^{\omega_m} F(\omega)\, d\omega} \tag{4}$$

and, hence, the speech power spectrum distribution is

$$S(\omega) = \left(\frac{F(\omega)}{\int_{\omega_0}^{\omega_m} F(\omega)\, d\omega} \right) cP_T. \tag{5}$$

The value used for the peak-to-rms ratio for speech depends on the amount of speech clipping that is to be permitted. If speech clipping is to occur no more than 0.1 percent of the time, then it should take place at a level 14.2 dB greater than the rms level, and the peak-to-rms ratio for the resulting clipped signal should be 14.2 dB [3]. This is a typical clipping level for high-quality systems.

In order to use more efficiently the dynamic range of the channel, the speech signal is usually clipped at some selected level to reduce the peak-to-rms ratio. In the calculations which follow, it is assumed that clipping occurs at a level that is 6 dB greater than the unclipped rms value of the speech. Such clipping produces a speech signal having a peak-to-rms ratio of approximately 8 dB and exhibiting negligible reduction in word intelligibility [4]. Since the peak-to-rms ratio of the sinusoidal tone is 3 dB, the value of c is 3 dB − 8 dB = −5 dB. This value will be slightly modified if the speech is pre-emphasized. In this event it is assumed that the clipping level will be modified to produce a final clipped speech signal with an 8-dB peak-to-rms ratio. Thus, the value of c is unchanged.

Speech Signal-to-Noise Spectral Density Ratios

To determine AI, the speech signal-to-noise power ratio in each of 20 different frequency bands in the speech spectrum range must be determined [2]. In the nth frequency band, defined by the range limits $\omega_n < \omega < \omega_{n+1}$, the speech signal power determined from equation (5) is

$$S_n(\omega) = q_n\, cP_T \tag{6}$$

where

$$q_n = \frac{\int_{\omega_n}^{\omega_{n+1}} F(\omega)\, d\omega}{\int_{\omega_0}^{\omega_m} F(\omega)\, d\omega}.$$

The noise power falling in the nth band can be determined as follows. If the test-tone-to-noise ratio, (T/N), is equal to R, then the total noise power relative to the power of the test tone, P_T, is P_T/R. If this power is distributed over the frequency range $\omega_0 < \omega < \omega_m$ in accordance with the function $N(\omega)$, then the amount of noise power falling in the nth band is

$$N_n = p_n \left(\frac{P_T}{R} \right) \tag{7}$$

where

$$p_n = \frac{\int_{\omega_n}^{\omega_{n+1}} N(\omega)\, d\omega}{\int_{\omega_0}^{\omega_m} N(\omega)\, d\omega}.$$

Thus, the speech signal-to-noise ratio in the nth band, given by the ratio of equations (6) and (7), is

$$\left(\frac{S}{N} \right)_n = \left(\frac{T}{N} \right) \left(\frac{q_n}{p_n} \right) c. \tag{8}$$

AI Computation

To compute the total AI, contributions from each of 20 frequency bands with equal AI weight must be summed. The maximum weight which can be given to each band is 1/20; hence, the maximum AI is unity.

The maximum AI is attained in any band when the speech signal-to-noise ratio in that band is equal to or greater than 20 dB. The AI contribution in any band for other values of speech signal-to-noise ratio is given by

$$\Delta_n(AI) = \begin{cases} \frac{1}{60}\left[\log\left(\frac{S}{N}\right)_n + 1\right]; & 0 \leq \log\left(\frac{S}{N}\right)_n + 1 \leq 3 \\ \frac{1}{20}; & \log\left(\frac{S}{N}\right)_n + 1 > 3 \\ 0; & \log\left(\frac{S}{N}\right)_n + 1 < 0 \end{cases} \quad (9)$$

Finally, the total AI is given by

$$AI = \sum_{n=1}^{20} \Delta_n(AI). \quad (10)$$

The spectrum limits for the nth channel are indicated in the following table.

n	ω_n	n	ω_n
1	200	12	1830
2	330	13	2020
3	430	14	2240
4	560	15	2500
5	700	16	2820
6	840	17	3200
7	1000	18	3650
8	1150	19	4250
9	1310	20	5050
10	1480	21	6100
11	1660		

The speech spectrum distribution used to compute AI is

$$F(\omega) = \frac{\omega_1^2}{\omega^2 + \omega_1^2} \cdot \frac{\omega_2^2}{\omega^2 + \omega_2^2}$$

$$f_1 = \frac{\omega_1}{2\pi} = 500 \text{ Hz}$$

$$f_2 = \frac{\omega_2}{2\pi} = 1500 \text{ Hz}. \quad (11)$$

This corresponds to a frequency response characteristic that descends at 6 dB per octave between 500 and 1500 Hz and at 12 dB per octave beyond 1500 Hz. As previously mentioned, the speech signal is considered to be bandpassed from 200 to 2500 Hz for the systems considered here. This will result in a maximum AI of 0.65 for all systems considered.

There are two noise spectra applicable to the various signal processing methods analyzed. The first is a flat noise model given by $N(\omega) = 1$, and the second is a 6-dB-per-octave descending model given by $N(\omega) = 1/\omega^2$. These models are considered to be flat or descending relative to the normal speech spectrum distribution. Curves 1 and 2 in Figure 1 are plots of AI versus T/N determined by using the previous method for flat noise and 6-dB-per-octave descending noise, respectively. These curves will be used to derive curves of AI versus carrier-to-noise density ratio in the following section.

FREQUENCY MODULATION

General

FM has been used extensively for communications via satellite repeaters. The use of FM in satellite communications has been encouraged since it provides an advantageous exchange of bandwidth and signal-to-noise ratio, and since terrestrial links are predominantly FM links. Because of its relative insensitivity to the amplitude nonlinearity of a satellite repeater TWT, angle modulation is preferred to amplitude modulation. However, frequency-modulated signals are not totally free from the effects of satellite repeater nonlinearity. For example, when many carriers appear simultaneously in the transponder TWT, intermodulation components occur which can result in intelligible crosstalk among the channels. To avoid this, it is necessary either to reduce the power output from the satellite repeater so that the intermodulation components are reduced to a level that is not objectionable, or to select carrier frequencies so that the objectionable intermodulation products fall in unused parts of the spectrum. However, for more than two carriers in a single transponder, this latter approach can result in inefficient bandwidth utilization, since much of the total bandwidth contains intermodulation products and is therefore unavailable for communication.

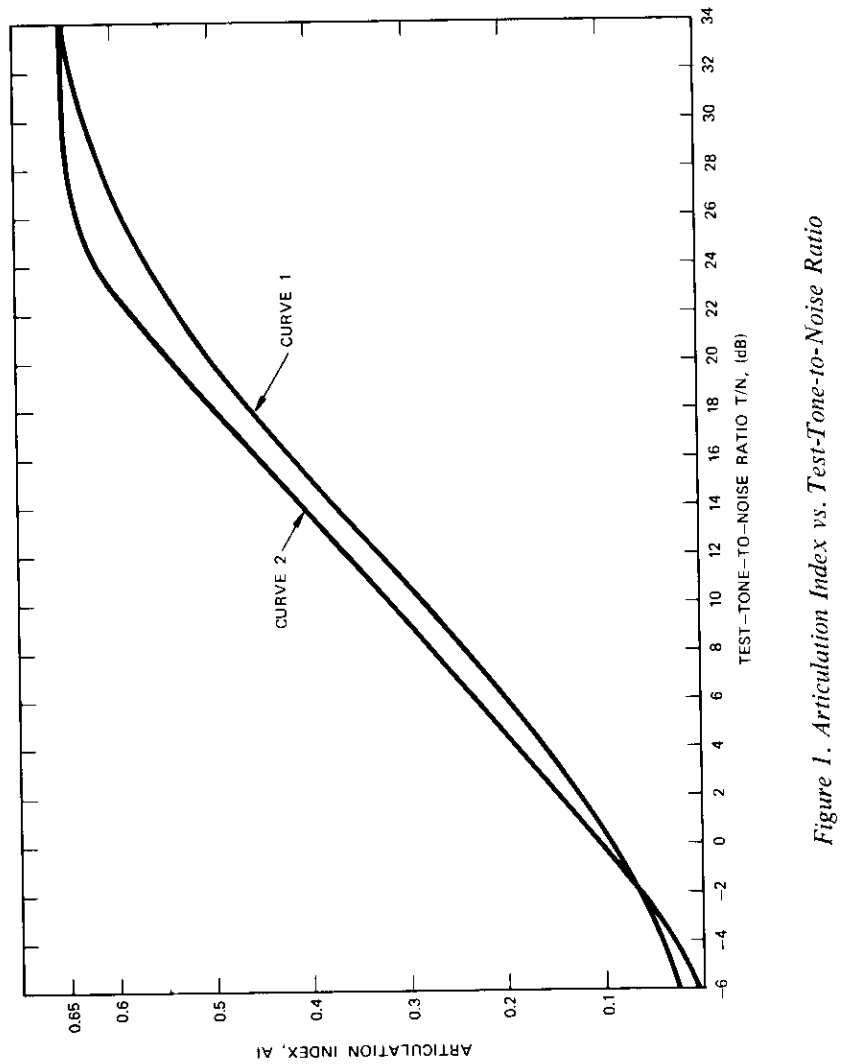


Figure 1. Articulation Index vs. Test-Tone-to-Noise Ratio

Expression for FM Performance

An expression for the T/N ratio resulting at the output of an FM discriminator (FMD) has been developed by Shimbo [5], [6], and is given by

$$(T/N)_{FMD} = \frac{3\beta^2\delta\gamma}{1 + 3.91[(\delta^2\sqrt{\gamma}\ \epsilon^{-\gamma})/(1 - \epsilon^{-\gamma})^2][1 + 3\,(\beta/\delta)^2\gamma]}. \quad (12)$$

For definitions applicable to this and following expressions, see List of Symbols.

FM receivers may incorporate phase-locked loop (PLL) demodulators rather than discriminators. Osborne and Schilling [7] have demonstrated that FM-PLL receivers provide approximately 2.5 dB of threshold extension for $\beta \simeq 3$. Since the values of β considered here are close to this value, the curves for FM–PLL receivers are obtained by extending the response of the FMD 2.5 dB below threshold.

20-kHz Bandwidth Case

The first FM case considered is that generally corresponding to a current design for aeronautical satellite receivers. In this case, $f_m = 2500$ Hz, $f_d = 6500$ Hz, $\beta = 2.6$, $\delta = 3.6$, and $B_{IF} = 18{,}000$ Hz. B_{IF} is determined by the Carson's Rule formula, $B_{IF} = 2(f_m + f_d)$. Speech clipping is assumed to result in a peak-to-rms ratio of 8 dB. For the calculations, an IF bandwidth of 20 kHz is used to provide a sufficient guard band for the 18-kHz RF signal. The resulting plots of $(T/N)_{FM}$ for the discriminator and PLL receivers are shown in Figure 2. In this plot, γ values are converted to values of carrier-to-noise density ratio, C/N_o, by adding 10 log B_{IF} to the values of γ expressed in dB.

To convert to *AI* values, it will be assumed that the speech spectrum is pre-emphasized at a positive slope of 6 dB/octave at the transmitter. In the receiver, de-emphasis at a negative slope of 6 dB/octave is introduced so that the noise produced by the FM discriminator is flat with respect to the original form of the speech spectrum. Therefore, in this case, the curve for flat noise shown in Figure 1 is used to convert the T/N values to *AI* values. The resulting curve of *AI* versus C/N_o is shown in Figure 3.

40-kHz Bandwidth Case

It is assumed that there is a 40-kHz RF bandwidth available to accommodate an FM signal with a Carson's Rule bandwidth of 36 kHz.

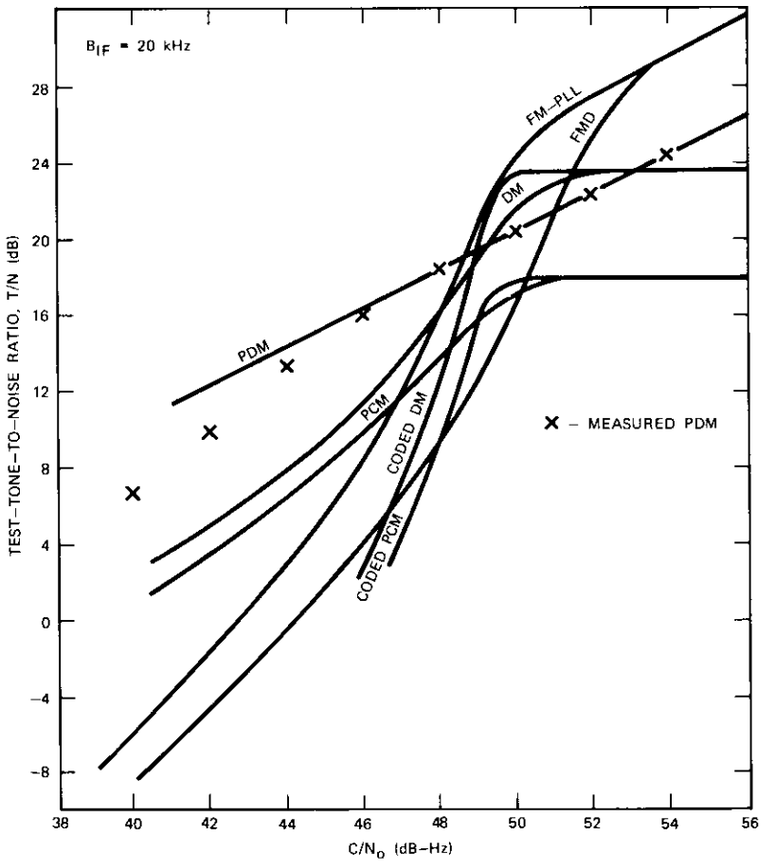


Figure 2. Test-Tone-to-Noise Ratio vs. C/N_o for 20-kHz Bandwidth

The parameters are identical to those used in the 20-kHz bandwidth case except that $B_{IF} = 36$ kHz and hence $\beta = 6.21$ and $\delta = 7.21$. The resulting curves of T/N versus C/N_o for FMD and FM–PLL are shown in Figure 4 and the corresponding AI versus C/N_o performance appears in Figure 5. The performance of FM compared to the other modulation techniques is similar to that observed in the 20-kHz case.

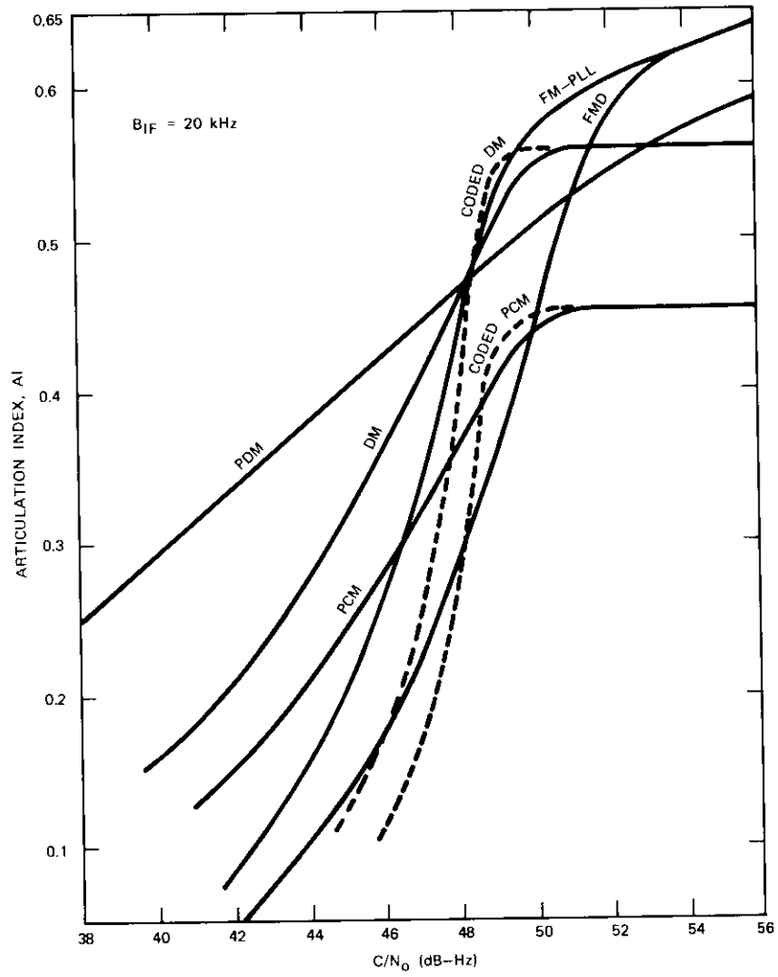


Figure 3. Articulation Index vs. C/N_o (B_{IF} = 20 kHz)

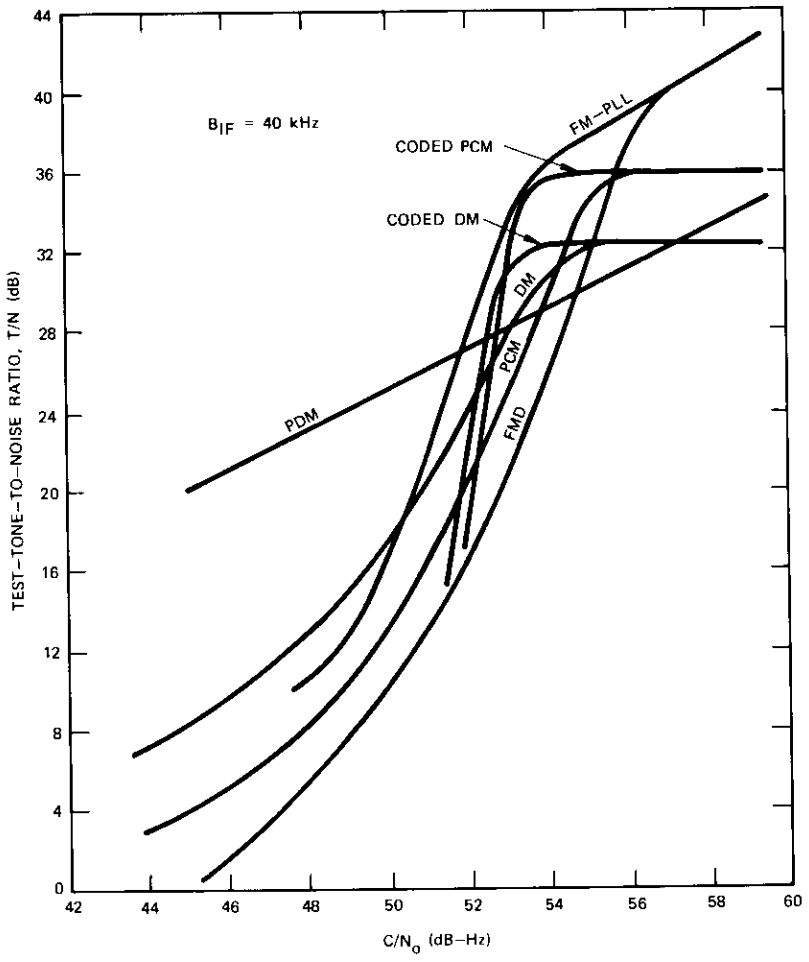


Figure 4. Test-Tone-to-Noise Ratio vs. C/N_o for 40-kHz Bandwidth

Narrowband FM Case (10 kHz)

The foregoing FM system designs are based on the use of the Carson's Rule bandwidth formula for the IF bandwidth. The next FM case of interest is a departure from this approach. It may be logically argued that

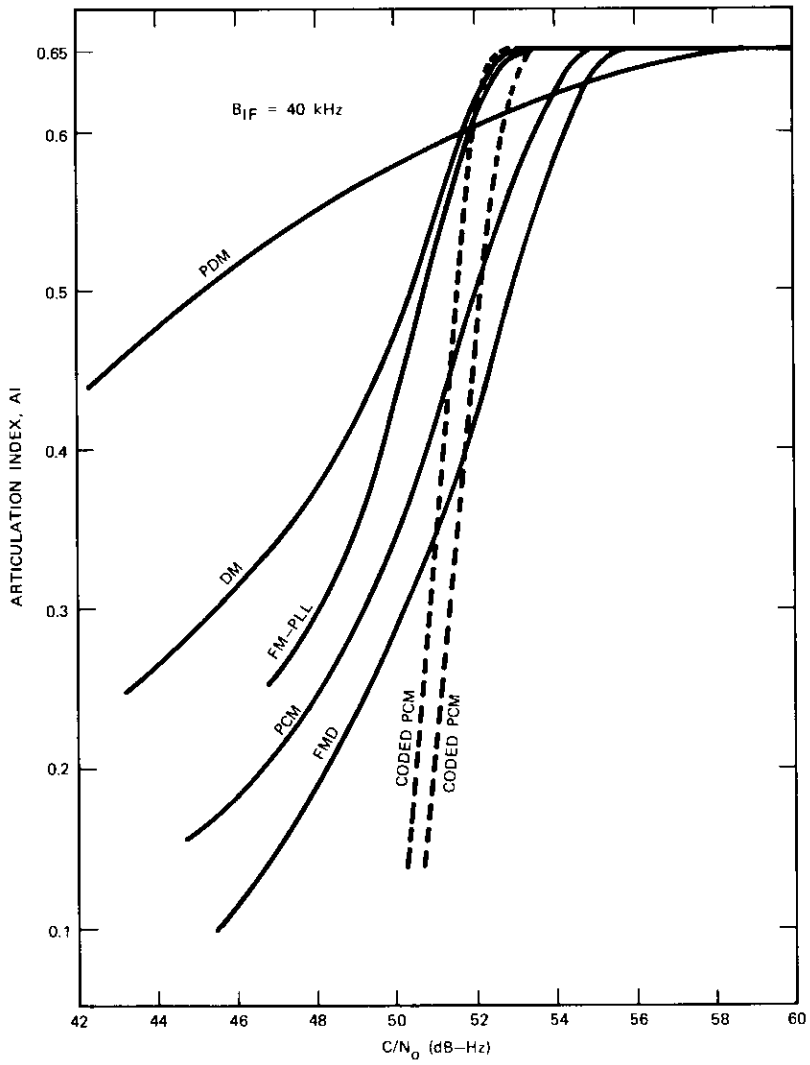


Figure 5. Articulation Index vs. C/N_o (B_{IF} = 40 kHz)

Carson's Rule does not strictly apply to speech because the spectral characteristic of speech is unlike white noise. Analysis of the speech spectrum reveals that the energy at high frequencies diminishes rather rapidly. This indicates that the IF bandwidth necessary to contain the FM carrier components needed to reconstitute the speech signal at the receiver is significantly less than that predicted by Carson's Rule. The vowel spectrum distribution model of speech begins to roll off at a rate of 6 dB per octave at approximately 500 Hz. At approximately 1000 Hz, the rate of roll off approaches 12 dB per octave. Thus, if components at 500 Hz are adjusted so that they cause a peak carrier deviation equal to frequency f_d, then those at 1000 Hz will cause the carrier to deviate by $f_d/2$, and those at 2000 Hz will cause it to deviate by only $f_d/8$.

If the value of f_d is 4000 Hz, then, because of speech components in the vicinity of 500 Hz, the FM carrier spectrum will have a deviation ratio of 8 and be approximately equal to the amplitude distribution of these components. These components will be contained in a bandwidth only slightly larger than the peak-to-peak deviation range, i.e., 8000 Hz. The components in the vicinity of 2000 Hz will cause the carrier to deviate by only 500 Hz, resulting in a modulation index of 0.25. Consequently, the components appearing in the FM carrier spectrum will occur principally in two sidebands positioned at ± 2000 Hz relative to the carrier.

If the speech spectrum is pre-emphasized by incorporating a filter at the transmitter that essentially differentiates the signal, the spectrum of speech will roll off at approximately 6 dB per octave at frequencies above 1 kHz. Thus, it can be expected that the FM carrier spectrum for the pre-emphasized voice signal can also be accommodated by using a bandwidth somewhat less than that calculated by using Carson's Rule.

To review the previous arguments, assume that a voice signal which has been pre-emphasized causes an FM carrier peak deviation $f_d = \pm 4$ kHz. To accommodate this FM carrier, let the IF bandwidth be 10 kHz. Also, assume that the voice signal is low-passed at $f_m = 2.5$ kHz. For these assumptions, β is not computed according to Carson's Rule but is simply defined as f_d/f_m. For the preceding assignments, the value of β is 1.6 and δ is 2.0. Using these values in equation (1) results in the test-tone-to-noise ratio versus C/N_o curves shown in Figure 6. These may be converted to plots of AI versus C/N_o by using curve 1 in Figure 1; the results are shown in Figure 7.

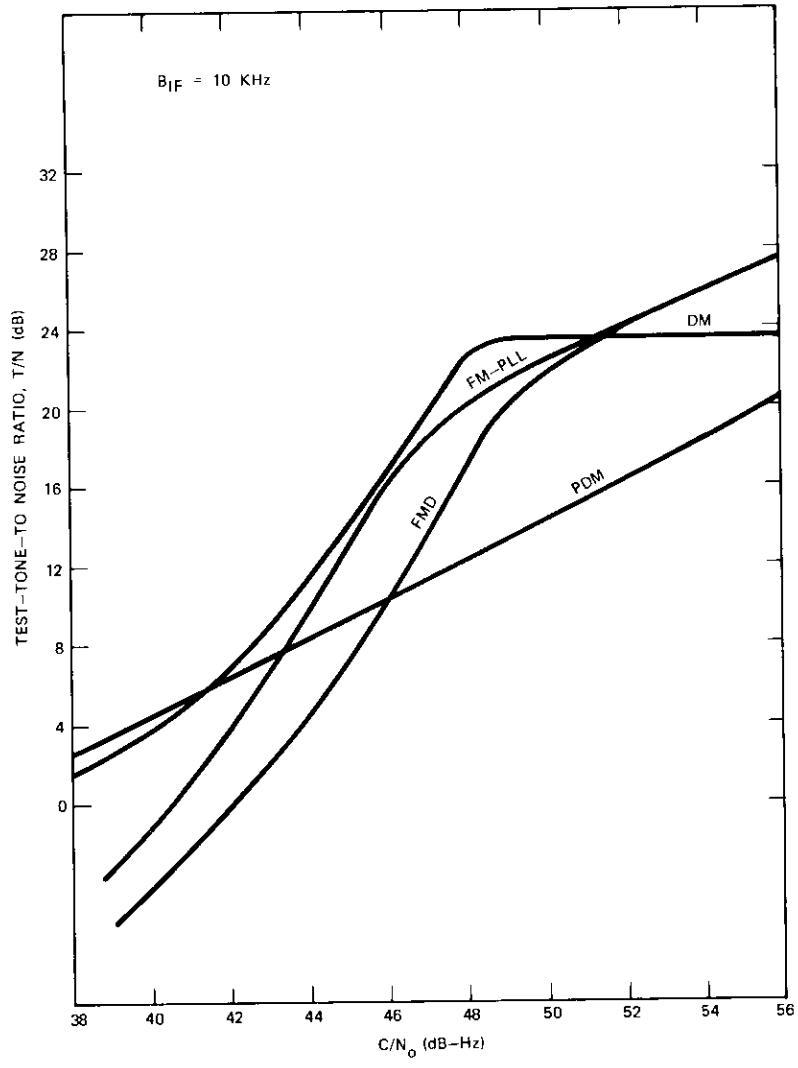


Figure 6. Test-Tone-to-Noise Ratio vs. C/N_o for 10-kHz Bandwidth

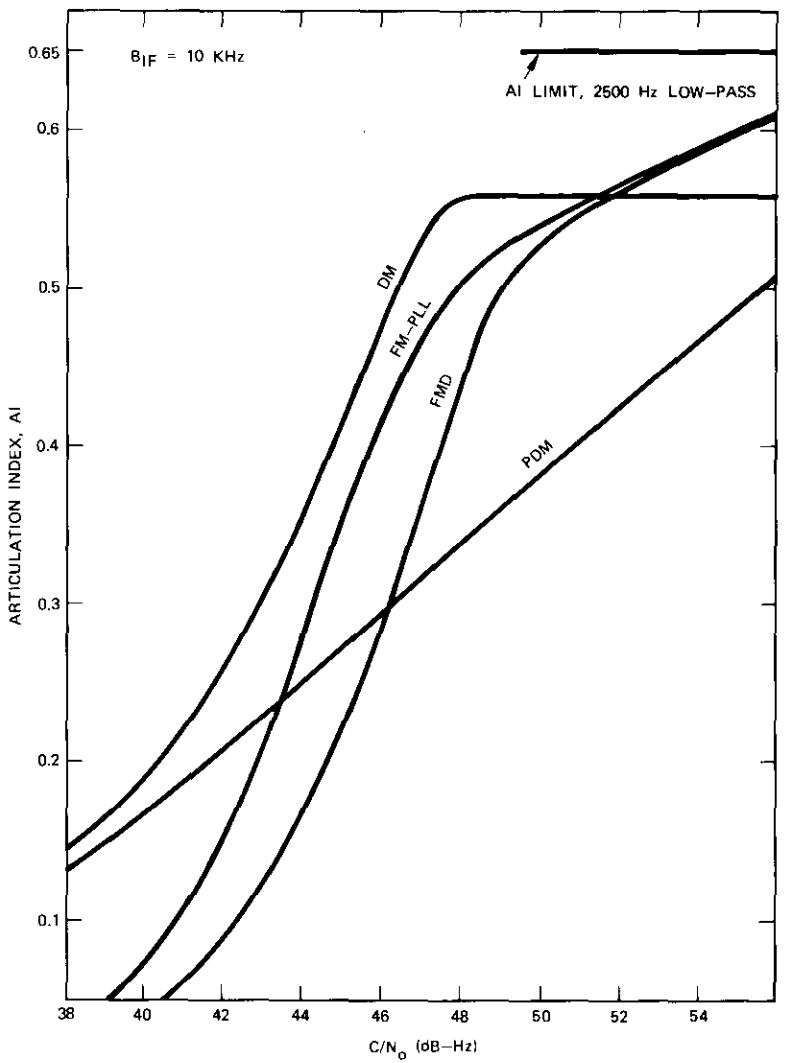


Figure 7. Articulation Index vs. C/N_0 (B_{IF} = 10 kHz)

PULSE-CODE MODULATION (PCM)/PHASE-SHIFT KEYING (PSK)

General

In the past two decades, pulse-code modulation (PCM) has been introduced for the transmission of a wide variety of signal sources including voice, data, and image signals. PCM is attractive because of an advantageous exchange of bandwidth and signal-to-noise ratio as well as the possibility of utilizing efficient and flexible multiplexing techniques. Similar to FM, the signal-to-noise performance of PCM can be substantially improved by increasing the transmission bandwidth.

A PCM signal is developed by sampling the voice baseband signal at no less than twice the rate of the highest voice frequency and quantizing the amplitude of each sample by assigning an M-bit binary code word. The resulting bit sequence is used to determine the phase states of the transmitted RF carrier, thereby producing a phase-shift keyed (PSK) signal.

Expression for PCM Performance

The parameters which determine the performance of the PCM/PSK system, and hence the received voice quality, are the sampling rate, f_s; the number of quantization bits per sample, M; and the bit-error probability, P_e, resulting from channel noise. As long as the sampling rate satisfies the Nyquist criterion, the noise caused by sampling error may be neglected. The system performance is therefore determined by quantization noise, N_q, and thermal noise, N_T. If it is assumed that these noise sources are independent and additive, the output signal-to-noise ratio of a PCM/PSK system can be expressed as

$$\left(\frac{S}{N}\right)_o = \frac{1}{(N_q/S) + (N_T/S)} . \tag{13}$$

In the following, expressions will be obtained for S/N_T and S/N_q. Uniform quantizing (i.e., equal step size throughout the quantizing range) is assumed. If companding is incorporated, then some sacrifice in performance at high signal levels will occur; however, this will be offset by improved performance for low-level talkers. First, the signal-to-noise contribution due to thermal noise and hence bit error is determined. For uniformly quan-

tized signals [8] and a test tone that occupies the peak-to-peak quantizing range, it can be shown [4] that

$$S/N_T \simeq \frac{1}{4P_e} \tag{14}$$

where P_e is the bit-error probability. For two-phase coherent PSK modulation,

$$P_e = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\gamma}\right) \tag{15}$$

The expression for the signal-to-quantization noise ratio depends on the type of input signal. For example, for a uniformly distributed signal,

$$\frac{S}{N_Q} = 2^{2M} \tag{16}$$

where M is the number of bits per sample. For sinusoidal signals it can be shown that

$$\frac{S}{N_Q} \simeq \frac{3}{2}\left(2^{2M}\right). \tag{17}$$

From equations (13), (14), and (17), the output test-tone-to-noise ratio for an M-bit PCM system is given by

$$\left(\frac{T}{N}\right)_{PCM} \simeq \frac{3(2^{2M})}{2[1 + 6P_e 2^{2M}]}. \tag{18}$$

PCM in 20-kHz Bandwidth With and Without Coding

Consider now the performance which may be obtained with PCM if a 20-kHz RF bandwidth is available. With two-phase PSK modulation it is assumed that the required transmission bandwidth is approximately 1.1 times the bit rate. Therefore, an 18-kbps PCM signal can be supported in a 20-kHz bandwidth. The system can be constructed by bandlimiting the voice spectrum from 200 to 2500 Hz, sampling at 6 kHz, and quantizing with $M = 3$ bits per sample. Figure 2, determined from equation (18), is a plot of output test-tone-to-noise ratio versus carrier-to-noise density ratio for this case. The flat portion of the curve (i.e., for $C/N_o > 51$ dB-Hz) is the quantization noise-limited region [9]. For $C/N_o < 51$ dB-Hz,

the performance of PCM is essentially determined by thermal noise in the channel. The region close to $C/N_o = 51$ dB-Hz is called the threshold and marks the transition from quantizing to thermal noise predominance. In this case the threshold corresponds to an error rate, P_e, of approximately 10^{-4}.

As will be discussed in a later section, rate one-half coding combined with four-phase PSK modulation makes it possible to enhance the performance of a digital communication channel for pre-detection error rates less than 10^{-3}. For the 20-kHz bandwidth, the enhancement thereby gained is shown by the curve marked "coded PCM" in Figure 2. Note that only performance near threshold is improved.

Figure 3 shows the performance of three-bit-per-sample coded PCM in terms of *AI* versus C/N_o. Curve 1 of Figure 1 is used to make the conversion. It is readily apparent that three-bit-per-sample coded PCM does not compare favorably with the other methods shown. The performance curves computed here assume an optimum speech signal level; hence, companding will provide no significant improvement [10]. Since the relative performance of PCM is poor at 18 kbps, it is not considered for use in the 10-kHz bandwidth.

PCM in 40-kHz Bandwidth With and Without Coding

The T/N versus C/N_o performance of PCM in a 40-kHz RF bandwidth is shown in Figure 4. The parameters are the same as those used in the 20-kHz bandwidth case except that a quantization resolution of six bits per sample is employed. This, of course, yields a bit rate of 36 kbps. The results of a PCM, two-phase PSK, uncoded system and a PCM, four-phase PSK, rate one-half coded system are shown. The *AI* versus C/N_o curve (Figure 5) demonstrates that the relative performance of PCM is improved in this bandwidth. In this case, companding would be advantageous for accommodating varying speech power levels caused by changes in vocal effort. Of course, the introduction of such companding causes some degradation in the signal-to-noise ratio performance of talkers whose speech occupies the full peak-to-peak range of the quantizer.

It should be noted that PCM transmission of telephone-quality voice is typically achieved by using a sampling rate of 8 kHz and nonuniform quantization of seven or eight bits, giving a final bit rate of 56 to 64 kbps. Also, PCM communication for military use is considered acceptable when using a sampling rate of 6 kHz and six or seven bits per sampling, yielding a bit rate of 36 to 42 kbps.

DELTA MODULATION (DM)/PSK

General

Similar to PCM, delta modulation (DM) encodes an analog signal into binary digits. DM is a differential encoding technique which accomplishes analog-to-digital conversion with a much simpler implementation than PCM. As with PCM, the performance of DM is determined by both quantization noise and thermal noise. Also, delta modulation with fixed slope exhibits a phenomenon known as slope overload, which occurs whenever the rate of change of input signal amplitude exceeds the slope rate during a sampling interval. This effect produces a source of error other than quantization and thermal noise. Use of variable slope delta modulation [11] greatly reduces this problem.

Expression for DM Performance

If slope overload is avoided, the signal-to-quantization noise ratio for DM can be expressed as [8]

$$\frac{T}{N_Q} = \frac{3\pi}{(\omega_T \tau)^3} \tag{19}$$

where $f_T = \omega_T/2\pi$ is the test-tone frequency and τ is the sampling interval. Slope overload can be avoided by assuming that the slope of correction introduced by the delta modulator at least equals the maximum slope expected of the signal. For a speech spectral shape that is assumed to be flat to 1000 Hz and thereafter to descend at a rate of 6 dB/octave or more, this maximum slope would occur for a 1000-Hz test tone which occupies the full peak-to-peak range of the delta modulator.

The signal-to-thermal noise ratio caused by errors in the bit stream is

$$\frac{T}{N_T} \simeq \frac{\omega_0 \pi}{8\omega_T^2 P_\epsilon \tau} \tag{20}$$

where $f_0 = \omega_0/2\pi$ is the low-frequency cutoff of the speech signal and P_ϵ is given by equation (14) for binary coherent PSK. T/N_T is dependent on f_0 because the power spectral density of the noise at the input to the baseband filter at the receiver is proportional to $1/\omega^2$. Since the noise power near $\omega = 0$ is inordinately high, the low-frequency cutoff must be carefully assigned. Using equation (13) with equations (19) and (20) results in the following expression for the DM output test-tone-to-noise ratio:

$$\left(\frac{T}{N}\right)_{DM} = \frac{1}{1 + \dfrac{3\pi/(\omega_T\tau)^3}{\{24P_\epsilon/[(\omega_T\tau)(\omega_0\tau)]\}}} \tag{21}$$

Newer versions of DM codecs employ adaptive step size control which results in delta modulation with a variable slope, and hence optimal adjustment of the delta modulator's slope for all signal levels. One method [11] which accomplishes this step size control, by using digital feedback, has significantly improved DM performance, especially with respect to the slope overload problem. In fact, it has been claimed that commercial-quality speech can be achieved by using DM at about 40 kbps, provided that the error rate is not excessively high. For aeronautical and marine mobile communications via satellite, this form of DM may provide acceptable voice communications at 18 kbps.

DM in 20-kHz Bandwidth With and Without Coding

Performance of DM for a sampling rate of 18 kHz, i.e., $\tau = 55$ μs; a low-frequency cutoff, f_0, equal to 200 Hz; a test-tone frequency, f_T, of 1 kHz; and two-phase PSK is shown in Figure 2. Also shown is a curve giving the performance of coded DM using rate one-half coding with four-phase PSK modulation. Applying the transformation of Figure 1 and using curve 1 above threshold (quantization noise has a flat spectrum) and curve 2 below threshold (thermal noise decreases at 6 dB per octave) yields the curves of AI versus C/N_0 shown in Figure 3.

DM in 10-kHz Bandwidth

It is possible to realize transmission of 18-kbps delta modulation by using four-phase PSK modulation in a bandwidth of 10 kHz. The relationship between T/N and C/N_0 for this case is shown in Figure 6, and the conversion to AI, using the same technique as in the 20-kHz bandwidth case, is shown in Figure 7.

DM in 40-kHz Bandwidth

The performance of DM at 36 kbps is shown in Figures 4 and 5 for a 40-kHz bandwidth. The parameters are identical to those used in the 10-kHz and 20-kHz cases except that the sampling rate has been increased to 36 kHz. Both a DM, two-phase PSK, uncoded system and a DM, four-phase PSK, rate one-half coded system are considered. Note that performance above threshold is better than in the 10-kHz and 20-kHz cases.

CHANNEL ENCODING AND MULTIPHASE SIGNALING

Satellite communication for aeronautical and marine applications is typified by a power-limited channel. Bandwidth limitation may also be a problem in some system designs. The performance of binary coded systems such as PCM and DM can be improved somewhat in such an environment through the use of channel encoding and/or multiphase signaling. For example, for coded systems it has been demonstrated [12] that, through the use of rate one-half convolutional encoding (one parity check bit per information bit) with threshold decoding, a performance improvement can be achieved for predetection error rates less than 10^{-3}. Of course, the use of coding increases the transmitted bit rate by a factor which is the inverse of the rate of the code. For example, rate one-half coding of an 18-kbps DM signal results in a 36-kbps transmission bit rate. Four-phase PSK modulation is then employed for the RF transmission; the resulting symbol rate (as well as the information rate) is 18 kHz. This symbol rate can still be supported in the original 20-kHz bandwidth.

The test-tone-to-noise and *AI* performance curves for PCM and DM at information rates of 18 and 36 kbps with rate one-half coding and four-phase PSK signaling were obtained by using the preceding method. Note that coding provides a 1.5-dB C/N_o improvement near threshold. Below threshold the performance of the coded systems degrades rapidly because the coding improvement decreases rapidly for predetection error rates less than 10^{-3}. In fact, for predetection error rates greater than 10^{-2}, the coded system results in performance degradation compared to an uncoded system.

Using four-phase PSK reduces the bandwidth of the binary coded channel by one-half. Thus, a rate one-half coded channel can be accommodated in the same bandwidth as the two-phase PSK modulated uncoded channel. Furthermore, an 18-kbps delta-modulated signal can be supported by a 9-kbaud symbol rate in a bandwidth of 10 kHz. This method is used to obtain the DM curves shown in Figures 6 and 7.

PULSE-DURATION MODULATION (PDM)

General

In conventional pulse-duration modulation (PDM), the values of samples taken periodically on an input time function are transmitted by modulating the width of a pulse occurring at the same periodicity [13]. In this method the leading edge of the pulse always occurs at equally spaced time intervals, whereas the trailing edge varies in accordance with the modulating signal amplitude. If the sampling rate is F, then the average interval between positive-going and negative-going transitions will be $(2F)^{-1}$.

Suppressing the periodic leading edges prior to transmission by developing a bipolar signal that reverses its polarity at the time of occurrence of each trailing edge of the conventional PDM pulse train makes it possible to realize improved performance. This doubles the average interval between positive-going and negative-going edges so that it becomes F^{-1}. The advantage of this operation is a reduction in postdetection noise at the receiver. The leading edge information is retrieved at the receiver by a phase-locked loop that averages received signal polarity changes.

The modulation method for transmission of the pulse is assumed to be two-phase PSK with coherent reception. In this case, for one state of the bipolar modulating signal, a sinusoidal carrier is transmitted and, for the other state, the carrier changes phase by π radians.

Performance of PDM in Bandwidths of 10, 20, and 40 kHz

Performance of the PDM method described previously for a test tone adjusted to cover the full peak-to-peak modulator range can be expressed by the following relationship:

$$(T/N)_{PDM} = \gamma \frac{B_{IF}^2}{4F^2 f_m} . \tag{22}$$

The derivation of this relationship, given in Appendix A, takes into account only the perturbation of the information-bearing axis crossings caused by noise. Other effects occur at low values of signal-to-noise ratio which are ignored; the values obtained by excluding these effects hence are optimistic.

For an IF bandwidth of 20 kHz, a maximum voice frequency of 2500 Hz, and a sampling rate of 6000 Hz, the performance curve of T/N versus C/N_o for PDM is shown in Figure 2. Performance in terms of *AI* versus C/N_o, determined by using curve 1 of Figure 1 for flat noise, is shown in Figure 3. As revealed by these curves, the PDM system exhibits good performance at low values of C/N_o. Also shown in Figure 2 are the results of measurements made on an experimental PDM unit designed for 20-kHz bandwidth.

The T/N versus C/N_o characteristic for PDM in a 10-kHz bandwidth is shown in Figure 6. Figure 7 is the corresponding AI versus C/N_o plot. In this case, the performance of PDM is not favorable at high values of C/N_o, but is comparable to FM and DM at very low values of C/N_o.

The results for PDM operating with $B_{IF} = 36$ kHz are shown in Figures 4 and 5. For $C/N_o < 51$ dB-Hz, PDM seems to deliver the best performance of all systems considered. It is important to note, however, that equation (22) does not include the degradation caused by the reduced performance of the carrier recovery loop at low signal-to-noise ratios.

CONCLUSIONS

The principal results of this paper appear in Figures 3, 5, and 7, which show the relative performance of DM, PCM, FM, and PDM in channel bandwidths of 20, 40, and 10 kHz, respectively, in terms of AI versus C/N_o. These bandwidths were chosen because they correspond to the useful bandwidths for the channel assignments that are currently of interest to proponents of aeronautical and marine satellite communications. The results are also summarized in Table 1 for low, intermediate, and high levels of C/N_o for all three bandwidths.

In the 20-kHz bandwidth case, it is seen that FM-PLL and DM provide excellent performance for values of $C/N_o > 49$ dB-Hz and that PDM holds up well for very low values of C/N_o. It should be noted that the performance predictions for PDM at low values of C/N_o are considered to be optimistic for the reasons stated in the text. The other methods degrade more rapidly for $C/N_o < 49$ dB-Hz, with DM ranking next to PDM. These relative performances are achieved by using the most suitable method of modulation for each technique and by assuming there has been no loss of lock of the PLL's required for bit-timing recovery. Frequency modulation with the discriminator detector (FDM) is not as effective as the other methods except at high values of C/N_o.

In the 10-kHz bandwidth case, the difference in performance between FM, DM, and PDM decreases. In fact, FM would probably provide acceptable performance. DM delivers the highest values of AI for all values of $C/N_o < 51$ dB-Hz. PDM is significantly limited in performance at values of $C/N_o > 44$ dB-Hz. PCM performance (not shown) is poor compared to that achieved by the other methods because of the small number of quantizing levels available and the attendant high level of quantizing noise. This is primarily a consequence of the bandwidth limitation.

A comparison of the relative shifting of curve positions in the 10-kHz and 20-kHz cases (compare Figures 3 and 7) indicates that both DM and PDM derive a greater advantage from the increase in bandwidth than either FMD or FM-PLL. Note also that PCM is included on the 20-kHz bandwidth curves, but that its performance for both uncoded and coded cases is poor.

Comparing the 20-kHz results with the 40-kHz results (compare Figures 3 and 5) shows that the digital techniques, particularly PCM, deliver significant improvements in performance in the vicinity of threshold. For both the 20- and 40-kHz cases, the general result is that the digital techniques (i.e., PCM and DM) deliver good performance at high and intermediate values of C/N_o, PDM gives the best performance at low

TABLE 1. AI PERFORMANCE SUMMARY

Modulation Techniques	ARTICULATION INDEX								
	10-kHz Bandwidth			20-kHz Bandwidth			40-kHz Bandwidth		
	42 dB-Hz	47 dB-Hz	52 dB-Hz	45 dB-Hz	50 dB-Hz	55 dB-Hz	48 dB-Hz	53 dB-Hz	58 dB-Hz
Delta Modulation (uncoded)	0.255	0.53	0.56	0.313	0.545	0.56	0.375	0.65	0.65
Delta Modulation (coded)	—		—	0.11	0.56	0.56	—	0.65	0.65
FM Discriminator	0.085	0.437	0.56	0.135	0.44	0.63	0.185	0.505	0.65
FM-PLL	0.15	0.36	0.56	0.205	0.575	0.63	0.295	.645	0.65
Pulse-Duration Modulation	0.237	0.33	0.425	0.4	0.507	0.58	0.55	0.61	0.645
PCM (uncoded)	—	—	—	0.245	0.438	0.455	0.245	0.52	0.65
PCM (coded)	—	—	—	0.175	0.45	0.455	—	0.625	0.65

C/N_o, and FM delivers the best performance only at the very high values of C/N_o (i.e., $C/N_o > 56$ dB-Hz).

Digital voice transmission techniques possess some additional characteristics desirable for mobile communications via satellite. First, because PCM and DM transmit samples in digital form, satellite interference effects due to intermodulation distortion in the satellite transponder will be manifested as a small random noise, rather than the intelligible crosstalk which occurs with FM. Digital communications systems such as DM and PCM can be efficiently applied in CDM, TDM, or FDM demand-assignment systems. Finally, digital speech processing systems utilize bit-by-bit synchronous transmission. Therefore, they can easily support high-rate data communications for a single source or on a multiplex basis. PDM possesses similar data communication advantages since a synchronous clock is part of its implementation.

List of Symbols

AI	Articulation Index
B_{IF}	IF bandwidth
c	Speech-rms-to-test-tone-rms ratio
C/N_o	Carrier-to-noise density ratio, dBHz
F	Sampling rate in PDM expressions
f_d	Peak deviation frequency
f_m	Maximum modulating frequency
f_s	Sampling rate in PCM expressions
$F(\omega)$	Normalized signal power spectrum
k_a	Gain constant
M	Number of bits per sample in PCM case
N_Q	Quantizing noise
N_T	Thermally induced noise
$N(\omega)$	Noise power spectrum in audio band
P_e	Bit error rate
P_S	Total speech power
P_T	Test tone power
p_n	Fraction of noise power in nth AI band
q_n	Fraction of speech power in nth AI band
$S(\omega)$	Signal power spectrum
T	Test tone power
$(T/N)_x$	Output test-tone-to-noise ratio for the technique designated by subscript x

β	f_d/f_m
$\Delta_n AI$	AI contribution in nth AI band
δ	$B_{IF}/2f_m$
γ	Carrier-to-noise ratio in the IF bandwidth
τ	Sampling interval
ω_1, ω_2	Speech spectrum corner frequencies, rad/sec.
ω_m	Upper audio cut-off frequency
ω_o	Lower audio cut-off frequency
ω_T	Test tone frequency in rad/sec.

REFERENCES

[1] K. D. Kryter, "Validation of the Articulation Index," *Journal of the Acoustical Society of America*, Vol. 34, No. 11, November 1962, pp. 1698–1702.

[2] K. D. Kryter, "Methods for Calculation and Use of A. I.," *Journal of the Acoustical Society of America*, Vol. 34, No. 11, November 1962, pp. 1689–1697.

[3] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *Journal of the Acoustical Society of America*, Vol. 19, No. 1, January 1947, pp. 90–119.

[4] W. Wathen-Dunn and D. W. Lipke, "On the Power Gained by Clipped Speech, *Journal of the Acoustical Society of America*, Vol. 30, No. 1, Jan. 1958, pp. 36–40.

[5] O. Shimbo and C. Loo, "A Simple Formula for the Threshold Characteristics of FM Signals," *Proceedings of IEEE* (Correspondence), Vol. 56, No. 7, July 1968, pp. 1241–1242.

[6] O. Shimbo, "Threshold Characteristics of FM Signals Demodulated by an FM Discriminator" [Correction to article in *IEEE Trans. Information Theory*, Vol. IT–15, Sept. 1969, pp. 540–549], *IEEE Transactions on Information Theory*, Vol. IT–16, No. 6, Nov. 1970, p. 769.

[7] P. W. Osborne and D. L. Shilling, "Threshold Analysis of Phase-Locked-Loop Demodulators Using Most Likely Noise," *IEEE Transactions on Communications Technology*, Vol. COM–19, No. 1, February 1971, pp. 31–41.

[8] H. Taub and D. L. Shilling, *Principles of Communication Systems*, New York: McGraw-Hill Book Co., 1970.

[9] W. R. Bennett, "Spectra of Quantized Signals," *Bell System Technical Journal*, Vol. 27, No. 3, July 1948, pp. 446–472.

[10] Members of Bell Laboratories Technical Staff, *Transmission Systems for Communications*, Bell Telephone Laboratories, 4th ed., 1970, Winston-Salem, N.C.: Western Electric.

[11] J. A. Greefkes, "A Digitally Controlled Delta Codec for Speech Trans-

mission," *IEEE 1970 International Conference on Communications*, San Francisco, June 8–10, 1970 [Proceedings], pp. 7–33 to 7–48.

[12] E. Cacciamani, "A Channel Unit for Digital Communication in the SPADE System," *IEEE 1971 International Conference on Communications*, Montreal, June 14–16, 1971 [Proceedings], pp. 42-15 to 42-20.

[13] M. H. Nichols and L. R. Rauch, *Radio Telemetry*, 2nd ed., New York: John Wiley & Sons, Inc., 1956, pp. 60–64.

APPENDIX A. DERIVATION OF A RELATIONSHIP FOR PDM PERFORMANCE

INTRODUCTION

The following derivation has been developed for computation of the sinusoidal test-tone-to-noise ratio of a PDM system. Terms used in the derivation are defined as follows:

F = sampling rate
B_{IF} = IF bandwidth
N_o = noise power spectral density in the IF bandwidth
$N = N_o B_{IF}$ = total noise power in the IF bandwidth
h = peak-to-peak amplitude swing following detection at the receiver
C = carrier power
σ_n = noise standard deviation, postdetector
S = maximum signal slope, postdetector
σ_Z = standard deviation in observation time of the zero-axis crossing.

The derivation proceeds by first determining an expression for the time variance of axis crossings at the receiver detector for a PDM system and then determining the mean-squared time variation due to a test-tone sinusoid that occupies the full peak-to-peak range of the modulator. Figure A-1 is a block diagram of the PDM system under consideration.

DETERMINATION OF VARIANCE IN RECEIVED AXIS CROSSINGS

For the moment, consider the output resulting from detection at the receiver. Assume that

a. the detected signal corresponding to the full range of modulation traverses the range from $-h$ to $+h$;

b. a zero-axis crossing detector is used to indicate the instant of occurrence of the zero-axis crossing;

c. a Gaussian noise of standard deviation σ_n is perturbing the detected signal; and

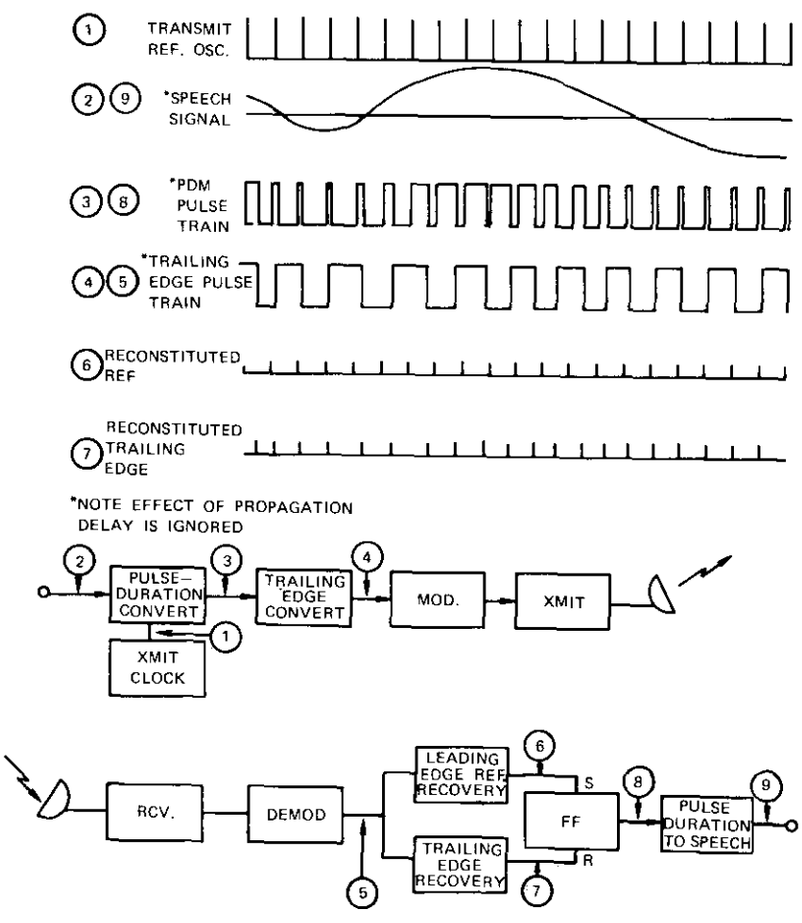


Figure A-1. Block Diagram of the PDM System

d. the bandwidth is limited by an IF filter of width B_{IF}, which at baseband is equivalent to a low-pass filter having a bandwidth of $B_{IF}/2$.

The standard deviation in the time of occurrence of the zero-axis crossing may be determined as follows. The maximum slope in making the transition from $-h$ to $+h$, which is a function of the effective low-pass signal bandwidth, $B_{IF}/2$, and the peak-to-peak swing, $2h$, is given by [A1]

$$S = 2hB_{IF}. \tag{A1}$$

This maximum value occurs at a point midway between the extremes. Hence, if the swing is symmetrical about zero, it occurs at the zero-axis crossing. Therefore, the standard deviation in the time of zero crossing is obtained by dividing the standard deviation in amplitude by the slope. That is,

$$\sigma_Z = \frac{\sigma_n}{2hB_{IF}}.$$

Values of σ_n and h must now be determined. To accomplish this, it is assumed that a coherent detector is employed on a signal modulated so that one state is represented by a CW carrier and the opposite state is represented by the same CW carrier reversed in phase. The coherent detector is assumed to consist of a perfect multiplier whose reference input consists of a sinusoid $\sqrt{2E}\sin(\omega_1 t + \theta_1)$ and whose signal input consists of a sinusoid $\pm\sqrt{2C}\sin(\omega_1 + \theta_1)$. For a carrier of power C, the power output of the coherent detector due to the signal is

$$\sigma_S^2 = E^2C. \tag{A3}$$

Note that this is also the value of h^2; i.e., $h^2 = \sigma_S^2$.

The output of the coherent detector due to noise of total power N in the bandwidth B_{IF} is

$$\sigma_n^2 = \frac{E^2N.}{B_{IF}T} \tag{A4}$$

Substituting the preceding values for h and σ_n into equation (A2) gives

$$\sigma_Z^2 = \frac{N}{4TB_{IF}^3C}. \tag{A5}$$

The value of T is equal to the average duration between phase changes of the carrier. For conventional PDM, the system in which the coherent leading edge of the width-modulated pulse is retained, the value of T is $1/2F$. For a PDM system

in which the leading edge of the width-modulated pulses is suppressed, the value of T is $1/F$. Suppressing the leading edges of the pulses thus provides a 3-dB reduction in noise.

DETERMINATION OF VARIANCE FOR A TEST TONE

Equation (A5) gives the variance in terms of axis crossings and is the expression for noise in the PDM detector. It is now necessary to determine the variance (or power) in a sinusoidal test tone that is scaled to occupy the peak-to-peak time range. The full peak-to-peak range is simply $1/F$. Thus, a sinusoid whose peaks reach the limits of this range has a peak amplitude of $(2F)^{-1}$ and, of course, a variance of $(8F^2)^{-1}$.

TEST-TONE-TO-NOISE RATIO

Equation (A4) for σ_n^2 gives the noise power residing in the Nyquist bandwidth, $F/2$. If, following detection, the detected signal is again low-passed by a filter with a cutoff at f_m, then the noise is further reduced in power by a factor of $2f_m/F$. The test-tone-to-noise ratio is then the ratio of the sinusoidal signal variance to the noise power in the low-pass filter. Thus,

$$\left(\frac{T}{N}\right)_{PDM} = \gamma\frac{B_{IF}^3}{4F^2f_m}$$

where $\gamma = C/N$. This derivation assumes perfect recovery of the leading edge phase reference by the PLL at the receiver.

REFERENCE

[A1] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *Journal of the Acoustical Society of America*, Vol. 19, No. 1, January 1947, pp. 90–119.

Dr. S. J. Campanella received a B.S.E.E. from Catholic University of America, a M.S.E.E. from the University of Maryland, and the Ph.D. in electrical engineering from Catholic University of America. Before joining COMSAT, he was Manager of the Electronics Research Center at Melpar, Inc.

Presently, he is Manager of the Baseband Processing Branch of the Communications Processing Laboratory, COMSAT Laboratories, where he is responsible for communications signal processing research and development.

Dr. Campanella is a member of IEEE, Sigma Xi, Phi Eta Sigma, and the Acoustical Society of America.

Joseph A. Sciulli received the B.S.E.E. from the University of Pittsburgh and the M.S.E.E. from the University of Maryland. From 1962 to 1968 he was employed by the National Aeronautics and Space Administration's Goddard Space Flight Center, where he engaged in image signal processing research and digital communications system development. In 1968 he joined COMSAT Laboratories, where he is presently head of the Information Techniques Section of the Baseband Processing Branch. His current efforts are directed toward the development of signal processing techniques for speech and image signals.

Mr. Sciulli is a member of IEEE, Eta Kappa Nu, Sigma Tau, and Omicron Delta Kappa.

Index: communication satellites, earth terminals, radio frequency interference, radio relay stations

Feasibility of collocating a radio relay station with a sharing earth station

JAMES B. POTTS

ABSTRACT

Sharing of frequency bands between earth stations of the Fixed Satellite Service and terrestrial stations of the Fixed Service is commonplace. Nevertheless, it is usually assumed that this situation is possible only with large geographic separations. Recent studies indicate that conditions could be established which would permit small geographic separations. In fact, this study shows that, for practical purposes, the respective stations can be collocated without experiencing unreasonable interference levels. A particular case is analyzed to demonstrate the feasibility of this concept.

INTRODUCTION

It is generally assumed that sharing of frequency bands between terrestrial stations of the Fixed Service and earth stations of the Fixed Satellite Service [1] requires relatively large separations between these stations to avoid unreasonable interference. The highly developed terrestrial network in many areas, especially common near traffic centers, presently results in remote locations for the earth stations. Consequently, the cost of providing interconnection facilities from the earth stations to the traffic centers becomes significant.

Because earth station interconnection facilities in the United States commonly use radio-relay systems operating in the 10.7- to 11.7-GHz frequency band, frequency sharing is apparently precluded in this band.

In the light of recent frequency allocations for satellite use in this band, it is of considerable importance to determine whether these two uses are really mutually exclusive. It appears that certain radio-relay techniques [2] might make it possible to collocate a receiving earth station antenna with a radio-relay station antenna. This paper is a study of the conditions under which this technique appears feasible.

COUPLING BETWEEN COLLOCATED ANTENNAS

Measured antenna far-field radiation patterns and measured "near-field" coupling may be used to determine the amount of antenna isolation which can be achieved in practice for various types of antennas. The data which will be given in the following subsections will provide the requirements for such a practical model.

Earth Station Antenna Patterns

For large antennas ($D/\lambda > 100$), the mean value of the peaks of sidelobe gain in the rear sector is usually assumed to be -10 dBi. This assumption is based on the CCIR reference radiation diagram [3], which is in turn a conservative estimate based on limited supporting data.

Several measured earth station antenna radiation patterns taken at different frequencies have been analyzed; the results are presented in Figure 1. The points represent the sidelobe peaks in the rear 180° (i.e., from 90° to 270° from the peak of the main beam) for two different antennas. From these data, it may be concluded that a rear sector gain of -15 dBi is quite reasonable, particularly at higher frequencies.

Radio-Relay Antenna Patterns

This study has assumed that horn antennas will be used in the radio-relay system to achieve the maximum isolation. The proposed reference pattern for these horns, taken from Reference [4], is

$$\begin{aligned} G(\phi) &= 41 - 35 \log \phi \text{ (dBi)} && 0 \leq \phi \leq \pm 40^\circ \\ &= -15 \text{ dBi} && \pm 40^\circ \leq \phi \leq \pm 100^\circ \\ &= -25 \text{ dBi} && \pm 100^\circ \leq \phi \end{aligned} \quad (1)$$

where $G(\phi)$ is the radio-relay antenna pattern gain. Measurements on a particular horn antenna design at 11 GHz yield sidelobe levels which are much better than those included in this reference pattern, i.e., generally lower than -40 dBi.

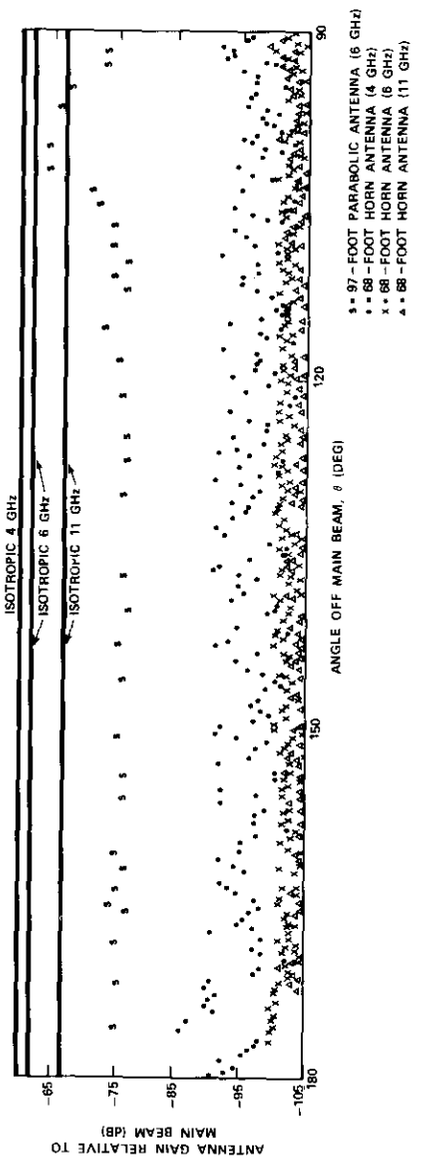


Figure 1. Peaks in the Rear Section of Large Earth Station Antennae

Near-Field Effects

Because the collocated antennas are so close together, it was felt that the use of far-field patterns to calculate the coupling might yield questionable results. However, as shown in the following example, coupling values computed by using far-field patterns and free-space transmission losses compare favorably with measured values, predicting coupling values higher than those actually measured.

Measured coupling loss values and far-field patterns for an 11-GHz antenna are described in Reference [5]. From the antenna geometry given in this reference, the coupling loss can be computed by assuming free-space transmission in conjunction with the far-field patterns. This computation has been performed for the two relative orientations: pointing in the same direction, and pointing in opposite directions. The results are as follows:

Case	Measured Value (dB)	Computed Value (dB)	Δ (dB)
1	82	72	10
2	107	105.4	1.6

A comparison of the measured and calculated values shows that the measured coupling loss exceeds the calculated coupling loss. Hence, it is concluded that "near-field" effects have not introduced errors which will lead to incorrect conclusions.

Coupling measurements were made in the rear sector of the earth station antenna at Etam, West Virginia, at 6 GHz. Figure 2 shows a cone whose surface intersects the circumference of both the antenna subreflector and main reflector and whose vertex is on the boresite axis. Measurements were made within the frustum formed by the intersection of the main reflector and the cone. Figure 3 indicates the values measured on an arc 100 feet behind the antenna. Figure 4 relates all of the measured values to the azimuth angle. Measured antenna gains were typically -20 to -30 dBi; these values are well below those indicated on the CCIR reference radiation diagram.

EXPECTED INTERFERENCE LEVELS

The expected interference level can be determined as follows:

$$I = P_t + G_t(\phi) + G_r(\theta) - L_b \qquad (2)$$

where I = interfering carrier power received, dBW

P_t = radio-relay transmitter power, dBW

$G_t(\phi)$ = gain of transmitting radio-relay antenna, dBi

$G_r(\theta)$ = gain of receiving earth station antenna, dBi

L_b = basic transmission loss, dB.

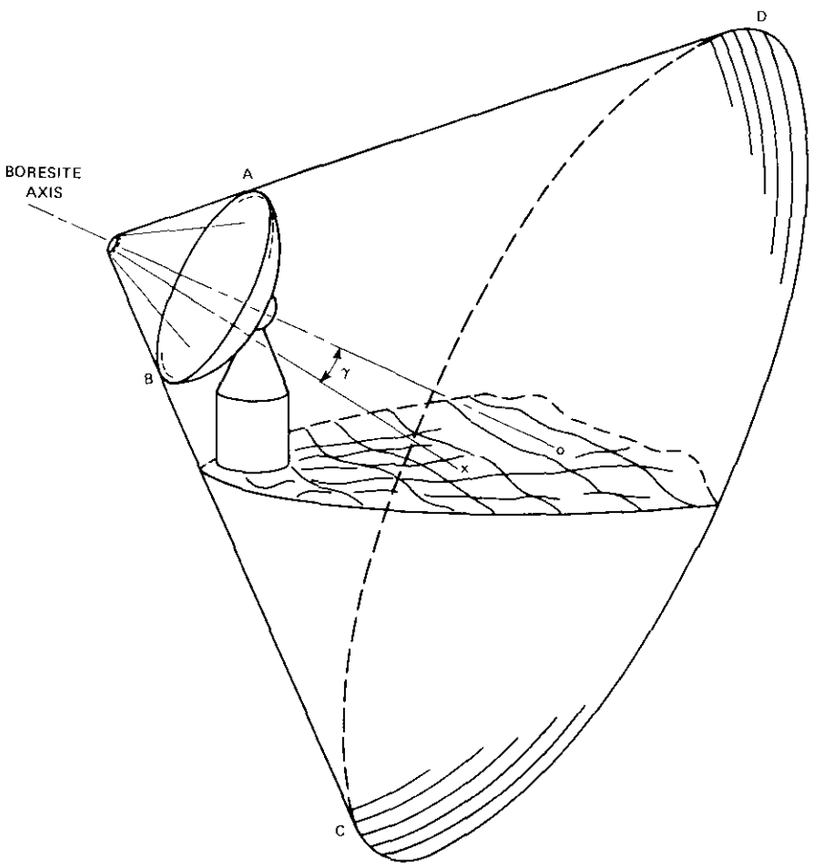


Figure 2. Region of Measurement (frustum = ABCD)

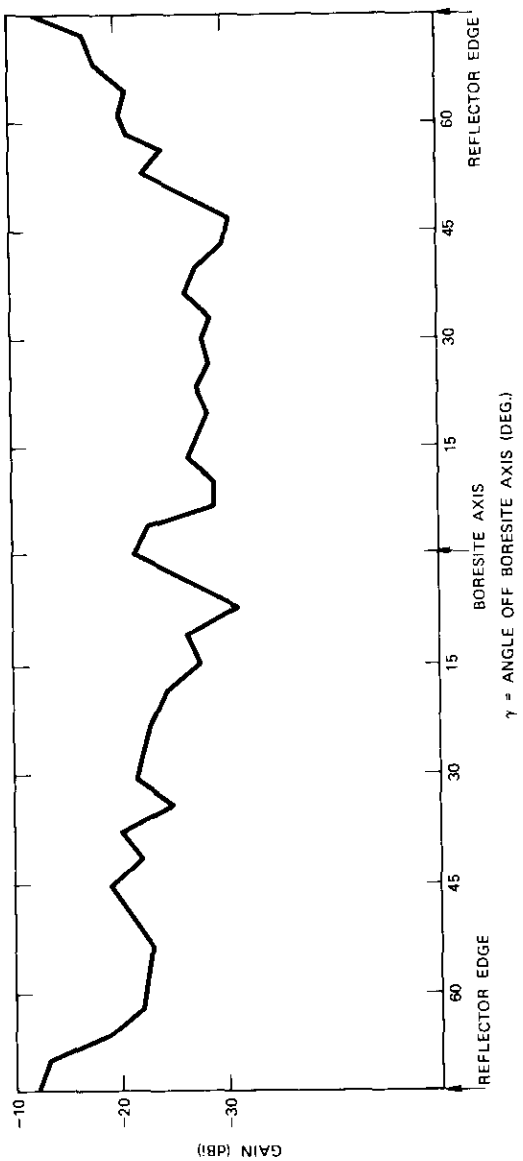


Figure 3. Ground Level Measurements on an Arc Approximately 100 Feet Behind the Antenna

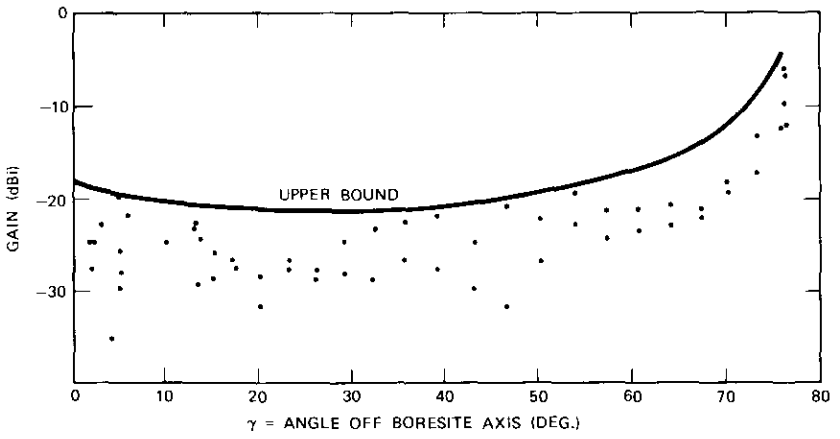


Figure 4. Measured Gain vs. Angle Off Boresite Axis

Interference into the receiving earth station from the radio-relay transmitter is the primary concern. Since all cases of interest involve line-of-sight transmission, L_b is simply the free-space transmission loss.

The interference levels predicted by equation (2) are evaluated for two physical configurations of interest using a radio-relay system at 11 GHz with one watt of transmitter power and an antenna with radiation patterns defined by equation (1). Configuration I is a simple direct link with a collocated radio-relay terminal. A practical separation between the earth station antenna and the collocated station is assumed to be 160 m, although for some unusual cases a separation of up to 1.6 km could be considered.

Configuration II is a remote radio-relay terminal within line of sight and pointing in the general direction of the earth station. The distance to the remote radio-relay station is assumed to be 16 km. For this configuration, the use of a passive reflector located about 1.6 km from the earth station antenna (as shown in Figure 5) can provide additional angular discrimination for the remote radio-relay station. Hence, equation (2) would be modified for this configuration as follows:

$$I = P_t + G_t(\phi) + G_p(\alpha) + G_r(\theta) - L_b(d_2) - L_b(d_3) \qquad (3)$$

where $G_p(\alpha)$ = total gain of passive reflector in dBi

$L_b(d_i)$ = basic transmission loss in dB for distance d_i.

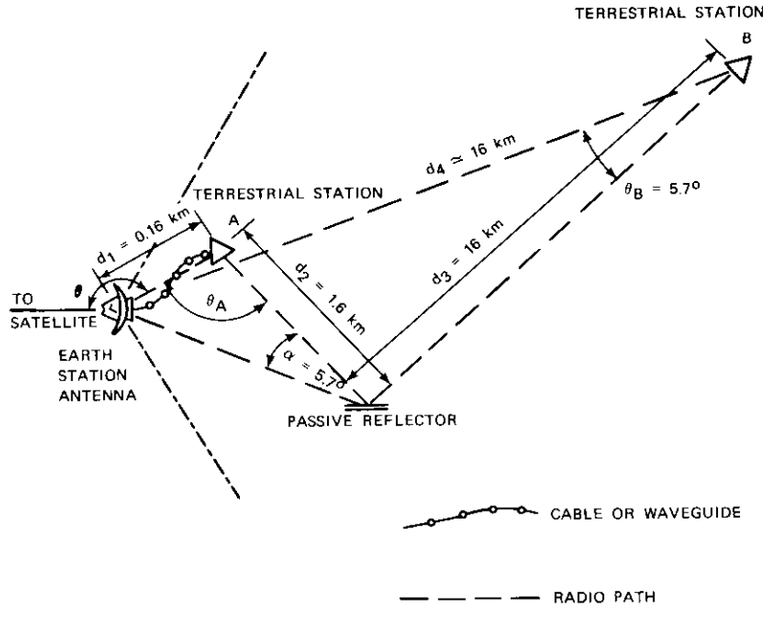


Figure 5. Entrance Link Configuration

Table 1 gives the results of analyzing these configurations, showing particular values of interference from the various sources. The gain of the passive reflector is taken from Figure 6. The efficacy of the direct path is seen to be limited by interference from the remote radio-relay station. Hence, the passive reflector appears necessary unless the antenna and the collocated terminal are separated by a relatively long distance, such as 1.6 km.

RADIO-RELAY LINK DESIGN AT 11 GHz

The radio-relay link design departs from conventional radio-relay link design in two important respects to result in more economical terrestrial links and reduced interference potential:

a. transmitter power is reduced to the lowest possible level consistent with fade margin requirements, and

b. modulation indices are increased to reduce thermal noise. (Indices are preferably the same as those in the satellite system.)

TABLE 1.

a. Configuration I

Source	Distance (km) d_1	Distance (km) d_2	Transmitting Radio-Relay Antenna Gain (dBi)	Receiving Earth Station Antenna Gain (dBi)	Received Interfering Carrier Power (dBW)
Collocated	0.16	—	−25	−15	−138
Remote	—	16	48	−15	−105
Passive	—	—	—	—	—

b. Configuration II

Source	Distance (km) d_1	Distance (km) d_2	Distance (km) d_3	Transmitting Radio-Relay Antenna Gain (dBi)	Gain of Passive Reflector (dBi)	Receiving Earth Station Antenna Gain (dBi)	Received Interfering Carrier Power (dBW)
Collocated	0.16	—	—	−25	—	−15	−138
Remote	—	—	16	8	—	−15	−138
Passive	—	1.6	—	+15	—	−15	−143

Sample System Parameters

System requirements are calculated for 1200 channels using the basic configuration of Figure 5 and horn antennas with a gain of 48 dB. A channel thermal noise performance of 12.5 pWp is assumed for each radio-relay link. The link equation is

$$S/N = 20 \log m + 228.6 - 10 \log T - 10 \log b + W + C \quad (4)$$

where m = rms test-tone deviation (unpre-emphasized)
T = system noise temperature = 4000°K
b = channel bandwidth (3.1 kHz)
W = psophometric weighting factor = 2.5 dB
C = received carrier power
= −66.6 dBW for the test-tone deviation of 600 kHz rms.

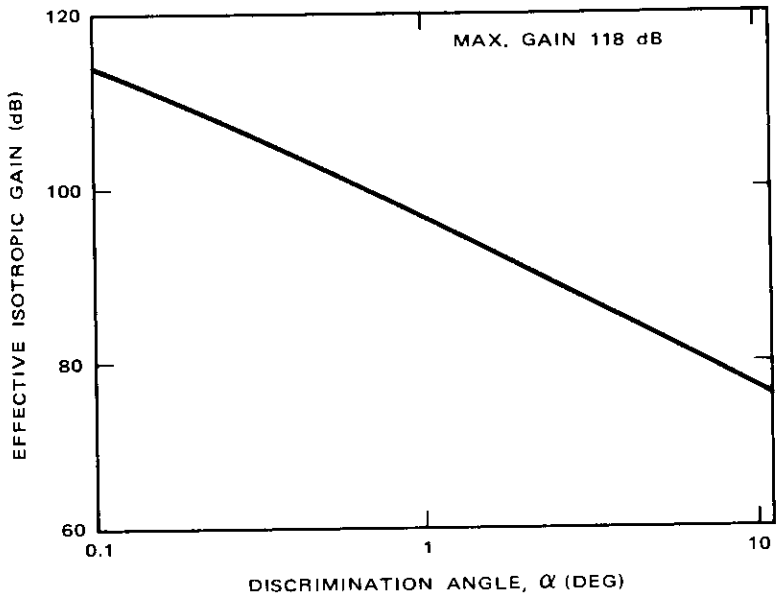


Figure 6. Gain Pattern of a Plane Reflector (24 x 30 ft) for an Incidence Angle of 45° as a Function of the Discrimination Angle Off the Reflected Beam Axis, 11 GHz

The transmitter power requirements are determined from

$$P_t = C + L_b(d_2) + L_b(d_3) + 2L_f - G_t - G_r - G_p \qquad (5)$$

for G_t, G_r, and G_p appropriate to the geometry of the configuration, where L_f is the loss in transmission feed lines in dB.

Table 2 summarizes these calculations for two different modulation indices. It can be seen that transmitter power requirements necessary to obtain the proper thermal noise yield an adequate C/N margin* and that 40-MHz channel bandwidths are also adequate.

* The fade margin requirements for a link of 16 km at 11 GHz are of the order of 40 dB above the threshold C/N, which is nominally about 10–12 dB.

TABLE 2.

Parameter	Test-Tone Deviation Ratio	
	0.187 (600 kHz)	0.062 [a] (200 kHz)
Free-Space Transmission Loss (dB)		
d_2	118	118
d_3	138	138
Loss in Transmission Feed Lines (dB)	1	1
Transmitting/Receiving Antenna Gain (dBi)	48	48
Passive Reflector Gain (dBi)	118	118
Radio-Relay Transmitter Power (dBW)	−23.6	−14
Received Carrier Power	−66.6	−57
Required Bandwidth [b] (MHz)	34	18.4
C/N in 40 MHz (dB)	50	59.6

[a] Standard CCIR deviation ratio.

[b] Computed from $B_c = 2(f_b + D_p)$, where B_c is the RF channel bandwidth in Hz, f_b is the top baseband frequency in Hz (5.3 MHz for 1200 channels), D_p is the multichannel peak deviation in Hz, and a peak factor of 10 dB is assumed.

It appears that the use of a higher modulation index is very advantageous, since it results in a link design that eliminates modulation-demodulation noise at the earth station by direct RF (or IF) connection between the earth station and the terrestrial station. For high-capacity systems, the noise is reduced by about 100 pWp, permitting increased budgets for other noise sources.

INTERFERENCE CONSIDERATIONS

The various interference cases can now be analyzed to determine expected interference levels for pre-emphasized FM, for which typical multichannel rms modulation indices, m_1 and m_2, and interference reduction factors, B, for wanted and interfering carriers are given in Table 3. The interference noise in a voice channel is determined from

$$S/N_I = C/I + B + W \qquad (6)$$

where all quantities are expressed in dB.

TABLE 3.

Number of Radio-Relay Channels, n	Multichannel rms Deviation (MHz)	Multichannel rms Modulation Index [b] (interfering carrier)	Multichannel rms Modulation Index (wanted carrier) [a]				
			0.5	0.8	1.0	1.5	2.0
			Interference Reduction Factor,[c] B (cochannel)				
1800	1.05	0.13	19	23	25	30	33
1200	1.23	0.24	19	23	25	30	33
600	0.87	0.34	19	23	25	30	33
240	0.55	0.53	19	23	25	30	33
1200 [d]	3.7	0.7	19	23	26	30	33

[a] Values are arbitrarily chosen and represent a range of useful values for satellite systems.

[b] All values are calculated from Reference [6] except for the special case of 1200 channels.

[c] Values are derived in Reference [7].

[d] Special case of more efficient design for the interconnection link.

Each parameter can take a range of values and tradeoffs between parameters can be made. However, as a specific example, the following values, based on Figure 5, have been chosen:

$P_t = -15$ dBW

$G_t(\phi) = -25$ dBi

$G_r(\theta) = -15$ dBi

m_1 = multichannel rms modulation index of wanted carrier = 0.7

m_2 = multichannel rms modulation index of interfering carrier = 0.7

$M = [m_1^2 + m_2^2]^{1/2} = 1.0$ (1200 channels, both systems)

$C = -109$ dBW (typical satellite received carrier for 1200 channels).*

* Corresponds to 37-dBW satellite e.i.r.p.; received at an earth station antenna with a gain of 60 dBi in clear weather.

The expected levels of interference from the terrestrial system to the earth station calculated for the cochannel case are:

a. For the near station, A, transmitting f_1, $I = -154$ dBW and $C/I = 45$ dB.

b. For the distant station, B, transmitting f_2, G_t is +15 dBi for 5.7° discrimination in the assumed reference pattern. Hence, $I = -154$ dBW and $C/I = 45$ dB.

c. For the passive reflector, which reflects both f_1 and f_2, the angular discrimination is again 5.7° and $G_p(\alpha)$ is 80 dB. The received interference level at f_2 is then $I = -159$ dBW and $C/I = 50$ dB. The interference level resulting from the reflection of f_1 by the passive system is negligible because of the much larger angular discrimination.

The interference reduction factor, B, is 23 dB. Hence, the following interference levels have been calculated for the three cases.

Station A	Station B	Noise (pWp) Reflection		Total	
f_1	f_2	f_1	f_2	f_1	f_2
91	91	neg.	29	91	120

Changing any of the three major parameters, P_t, G_t, or G_r, results in corresponding changes in noise levels. If an interleaving frequency arrangement, such as a 20-MHz spacing of interleaved carriers, is used, B increases to 40 dB and interference levels are 17 dB lower than those obtained for the cochannel case.

TV Interference

For TV signals, the lowest equivalent C/I ratio obtained from any of the sources is 45 dB; this is substantially higher than the usual 30-35 dB required for adequate protection.

CONCLUSIONS

The foregoing analysis has established the feasibility of frequency sharing between an earth station and a collocated radio relay station used for interconnection to the earth station. The interference conditions that result from collocation are tolerable, and since they are controlled by free-space propagation, one can calculate the actual interference levels rather simply and with a high degree of confidence. Both television and telephony

appear to be well protected and should be capable of satisfactory performance under these conditions.

It may also be feasible for the earth station to transmit in the shared frequency band. To provide experimental verification of this possibility, it is planned to conduct a "live" experiment at one of the present U.S. earth stations at both 4 and 6 GHz. This should provide further clarification of the "near-field" coupling, as well as additional data on antenna backlobe levels. If the radio-relay system is designed as described herein, use of a frequency band in this way could lead to unprecedented spectrum utilization as well as cost advantages.

Additional side issues of interest uncovered by this brief study include:

a. the relative accuracy attained by using far-field pattern data as compared to near-field coupling measurements,

b. the need for control of antenna pattern sidelobes in the rear sector for both large and small antennas,

c. the use of satellite frequency deviations on the radio-relay interconnection links, and

d. the applicability of these concepts to other shared frequencies.

REFERENCES

[1] RR No. 84–AG, SPA 2, Annex 1, p. 40, *Final Acts of the World Administrative Radio Conference for Space Telecommunications*, Geneva: International Telecommunication Union, January 1972.

[2] H. E. Curtis, "Radio Frequency Interference Considerations in the TD–2 Radio-Relay System," *Bell System Technical Journal*, Vol. 39, No. 2, March 1960, pp. 369–387.

[3] International Radio Consultative Committee (C.C.I.R.), "Radiation Diagrams of Antennae at Communication-Satellite Earth Stations for Use in Interference Studies," Report 391–1, *XIIth Plenary Assembly, New Delhi, 1970*, Vol. IV, Part 2, pp. 183–188, Geneva: International Telecommunication Union, 1970.

[4] Unpublished U.S. Submission to International Radio Consultative Committee (C.C.I.R.), 14 December 1971.

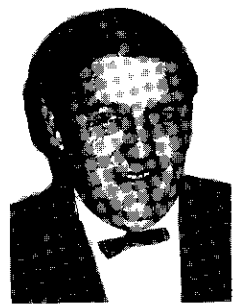
[5] A. B. Crawford and R. H. Turrin, "A Packaged Antenna for Short Hop Microwave Radio Systems," *Bell System Technical Journal*, Vol. 48, No. 6, July-August 1969, pp. 1605–1622.

[6] International Radio Consultative Committee (C.C.I.R.), "Radio-Relay Systems for Telephony Using Frequency-Division Multiplex," Recommendation 404–2, *XIIth Plenary Assembly, New Delhi, 1970*, Vol. IV, Part 1, pp. 147–148, Geneva: International Telecommunication Union, 1970.

[7] International Radio Consultative Committee (C.C.I.R.), "Techniques for Calculating Interference Noise in Terrestrial Radio-Relay Systems and Communication-Satellite Systems Carrying Multi-Channel Telephony," Report 388–1, *XIIth Plenary Assembly, New Delhi, 1970*, Vol. IV, Part 1, pp. 267–282, Geneva: International Telecommunication Union, 1970.

ACKNOWLEDGMENT

The author would like to thank Mr. Hans Weiss for his contributions to this study.



James B. Potts received B.S. and M.S. degrees in electrical engineering from Drexel University after service in the U. S. Navy during World War II as an electronic technician. He was employed by RCA from 1950 to 1963 in development engineering and communications systems engineering. He joined COMSAT in September 1963 as Chief of the Terminal Facilities Department. He assumed his present position as Director, Earth Station Implementation Division, in 1969. Mr. Potts has participated actively in the C.C.I.R. since 1963, and has been named to a number of United States delegations. He was named National Chairman of CCIR Study Group 4 in 1970.

Index: communication satellites, data transmission, earth terminals, ground support equipment, multi-channel communications, multiplexing

SPADE system progress and application

By BURTON I. EDELSON AND ANDREW M. WERTH

ABSTRACT

The SPADE demand-assignment multiple-access system will bring increased flexibility and efficiency to satellite communications. The system employs digital coding and modulation with single-channel-per-carrier frequency-division multiple access. Circuits are established on demand through use of a frequency-division switching matrix and a common signaling channel. SPADE equipment has been developed and tested over satellite links, and operational equipment is being installed at earth stations in the Atlantic region to enter into operational service with INTELSAT IV satellites. Use of the SPADE system is expected to encourage the establishment of new satellite links, as well as to increase traffic over existing links. The techniques developed in the SPADE system may have important future applications to telegraphy and data transmission, to small earth stations, and to new multidestinational services.

INTRODUCTION

The SPADE system represents the first introduction of a demand-assignment multiple-access capability in satellite communications. The

---

This paper is based upon work performed at COMSAT Laboratories under sponsorship of the International Telecommunications Satellite Consortium (INTELSAT). Views expressed herein are not necessarily those of INTELSAT.

need for such a capability stems from the desire to achieve efficient satellite capacity utilization and flexible traffic routing. These are important goals. Efficiency means a larger number of channels per satellite, with the possibility of a lower cost per channel. It could also lead to fewer satellites and thus fewer earth station antennas in a geographical area. Routing flexibility can result in lower equipment costs through use of common units to serve many links and would also allow transponders to be "filled" more efficiently.

EFFICIENT USE OF SATELLITE CAPACITY

There are two ways to improve the traffic carrying efficiency of a satellite. One is to employ efficient modulation and multiple-access techniques to obtain a larger number of channels per transponder. The second is to assign traffic to channels so that each channel carries more traffic. The SPADE system uses both approaches.

Until recently, the INTELSAT mode of operation has been to transmit multidestination traffic using preassigned frequency-modulated carriers. FM carriers have been standardized at certain sizes (e.g., 24, 60, and 132 channels) to limit the variety of receivers that earth stations must provide. The multiple-access system now in use, known as FDM/FM/FDMA,* provides excellent quality and service but becomes inefficient when two or more carriers must pass through a transponder.

Table 1 shows the capacity of an INTELSAT IV global-beam transponder when accessed by standard earth stations using multiple FM carriers. The multiple-access penalty is attributable to the nonlinear characteristics of the traveling wave tube amplifiers in the satellite transponders [1]. Note that, for a single-carrier access, the capacity of a global-beam transponder (whose usable bandwidth is 36 MHz) is 900 channels, and that the transponder capacity drops to 500 channels or less in typical multicarrier FDM/FM applications. The table shows that the capacity would be only 336 channels using fourteen 24-channel carriers of the type that would probably be used for preassigned multidestination service to light links. This assumes a 100-percent fill factor.† This table also shows that the use of digital modulation (PCM/PSK) and a single channel

* FDM/FM/FDMA: frequency-division multiplex (baseband), frequency modulation (carrier), and frequency-division multiple access (RF band).

† The "fill factor" in this context is the fraction of baseband channels in an RF carrier that are actually in use.

TABLE 1. INTELSAT IV TRANSMISSION CHARACTERISTICS (GLOBAL BEAM)

RF Bandwidth per Accessing Carrier, MHz	Channels per Carrier [a]	Total Accesses per Transponder	Total Channels per Transponder
FDM/FM/FDMA (multichannel carriers)			
2.5	24	14	336
5	60	7	420
10/5 [b]	132	4	456
36	900	1	900
PCM/PSK/FDMA (SPADE)			
0.045	1	800	800

[a] Shows standard carrier capacities employed in 1971. Expanded carrier use will provide somewhat higher capacities in the future.

[b] Three carriers at 10 MHz and one carrier at 5 MHz.

per carrier can result in a transponder capacity as high as 800 channels. This is the technique used in SPADE.

Demand assignment may also serve as a traffic "concentrator." Allocating traffic to available channels and assigning channels to appropriate links permit each channel to carry more traffic than it could carry on a preassigned basis. It is possible to calculate the number of satellite channels which would be required by preassignment and demand-assignment systems to handle various levels of traffic intensity to one or more destinations. For a traffic intensity of 0.5 erlang (30 channel-minutes per hour) to each destination, 40 channels would be required to provide preassigned service to 10 destinations, while only 11 channels would be required if the service were demand assigned. In this case, the ratio, or "traffic concentration factor," would be 3.7.

Traffic concentration factors have been calculated for a range of destinations and for different levels of traffic intensity per destination [2]. It should be noted that the concentration factor is higher for more destinations and lighter traffic intensity per destination. As an example, in typical cases which might be encountered in the Atlantic region of the INTELSAT system in the early 1970's (grade of service $P = 0.01$, 10 to 20 destinations, 0.1 to 0.5 erlang per destination), concentration factors of about four would prevail. On this basis, 800 PCM/PSK demand-assigned channels carried by a single transponder in an INTELSAT IV

satellite in the Atlantic region would be equivalent in traffic intensity to about 3200 preassigned channels.

SPADE DEVELOPMENT AND IMPLEMENTATION

Development of a demand-assignment system started at COMSAT in 1965. After the formation of COMSAT Laboratories in 1967, a demand-assignment system became a formalized project in the INTELSAT R&D program. Its stated goals were as follows:

a. to provide efficient service to light traffic links,

b. to handle overflow traffic from medium-capacity preassigned links,

c. to allow establishment of a communications link from any earth station to any other earth station within the same zone on demand,

d. to utilize satellite capacity efficiently by assigning circuits individually, and

e. to make optimum use of existing earth station equipment.

As system development proceeded, the following important characteristics were incorporated in the system, which came to be known as SPADE:*

a. a single channel on each RF carrier to allow easy establishment of links with any small number of channels,

b. digital coding and modulation for satellite bandwidth conservation,

c. frequency-division multiple access for efficiency in link connections,

d. carrier on-off for satellite power conservation, and

e. autonomous station operation to provide freedom from central control for reliability and earth station independence.

Circuits were developed, breadboarded, and tested, and experimental units were integrated at COMSAT Labs. Units were operated back-to-back under laboratory conditions, first at baseband, then at IF. Finally, experimental equipment was successfully tested over satellite links between earth stations at Andover, Maine, and Cayey, Puerto Rico, and between Goonhilly, U.K., and Etam, West Virginia, in 1968 and early 1969. In April 1970, INTELSAT made a decision to employ the SPADE demand-assignment system for operational use in the Atlantic region.

Although the SPADE system was developed by INTELSAT, the procurement, installation, and operation of SPADE equipment at any earth station is the responsibility of the owner of that station. The United States was the first nation to place an order for SPADE equipment. This was done on the basis of a request for proposal issued by COMSAT on behalf of the U.S. Earth Station Owners' Consortium on an international bid basis in November 1969. Performance specifications for the equipment were those developed by INTELSAT as a standard for all earth stations for regular use in the INTELSAT system.

Nippon Electric Company of Japan was awarded a $313,000 contract for fully redundant equipment with 24 channel units to be used at the Etam, West Virginia, earth station. The equipment, which was installed at Etam and tested in late 1971, will enter operational service in 1972. Figure 1 is a photograph of the NEC equipment for the Etam station.

At the end of 1971, over 35 countries in the Atlantic region had indicated their intention to procure SPADE equipment for use over more than 400 links.

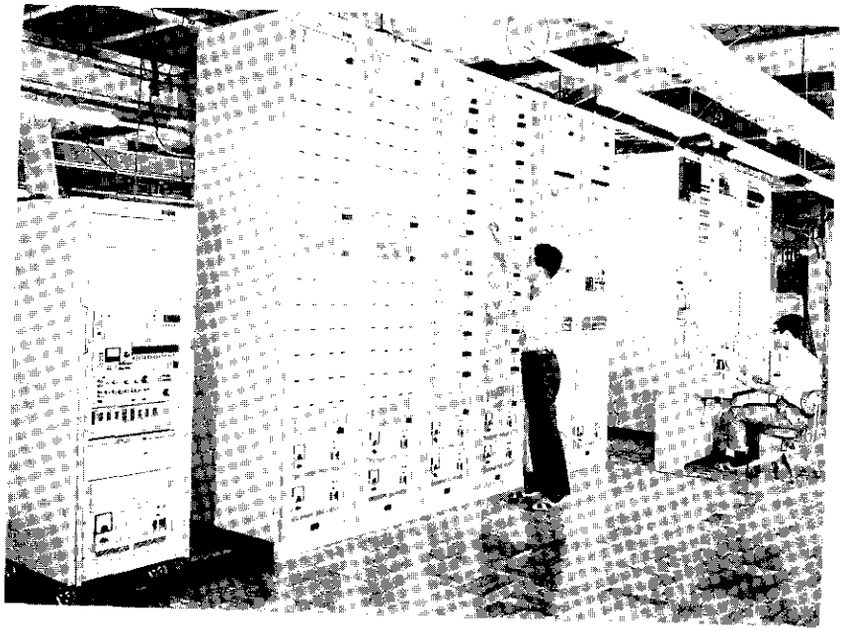


Figure 1. NEC SPADE Equipment at Etam

* SPADE was derived from *s*ingle-channel-per-carrier, *P*CM multiple-*a*ccess *d*emand-assignment *e*quipment.

SYSTEM CONFIGURATION

Figure 2 is a block diagram of a typical SPADE installation developed for use in the INTELSAT system [3], [4]. An installation consists of a terrestrial interface unit, a number of transmit/receive channel units, a demand-assignment signaling and switching (DASS) unit, and an intermediate frequency (IF) subsystem. The system operates with a single voice channel assigned to each radio frequency carrier and each carrier assigned to a separate frequency. Demand assignment of channels is provided without central control. Pulse-code modulation (PCM) is used for channel encoding and four-phase coherent phase-shift keying (PSK) is used for modulating each carrier.

System coordination is maintained through a high-quality common signaling channel (CSC) employing time-division multiple access (TDMA). The characteristics of the communications and signaling channels are shown in Table 2. Note that the CSC is much more highly protected than the communications channels. It operates at a data rate of 128 kbps with coding, uses two-phase PSK, and attains an extremely low error rate of 10^{-7}.

The SPADE terrestrial interface unit provides the interface for call

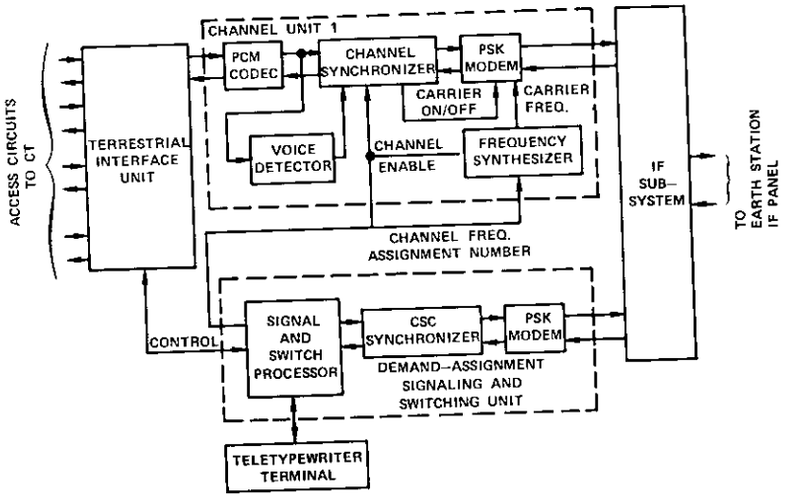


Figure 2. SPADE Terminal Functional Block Diagram

signaling to initiate, supervise, and terminate all calls systematically. When a call request is received from a local transit center (CT), the DASS unit automatically selects a frequency pair from the pool of available channels and alerts the destination station to the incoming call and frequency assignment. All DASS units utilize the signaling information disseminated by the CSC to update a channel table so that the frequencies just assigned are unavailable for new calls.

The frequency selected for the outgoing call is provided to the channel unit by a frequency synthesizer that is capable of generating any of the 800 discrete frequencies using digital codes provided by DASS. This frequency is used for both the outgoing carrier and the received-signal local oscillator. Channel pairings are based on the common use of the synthesizer for received and transmitted signals.

Upon activation of the modem, the DASS unit conducts a two-way circuit continuity check. Once the call has been established, the analog signal received by the channel unit is sent to a PCM codec, which transforms it to a digital signal for outgoing transmission and transforms returning signals from digital to analog form.

The content of the voice channel coming from the CT is detected by a voice detector, which is used to gate the channel carrier on or off. This conserves satellite power as a function of talker activity. The digital bit

TABLE 2. SOME TECHNICAL CHARACTERISTICS OF THE SPADE SYSTEM

a. Communications Channel Characteristics

Channel Encoding	PCM
Modulation	4-Phase PSK (coherent)
Bit Rate	64 kbps
Bandwidth per Channel	38 kHz
Channel Spacing	45 kHz
Stability Requirement	±2 kHz (with AFC)
Bit-Error Rate at Threshold	10^{-4}

b. Common Signaling Channel Characteristics

Access Type	TDMA
Bit Rate	128 kbps
Modulation	2-Phase PSK
Frame Length	50 ms
Burst Length	1 ms
Number of Accesses	50*
Bit-Error Rate at Threshold	10^{-7}

* 49 stations plus 1 reference.

streams in and out of the PCM codec are synchronized by the transmit-receive synchronizer, where timing, buffering, and framing functions are performed. The PSK modem modulates the assigned carrier frequency with the outgoing bit stream and coherently demodulates the incoming bursts by recovering carrier and bit timing associated with the received signals. The modulated carriers, both outgoing and incoming, are then passed through a common IF subsystem, which interfaces with the earth station up- and down-converters at IF. The carrier used for the CSC modem is also passed through the IF subsystem.

When the call is completed, a control signal from the CT allows the DASS unit to return that circuit to the frequency pool for reassignment. This information is passed to all stations via the CSC.

The design of the SPADE system has taken into account the variety of signaling and switching systems that will interface with each SPADE terminal. The terrestrial interface and DASS units in the SPADE terminal will be designed to accommodate the CCITT signaling and switching systems in local use. Communications between countries having different signaling and switching systems, such as CT-1, CT-5, CT-5 bis, R-2, manual, or even the new CT-6, can then be established using the SPADE DASS unit to provide compatibility.

SYSTEM OPERATION

The flow of signaling and communications information through the SPADE system may be seen in Figures 3 and 4, respectively. The dots indicate a terrestrial signal from a subscriber arriving via a CT at the DASS unit of earth station A. The DASS unit then routes the signaling information via the CSC, operating in a TDMA broadcast mode, to all participating earth stations. At the initiation of this routing, station A requests allocation of a particular frequency for this call from a busy-idle table of frequencies continually updated via the CSC. Station C, for which the call is destined, monitors the common signaling channel and notes the arrival of station A's request and the allocated frequency. If no other station has requested that frequency before station A, station C transmits an acknowledgment of the call request, assigns the frequency to the channel equipment, and proceeds to verify continuity and establish ultimate connection to the subscriber via the CT.

Station A monitors the common signaling channel between the time when it requests a frequency and the time when it receives its own request back (about 240 ms). If, during that time, another station had requested

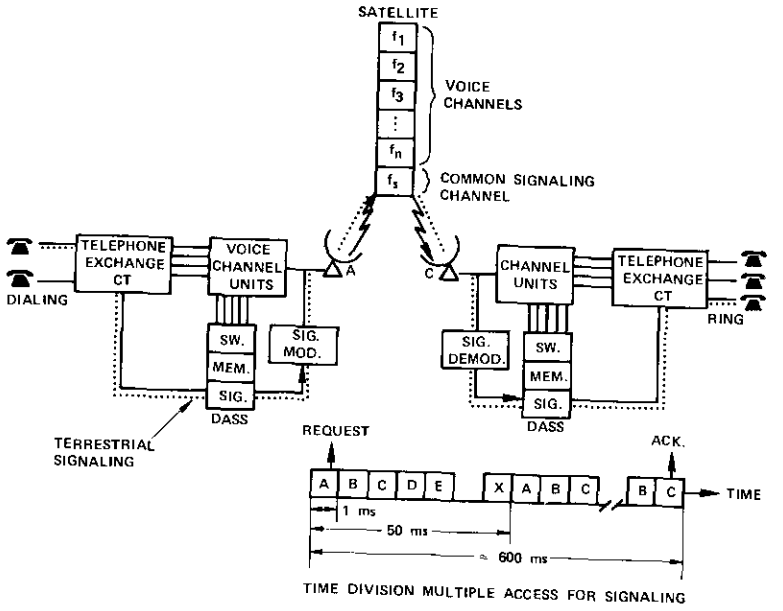


Figure 3. Signaling Flow in a SPADE Demand-Assignment System

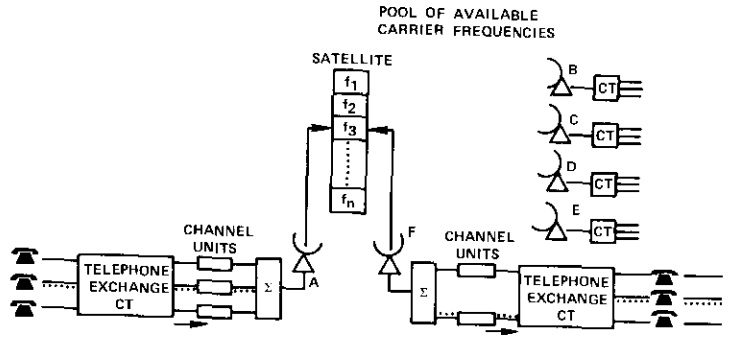


Figure 4. Voice Signal Flow in a SPADE Demand-Assignment System

the same frequency, both stations A and C would have registered a busy to that allocation on their busy-idle tables and station A would have initiated a new request. To minimize conflicting requests for simultaneous frequency assignment, each station chooses randomly from a list of available frequencies stored in the DASS unit.

Note on the time scale of Figure 3 that access to the CSC occurs once every 50 ms for each station and that the access duration is 1 ms. Station A may therefore initiate other calls while awaiting its own and station C's response. At the end of the call, the disconnect signal is broadcast over the CSC and all stations note that this frequency is again available for assignment.

The communications signal (Figure 4) is received by a channel unit at the SPADE installation at earth station A. Upon receipt of the appropriate signal from the DASS unit, as just described, the modem of that channel unit is assigned to a particular carrier frequency pair (f_3 in this case). The modem at destination station F is assigned the corresponding transmit/receive pair of frequencies. For the duration of that connection, f_3 remains assigned to that circuit and is "busied out" to other requesters. At the end of the communication, f_3 is returned to the network pool for subsequent reallocation. At earth station F, the call arriving on frequency f_3 is demodulated in the particular unit to which circuit f_3 has been assigned and is forwarded, via the CT, to the subscriber over the terrestrial link.

It might be noted that, at both the origin and destination stations (A and F in this case), the terrestrial connections between CT and earth station are assigned arbitrarily, since the equipment assignment is independent of either call source or destination. This extends the advantage of demand assignment to the CT's and allows more efficient use of the CT-to-earth-station capacity.

The SPADE system has been assigned to transponder number 10 of an Atlantic INTELSAT IV (F-2) satellite. This transponder, which has a nominal RF bandwidth of 36 MHz, lying between 6.302 and 6.338 GHz for reception and 4.077 and 4.113 GHz for transmission, is connected to a global-coverage antenna.

Figure 5 shows the distribution of SPADE carriers over the 36-MHz band of transponder 10 and the corresponding band centered at the 70-MHz IF interface between the SPADE terminal and the earth station up- and down-frequency converters. At the lower end of the band, space is allocated for the common signaling channel. A system pilot is located at band center.

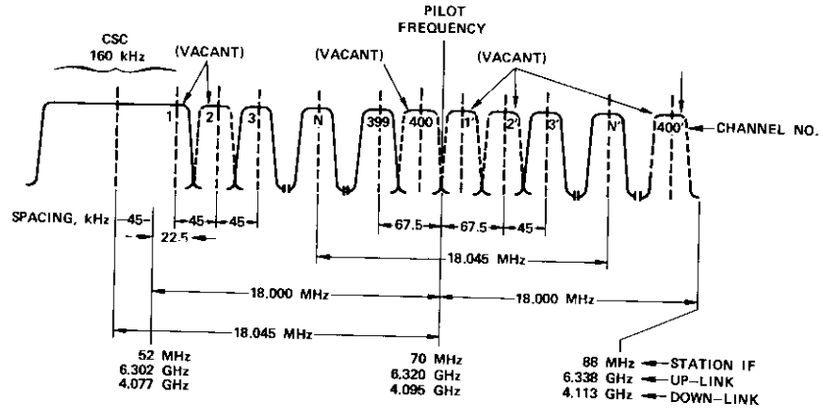


Figure 5. SPADE Multichannel Frequency Allocation Spectrum

SPADE DATA RECORDING SYSTEM

A special system for recording traffic data from the operational SPADE network in the Atlantic region has been developed for INTELSAT by COMSAT Laboratories. This system will make it possible to record and analyze event-time data on each call offered to and carried by the SPADE system. This information will provide a data source for system performance and growth measurement, development of demand-assigned call statistics, and determination of grade of service. The system can also generate data for billing if required.

The first data recording terminal is being located at the Etam earth station. The recorded data will be periodically transmitted to and processed by a general-purpose (IBM 360/65) digital computer at COMSAT. Only one unit is presently being installed, but another may be added at another location, particularly if a per-minute billing charge now under study is instituted by INTELSAT.

SPADE SYSTEM PERFORMANCE

As was noted previously (Table 1), an attractive feature of the SPADE mode of operation is the high achievable channel capacity per transponder relative to present FDM/FM/FDMA operation. It has been shown [5] that, when voice activation is used on each carrier, there is sufficient

satellite power to support more than 800 SPADE-type channels in one INTELSAT IV transponder using standard earth stations. Since each channel occupies 45 kHz of RF bandwidth, it is evident that the transponder is bandwidth limited to 800 channels (actually 794 channels after housekeeping channels are excluded).

It has been estimated that the voice activation feature provides a 4-dB satellite power saving. Of course, as in the SPADE system, voice-activated carriers can be used only when a single channel is assigned to each carrier. A similar advantage could be realized by a system using single-channel FM carriers, but PSK modulation is more efficient in terms of bandwidth for a single channel, thus providing more channels per transponder. Calculations show that single-channel-per-carrier FM/FDMA with voice activation would provide a maximum of 450 channels per transponder.

The capacity of the transponder for SPADE-type channels would be reduced if voice activation were not used –for example, if all channels were carrying continuous data streams. On an absolute power basis, 350 simultaneous channels, continuously activated, could be supported in an INTELSAT IV global-beam transponder with an optimized operating point. In such a case, however, half of the transponder bandwidth would be unused. Thus, if such an application existed, it would be practical to use the excess bandwidth for forward-acting error coding on the 350 continuous channels. With a rate three-quarter convolutional code, for example, each channel would occupy a bandwidth of 60 kHz, allowing 600 channels in a bandwidth-limited case [6]. Actually, the system would still be power limited and the capacity which could realistically be attained with this method would be only 500 channels.

The 500-channel-per-transponder limitation would apply to preassigned channels without voice activation. The demand-assignment concentration factor (as shown in Figure 1 of Reference [2]), however, is independent of voice activation and would apply whenever demand assignment is used.

Table 3 summarizes the capacity in channels of the INTELSAT IV global-beam transponder using standard earth stations and various modulation and access techniques with and without voice activation and coding.

SPACE SYSTEM APPLICATIONS

Demand-Assignment Operation with Standard Stations

Earth stations accessing the Atlantic region INTELSAT IV satellite can participate in the SPADE demand-assignment service through transponder 10. Demand-assigned service to the INTELSAT network of standard stations, which was the motivation for development of the SPADE system, will be its first and most important application.

As has been shown, this service will find ready application to lightly loaded links in the Atlantic region, serving in many cases where the number of channels per link is 12 or less. Although no commitment has been made, this same service can be expected to extend later to the Indian Ocean region.

Introduction of demand-assigned service is expected to encourage new traffic by simply increasing traffic, channel by channel, if and when re-

TABLE 3. SUMMARY OF THE CHANNEL CAPACITY OF THE INTELSAT IV GLOBAL-BEAM TRANSPONDER USING STANDARD EARTH STATIONS

Modulation Technique	Access Mode	Transponder Capacity (channels)
FDM/FM	Single Access/Single Carrier	900
PCM/PSK	Single Access/Single Carrier (64 Mbps)	1000
FDM/FM/FDMA	Multiple Access (average, mix of various size carriers)	450
FDM/FM/FDMA	Multiple Access, 14 24-Channel Carriers	336
PCM/PSK/FDMA (voice)	Single Channel/Carrier (64 kbps)	With Voice-Activated Carrier: 800 Without Voice-Activated Carrier: 350 Without Voice-Activated Carrier, with 3/4 Coding: ~550
PCM/PSK/FDMA (1×10^{-6} data)	Single Channel/Carrier (64 kbps)	Without Coding: 250 With 3/4 Coding: 450
PCM/PSK/FDMA (1×10^{-6} data)	24 Multiplexed Data Channels/Carrier (total rate equal to 1.544 Mbps)	With Coding: 450

quired on existing links. By the same token, new single-channel links can be readily established between participating stations even if no traffic has passed between them previously.

Also, and this is more speculative, the availability of SPADE in the INTELSAT network may provide an incentive for new earth stations to enter the system. There may well be cases in which a country's traffic requirements to one or two destinations may not have provided sufficient economic justification for an earth station, but in which the total traffic foreseen to all SPADE members may tip the scales in favor of entering the system.

Use of the SPADE demand-assignment system is foreseen throughout the INTELSAT IV era. In fact, INTELSAT V system studies, extending into the 1980's, indicate that SPADE will be used as the common global all-user telephony system.

Preassigned Single-Channel-Per-Carrier Operation

Normal use of the SPADE demand-assignment system will provide a means for orderly traffic growth on existing links and easy introduction of new links. As traffic builds up on demand-assignment routes, however, users may wish to establish preassigned links to certain destinations. For this purpose, a SPADE participant may elect to establish a single full-time PCM/PSK channel. He could then use single-channel-per-carrier (SCPC) equipment which is identical to SPADE hardware except that its frequency allocation is fixed. These fixed SCPC links will use transponder 10 until demand-assigned links require the entire capacity. At that time, the fixed routes will be transferred to another transponder.

The use of preassigned SCPC links will mean that capacity can be assigned in the minimum unit size (i.e., one voice channel) without wasting transponder capacity by applying multichannel standard size blocks. It is expected that additional requirements for SCPC will appear soon after the introduction of SPADE for use in transmitting such services as telegraphy bearer links, alternate voice/data, and point-to-point dedicated subscriber channels between stations not connected by preassigned FM carriers.

Operation with Small Earth Stations

The use of a single voice channel on each radio frequency carrier makes it possible to set the transmitted power of each carrier at a level appropriate to the size of the earth station for which it is destined. In a network of standard earth stations,* the power of each carrier will be the same. If one or more stations which are smaller than the present standard size are introduced into the network, however, the power levels to carriers destined for them will have to be increased to provide the same signal quality. Since only those carriers destined for smaller stations will require additional power, the number of channels per transponder will be higher than if power for all links had to be increased.

If all stations using a transponder are the same size, both power and bandwidth per carrier can be selected for efficient utilization of the transponder. For example, if all earth stations in a SPADE network using a global-beam transponder and uncoded transmissions had a G/T of 35 dB/°K, the capacity would be reduced to 400 channels and the system would be power limited. If rate three-quarter coding were then introduced to take advantage of the excess bandwidth available, the capacity would be increased to 600 channels.

If even smaller earth stations were employed, e.g., earth stations with a G/T of 29 dB/°K (equivalent to a 10-m antenna with an uncooled receiver), the transponder capacity in a SPADE network using uncoded transmissions would be only about 150 channels; with rate three-quarter coding, 250 channels; and with rate one-half coding, 350 channels. This is shown in Figure 6. (See also Reference [7].)

If a small station regional SPADE network were to operate through an INTELSAT IV transponder connected to a spot-beam antenna, it could immediately trade the 12 dB of additional satellite antenna gain for a corresponding reduction in the size of the earth station antennas. The transponder capacity for earth stations with a G/T of 29 dB/°K would still be 800, as in the standard SPADE system. If it were desired to operate a regional network with much smaller earth station sizes (of the order of 5 m in diameter and G/T = 21 dB/°K), the transponder could provide approximately 250 channels without coding. Rate three-quarter coding could increase this capacity to approximately 450 channels. Spot-beam coding advantages are shown in Figure 7.

* Standard earth stations are those that comply with the INTELSAT standard requirements in ICSC–45–13 (nominally 30-m diameter antenna and G/T = 40.7 dB/°K).

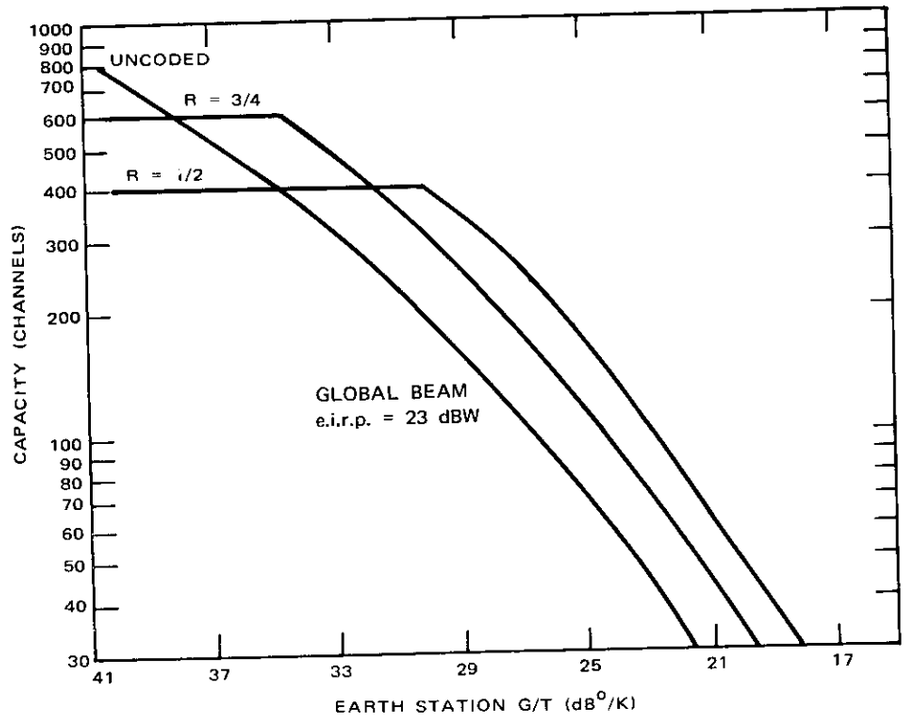


Figure 6. INTELSAT IV Transponder Capacity (global beam)

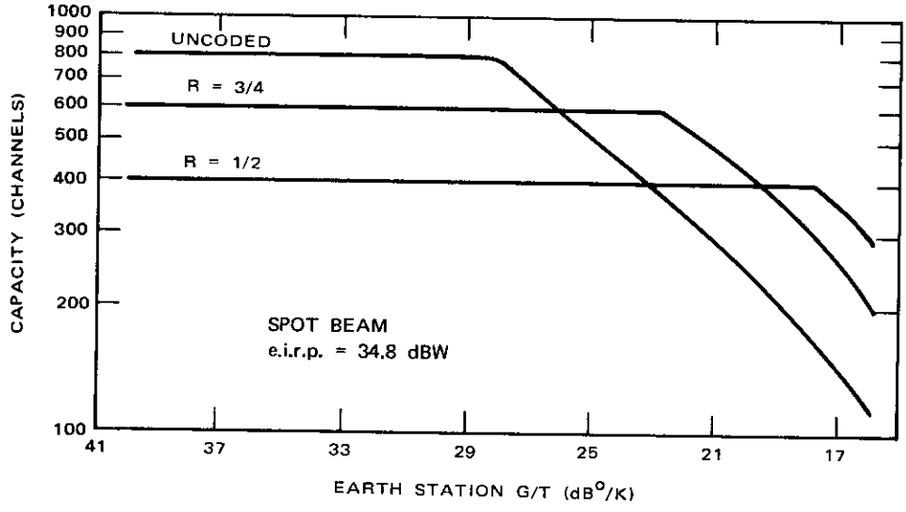


Figure 7. INTELSAT IV Transponder Capacity (spot beam)

Telegraphy

In the initial months of operation of the SPADE network, the present practice of accommodating telegraphy traffic on voice bearer channels will be continued. Thirty telegraphy channels are presently frequency-division multiplexed in the voice bandwidth and then combined in a multichannel FM carrier for transmission over the satellite. A single SPADE channel can also carry 30 telegraphy channels addressed to multiple destinations or to a single destination. Obviously, however, there are more efficient methods of accommodating telegraphy than the voice bearer method. The basis of these more efficient methods is the digital signal. To this end, under an INTELSAT R&D contract with the S.I.T. Siemens Company of Italy, equipment is being developed that will efficiently multiplex telegraphy signals for access to a SPADE-type channel unit or common signaling channel.

The equipment to be delivered under this contract consists of two basic units. The first, designed to accommodate larger users, multiplexes all of an originating station's telegraphy signals into a single 48-kbps bit stream which accesses a SPADE channel unit after appropriate stuffing or coding to a 56-kbps rate. The transmission of that bit stream over the satellite is then received at one or several stations, the bit stream is demultiplexed at these destinations, and the individual telegraphy channels are reconstituted into their normal format. It is expected that such a channel could carry the equivalent of 1100 fifty-baud telegraphy channels and that the equipment could accommodate standard telex signaling as well as point-to-point telegraphy links.

The second type of telegraphy equipment being developed as part of this program is tailored to the user with many destinations but with light traffic to each. The technology to be used for this application is modeled after the SPADE common signaling channel. Access to the satellite is on a time-division basis. Each user is allocated a burst time during which he transmits to the satellite all of his traffic destined for various correspondents. After the satellite transmits this information, each participating station in this telegraphy network receives the burst transmissions, extracts from each burst the specific channels that have been assigned to it, and reconstitutes the original telegraphy signal from the received bit stream. It is expected that, in this mode, the channel bit rate will be on the order of 128 kbps, and that the total capacity for this telegraphy network will be in excess of 2500 total telegraphy or telex channels operating at a nominal 50-baud rate or multiples thereof. Thus such a network could

accommodate 60 earth stations, each having a capacity of 40 multi-destination telegraphy channels. These numbers indicate that digital technology will introduce significant gains in efficiency over the methods presently being used. Figure 8 portrays the two telegraphy data systems just described.

Data Transmission

The SPADE channel operates at a fundamental bit rate of 56 kbps to accommodate a digitized voice signal. It is evident that this same channel can be used for transmission of data up to the same 56 kbps. Again the judicious use of forward-acting error coding can provide data transmis-

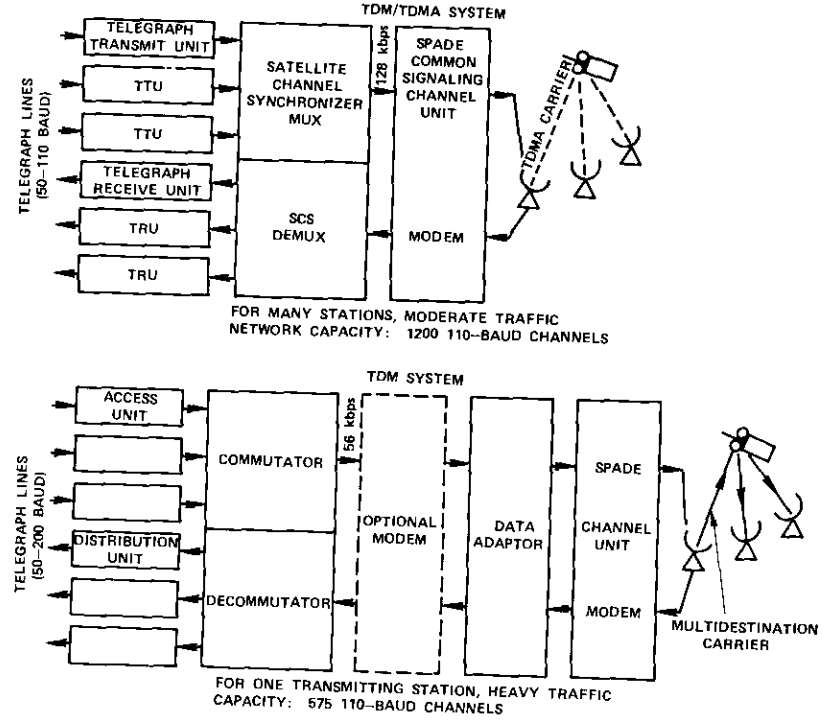


Figure 8. Telegraphy Systems for Use in the SPADE Network

sion at very low bit-error rates and with an efficiency not previously possible using frequency-modulation techniques.

Today, a standard analog voice channel is capable of carrying data at rates up to 4800 bps, and with equalization up to 9600 bps. The SPADE voice channel, capable of carrying data at rates as high as 56 kbps, can handle a single wideband data stream (e.g., 40.8, 48, or 50 kbps) or a multiplex of lower bit rate data transmissions. Techniques for multiplexing data into a single bit stream capable of accessing the SPADE channel are under study [6]. In addition, low-cost equipment is being developed to optimize the "through-put" of data links operating over the time delay of satellite links.

Equipment similar to that being used for the SPADE channel unit (in particular, the PSK modem) is being developed to allow efficient transmission of digital signals at higher rates than are presently possible over the satellite. The transmission performance of 24 PCM voice channels multiplexed to 1.544 Mbps has been successfully tested over INTELSAT IV. Further tests will be conducted at bit rates of 6.3 Mbps. Using phase-shift-keyed techniques, such transmission will require no more power and bandwidth than an equivalent number of nondigitized voice channels.

Multidestination Distribution Service

In a SPADE network it is easy to visualize several types of one-way service destined for many users and carrying information of wide interest, such as weather, financial data, international meeting proceedings, and world news coverage. A set of SPADE channels might also be set aside for governmental or commercial use, such as airline or hotel reservation service, international corporate accounting, inventory control, sales directives, and government service to consulates or embassies. There is a large number of such possibilities. Each of these applications could be established through the SPADE system, either on a demand-assignment or preassignment basis through proper programming.

CONCLUSIONS

The SPADE demand-assignment system which has been described in this paper is in the process of entering operational service in the INTELSAT network. It is expected to provide reliable, flexible, and economic service, particularly over lightly loaded links, although preassigned FM service will continue to be used for moderate and heavily loaded links. The SPADE system and the equipment and techniques developed for it will find several other important applications in the future: preassigned PCM/PSK channels, telegraphy and data transmission over a wide range of data rates, small and mixed sized earth station operation, and multidestination links.

REFERENCES

[1] A. L. Berman and C. E. Mahle, "Nonlinear Phase Shift in Traveling-Wave Tubes as Applied to Multiple-Access Communications Satellites," *IEEE Transactions on Communications Technology*, COM-18, No. 1, February 1970, pp. 37–48.

[2] G. D. Dill, "Comparison of Circuit Call Capacity of Demand-Assignment and Preassignment Operation," *COMSAT Technical Review*, Vol. 2, No. 1, Spring 1972 (this issue).

[3] J. G. Puente and A. M. Werth, "Demand-Assigned Service for the INTELSAT Global Network," IEEE *Spectrum*, Vol. 8, No. 1, January 1971, pp. 59–69.

[4] J. G. Puente, W. G. Schmidt, and A. M. Werth, "Multiple-Access Techniques for Commercial Satellites," *Proceedings of the IEEE* (U.S.A.), Vol. 59, No. 2, February 1971, pp. 218–229.

[5] R. B. McClure, "The Effect of Earth Station and Satellite Parameters on the SPADE System," *IEE Conference on Earth Station Technology*, London, England, October 14–16, 1970, IEE Conference Publication 72, pp. 17–23.

[6] E. R. Cacciamani, "A Channel Unit for Digital Communications in the SPADE System," *IEEE 1971 International Conference on Communications*, Montreal, Canada, June 14–16, 1971 [Proceedings], pp. 42–15 to 42–19.

[7] E. R. Cacciamani, "The SPADE System as Applied to Data Communications and Small Earth Station Operation," *COMSAT Technical Review*, Vol. 1, No. 1, Fall 1971, pp. 171–182.

ACKNOWLEDGMENTS

The authors appreciate the assistance of Messrs. E. Cacciamani, G. D. Dill, A. Mark, R. McClure, J. Puente, and P. Troutman of COMSAT, all of whom contributed technical information for this paper.

Dr. Edelson is Assistant Director of COMSAT Laboratories. Before joining COMSAT, he served in the U. S. Navy, where he was project manager for Navy space communications, navigation, and astronautics ship programs; was on the staff of the National Aeronautics and Space Council; and was Technical Officer at the Office of Naval Research in London.

Dr. Edelson is a U. S. Naval Academy graduate and received his M.S. and Ph.D. from Yale University. He is an Associate Fellow of AIAA, a senior member of IEEE, and the recipient of the Henry Marion Howe Research Medal and the U. S. Navy Legion of Merit.

Andrew M. Werth received his Bachelor's and Master's degrees from Columbia University in New York. Before joining COMSAT, he was a project engineer at ITT Federal Laboratories, where he directed several tasks in applied research in satellite communications.

At COMSAT Laboratories he was Manager of the Modulation Techniques Branch. Much of the SPADE demand-assignment system was developed under his direction. He was also responsible for the design and development of digital communications techniques, with emphasis on modems, application of coding techniques, and optimization of channel capacity.

Mr. Werth is presently Vice-President of Digital Communications Corporation, Rockville, Maryland. He is a member of IEEE.

Index: communication satellites, computerized simulation, mathematical models, radio relay systems, systems engineering, telecommunication

Comparison of circuit call capacity of demand-assignment and preassignment operation

GEORGE D. DILL

ABSTRACT

A model of the traffic handling capacity of low-density telecommunication trunks is developed. This model is used to derive the relative improvement to be gained on the terrestrial access circuits and in the satellite transponders by using fully variable destination demand-assigned circuits to service the low-density links. The Atlantic Ocean region traffic for the second quarter of 1971 is used to evaluate the model.

INTRODUCTION

A primary purpose of a demand-assignment communications system* is to provide a direct and economical service for low-density international

* Calculations in this paper are based on the use of a fully variable demand-assignment operational technique. Appendix A describes one example of such a multiple-access technique.

This paper is based upon work performed at COMSAT Laboratories under the sponsorship of the International Telecommunications Satellite Consortium (INTELSAT). Any views expressed in this paper are not necessarily those of INTELSAT.

telephony trunks.* One economic improvement provided by such a system is that traffic to all destinations can be serviced by the same equipment. No additional transmit or receive chains or other earth station equipment is required to service calls to or from up to 48 other earth stations operating in the system. A second and equally significant economic improvement brought about by a demand-assignment communication system is that calls can be carried more efficiently, both within the terrestrial facilities and through the satellite. The purpose of this paper is to quantify the operational improvement to be gained by individual administrations using demand-assignment communication terminals to service their low-density traffic trunks, and by INTELSAT, which can serve more traffic through its satellites. An example of such a demand-assignment communication system is the SPADE system described in References [1] and [2].

Large trunk groups have traditionally been used for hauling telephony traffic on a point-to-point, preassigned circuit basis. This mode has been found to be very economical for high-density trunks. However, because the efficiency of a preassigned circuit is proportional to the trunk size (total number of circuits per trunk), preassigned circuits are less economical for low-density trunks.

A new technique is used for fully variable demand-assigned telephony circuits; namely, each terrestrial access circuit appearance at one terminal can be routed to any other terminal's terrestrial access circuit appearance on a per-call basis. Thus, each call may use any free terrestrial access circuits, and satellite circuits are used only on demand. As soon as the call is completed, the circuits are returned to the idle state to await selection for subsequent calls.

The advantages of demand-assignment operation are:

a. the terrestrial access circuits between the earth terminal and the international exchange operate as a single trunk, and

b. the satellite circuits also operate as a single trunk.

As a result, the operational efficiency (or call capacity) of each circuit may be greater in both the terrestrial and satellite links.

* A trunk is defined as an individual assembly of one or more telephony circuits required to carry traffic between two points. A low-density trunk is a trunk requiring 10 or less preassigned circuits to support the offered traffic at the specified grade of service.

TRAFFIC ENGINEERING

To properly compare demand-assigned service with preassigned service for low-density traffic trunks, it is first necessary to establish the assumptions under which the comparison is to be made and to define the models of each service to be compared.

Mode of Operation

Today, many of the low-density preassigned trunks are operated on a manual ringdown basis. It is believed that this condition exists primarily because of the difficulty of providing multidestination interfacing, which is required for preassigned service. However, because the CCITT recommends that all administrations should install fully automatic or semi-automatic service in the future, this analysis will assume that such a service is also available for the low-density preassigned trunks [3]. It should be noted that both manual and automatic international exchanges can be interfaced to a SPADE terminal [4], [5].

Traffic Characteristics

To compare demand-assignment and preassignment operation, two assumptions regarding traffic characteristics are necessary:

a. a Poisson distribution of call arrivals at the earth station, and

b. an exponential distribution of call holding times.

Grade of Service

The performance of a trunk is generally measured in terms of the average ratio of calls blocked to the total calls offered to the trunk per unit time (normally during the busy hour). This ratio is designated as the grade of service and may be theoretically derived by using the erlang B formula:

$$P_B = \frac{a^n/n!}{1 + a + a^2/2! + \ldots + a^n/n!} \qquad (1)$$

where a = average traffic load per unit time (erlangs)

n = total number of two-way circuits available in the trunk

P_B = probability of a call encountering all circuits busy per unit time.

However, it is generally more convenient to use an erlang table [6], [7].

Average Traffic Load

Normally, the average telephony traffic load per destination is already known or is predicted from a traffic data base. Given an average traffic load to be offered by each trunk, the total load offered to an earth station (for either preassigned or demand-assigned service) may be found by using the following equation:

$$A_t = \sum_{j=1}^{t} (a_j) \tag{2}$$

where a = average two-way traffic load per unit time (erlangs)
t = total number of individual trunks (destinations) served by the kth earth station, where k is the total number of earth stations per community.

The total traffic load offered to the satellite during the busy hour may be found from the following expression:

$$A_s = \frac{1}{2}\left[\sum_{i=1}^{k} \sum_{j=1}^{t} (a_j) \right] \tag{3}$$

Terrestrial Access Circuit Requirements

Given a specified grade of service and traffic load per trunk or destination, one can use a standard erlang table [6], [7] to determine the minimum number of preassigned circuits, n_{PA}, required between each pair of correspondents. The sum of these requirements, i.e., the total number of preassigned terrestrial access circuits per earth station, may be determined from the expression

$$n_{PAT} = \sum_{j=1}^{t} (n_{PA})_j \tag{4}$$

where n_{PA} = total number of two-way preassigned terrestrial access circuits required per trunk (or destination)
n_{PAT} = total number of two-way preassigned terrestrial access circuits for the kth earth station.

Because the demand-assigned circuits are fully variable, each call may use any idle access circuit to reach its destination. Thus the access circuits of each earth terminal may operate as a single trunk, and the total number of demand-assigned terrestrial access circuits per earth terminal may be found from either an erlang table or from equation (1).

Satellite Circuit Requirements

The total number of two-way preassigned satellite circuits required to carry a specified load of traffic at a particular grade of service may be found as follows:

$$n_{PAS} = \frac{1}{2}\left[\sum_{i=1}^{k} \sum_{j=1}^{t} (n_{PA})_j \right]. \tag{5}$$

Because the demand-assigned service is fully variable, the INTELSAT IV transponder may operate as a single trunk for all calls offered. Therefore, the total number of satellite circuits required for this service may be found by using an erlang table or equation (1) to find the appropriate number of satellite circuits required to provide the desired grade of service to support a total traffic load equal to one-half of the sum of the individual traffic loads offered by each earth terminal [see equation (2)].

IMPROVEMENT FACTOR

It is clear that the use of demand-assigned service to support low-density trunks requires fewer terrestrial access circuits and satellite circuits in the network. It is easiest to first compute this improvement factor in the general parametric sense and to then evaluate it for each individual case.

Table 1 is a parametric tabulation of the total number of two-way telephony circuits for preassigned service (n_{PA}) and demand-assigned service (n_{DA}) as a function of offered traffic (a) per trunk (or destination), number of trunks per offered traffic density (t), and grade of service to be provided by each trunk. It may be noted that the number of preassigned circuits is sensitive to both the number of individual trunks or destinations to be supported and the offered traffic per trunk. On the other hand, the number of demand-assigned circuits is sensitive primarily to offered traffic per trunk. The improvement factor can then be defined as

$$I = \frac{n_{PA}}{n_{DA}} \tag{6}$$

where I = improvement factor
n_{PA} = number of preassigned two-way circuits
n_{DA} = number of demand-assigned two-way circuits.

TABLE 1. TOTAL NUMBER OF TWO-WAY CIRCUITS PREASSIGNED VS. DEMAND-ASSIGNED

OFFERED TRAFFIC PER DESTINATION		$a = 0.1$		$a = 0.5$		$a = 1.0$	
Number of Preassigned Trunks Per Offered Traffic Density	t	n_{PA}	n_{DA}	n_{PA}	n_{DA}	n_{PA}	n_{DA}
$P_B = 0.01$	4	8	3	16	7	20	10
	8	16	4	32	10	40	15
	10	20	5	40	11	50	18
	20	40	7	80	18	100	30
	40	80	10	160	30	200	53
	80	160	15	320	53	400	96
	100	200	18	400	64	500	117
	200	400	30	800	117	1000	221
	400	800	53	1600	221	2000	426

a = offered traffic per destination in busy hour (erlangs)
n_{PA} = total number of preassigned access circuits per earth station
n_{DA} = total number of demand-assigned access circuits per SPADE terminal
t = total number of preassigned trunks

Figure 1 is a parametric plot of the improvement factor for a grade of service, P_B, equal to 0.01. Empirically, the average number of trunks terminating at each earth station will lie between 2 and 20 (see Figure A-1 in Appendix A). It may also be seen that there are approximately 30 to 60 earth stations per communications satellite region. Thus, the total number of trunks handled by the satellite transponder will lie between 30 and 600. Consequently, for the cases shown in Figure 1, a terrestrial access circuit improvement factor of 2 to 6 may be experienced by individual earth station administrations using demand-assignment operation. That is, demand-assigned service requires from one-half to one-sixth as many individual access circuits as preassigned service. This reduction factor also applies to termination equipment, transmission facilities, initial installation costs, and continuing operation and maintenance costs, although it may vary for different traffic loads per trunk or grades of service.

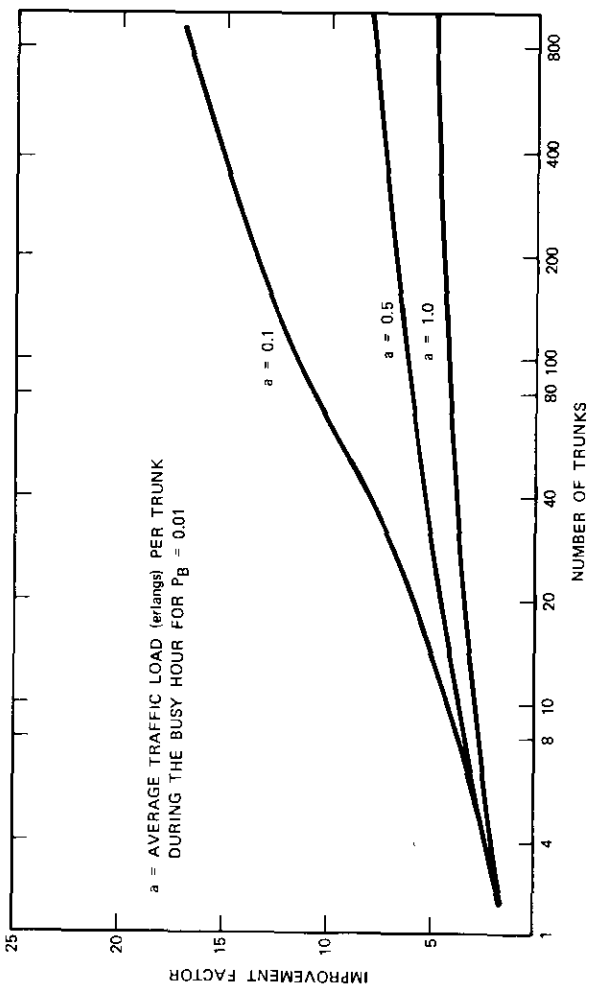


Figure 1. Parametric Plot of the Improvement Factor for Demand-Assignment Operation

Improvement factors from 5 to 15 may be achieved on the satellite by using fully variable demand-assignment operation to support low-density trunks. This greater improvement factor is attributable to the greater number of trunks to be serviced (if it is assumed that not all trunks are supported by one earth station). Thus, if one INTELSAT IV transponder is assigned to this service and if all earth stations in the community have a demand-assignment terminal, the equivalent of 2000 low-density preassigned circuits (if there are that many in the network) can then be transferred to that transponder. This would be a most significant factor in postponing the date when the Atlantic Ocean region INTELSAT IV satellite becomes fully loaded.

ADDITIONAL CONSIDERATIONS

The improvement factors noted previously are only theoretical. Some additional improvement may be gained in the actual networks. The following considerations may contribute to an even greater circuit improvement factor and may also have an impact on the future growth of satellite telecommunications:

a. Some trunks are now being served via transit exchanges or not served at all because of the relatively high cost of implementing a new preassigned down-link chain at each destination for the low-density trunks.

b. The previously noted improvement factors assume that the low-density trunks may occupy free channels in existing carriers. This is not always the case, and, in fact, these trunks may require the use of a larger carrier. In this case, the fill factor will be much less than 100 percent. Because of the day-to-day variation in the fill factor, the unused channels in such a carrier were not included in the total preassigned circuit requirement for comparison with demand-assigned circuit requirements.

c. The relatively high cost of implementing new low-density preassigned trunks is a deterrent to the inauguration of new trunks via preassigned FDM/FM carriers. On the other hand, demand-assignment terminals make it possible for each administration to conduct new traffic in any quantity to any of the other administrations which have demand-assignment terminals. Hence, demand-assignment operation provides an incentive for the establishment of new trunks, since the additional traffic will contribute to a lower cost per unit time. Because telecommunications via satellite is relatively new, this factor will also have a significant effect on future INTELSAT satellite requirements.

CONCLUSIONS

It has been found that a relative improvement factor of 2 to 6 may be obtained by a demand-assignment user; i.e., demand-assigned service requires only one-half to one-sixth as many terrestrial facilities as preassigned service to handle the same amount of low-density traffic. Similarly, demand-assigned service provides a satellite improvement factor of 5 to 15; i.e., it requires one-fifth to one-fifteenth as many satellite circuits as preassigned service to handle this low-density traffic. In terms of actual circuits, one INTELSAT IV transponder can handle the equivalent of 2000 to 6000 low-density preassigned circuits, depending on the average traffic load per trunk and grade of service to be provided.

An example of the improvement offered via fully variable demand-assigned circuits is evaluated by using the mid-1971 preassigned network in the Atlantic Ocean region (see Appendix A). If a 30-percent growth rate is assumed for the light traffic routes in the Atlantic Ocean region, a single INTELSAT IV transponder is anticipated to have sufficient capacity to handle all demand-assigned circuits at least up to 1978 and possibly up to 1980.

The circuit improvement offered by a fully variable demand-assignment communication system appears sufficient to justify its use by most administrations. It is believed that the satellite resource improvement provided by such a system justifies the establishment of an operational requirement that all trunks requiring 10 or less preassigned satellite circuits must be serviced via a demand-assignment system.

REFERENCES

[1] B. I. Edelson and A. M. Werth, "SPADE System Progress and Applications," *COMSAT Technical Review*, Vol. 2, No. 1, Spring 1972 (this issue).

[2] J. G. Puente and A. M. Werth, "Demand-Assigned Service for the INTELSAT Global Network," IEEE *Spectrum*, Vol. 8, No. 1, January 1971, pp. 59–69.

[3] International Telegraph and Telephone Consultative Committee (C.C.I.T.T.), Recommendation Q.5, "Advantages of Semi-automatic Service in the International Telephone Service," Recommendation Q.6, "Advantages of International Working," Recommendation Q.7, "Signalling Systems to Be Used for International Automatic and Semi-automatic Telephone Working," *IVth Plenary Assembly, Mar Del Plata, 1968, White Book*, Vol. VI, Part II, Chapter I, p. 1; Geneva: International Telecommunication Union, 1969.

[4] G. D. Dill and N. Shimasaki, "Signaling and Switching for Demand Assigned Satellite Communications," *INTELSAT/IEE International Conference on Digital Satellite Communications*, London, England, November 25–27, 1969, IEE Conference Publication 59, pp. 297–308.

[5] N. Shimasaki and G. D. Dill, "The Terrestrial Interface at SPADE Terminals," *AIAA Third Communication Satellite System Conference*, Los Angeles, California, April 6–8, 1970, Paper 70–413.

[6] International Telegraph and Telephone Consultative Committee (C.C.I.T.T.), "Table of the Erlang Loss Formula," *IVth Plenary Assembly, Mar Del Plata, 1968, White Book*, Vol. VI, Supplement 7, p. 1; "Curves Showing the Relation Between the Traffic Offered and the Number of Circuits Required," *op. cit.*, Supplement 8, pp. 1–2; Geneva: International Telecommunication Union, 1969.

[7] *Communication System Engineering Handbook*, edited by D. H. Hamsher, New York: McGraw-Hill, 1967, pp. 1–16, 7–2.

APPENDIX A. APPLICATION OF A FULLY VARIABLE DEMAND-ASSIGNMENT COMMUNICATION SYSTEM TO THE ATLANTIC OCEAN REGION

It is informative to compare the total number of low-density preassigned circuits actually in use in the Atlantic Ocean region with the number which would be required if any trunk having a total of 10 or less circuits were carried via a fully variable demand-assignment communication system.

Figure A–1 is a traffic matrix for the Atlantic Ocean region during the second quarter of 1971 [A1]. At the present time, there are 42 earth stations operating in the Atlantic Ocean region. Four additional destinations (Antigua, Barbados, Dominican Republic, and Trinidad) are serviced via Puerto Rico. Since these destinations are independently serviced, a total of 46 low-density destinations was assumed for the analysis. The total number of individual low-density trunks (n_t) in this network, each having n circuits, and the total number of two-way circuits (n_c) for each trunk size are listed in Table A–1. From Table A–1, the mean number of circuits per low-density trunk is

$$\bar{n} = \frac{\sum_{i=1}^{10} n(n_t)_i}{\sum_{i=1}^{10} (n_t)_i} \simeq 3.7 \qquad \text{(A1)}$$

and the average number of low-density trunks terminating at each earth station is

$$t = \frac{\sum_{i=1}^{10} (n_t)_i}{46} \simeq 2.3 \qquad \text{(A2)}$$

where n_t is the total number of two-way trunks per trunk size.

Given the mean trunk size and the requirement that these circuits shall provide a grade of service, P_B, of at least 0.01, one may use the erlang B equation to find that the average traffic load per trunk is approximately $a = 0.75$ erlang. Therefore, the computed total traffic load now being carried via preassigned low-density FDM/FM carriers is

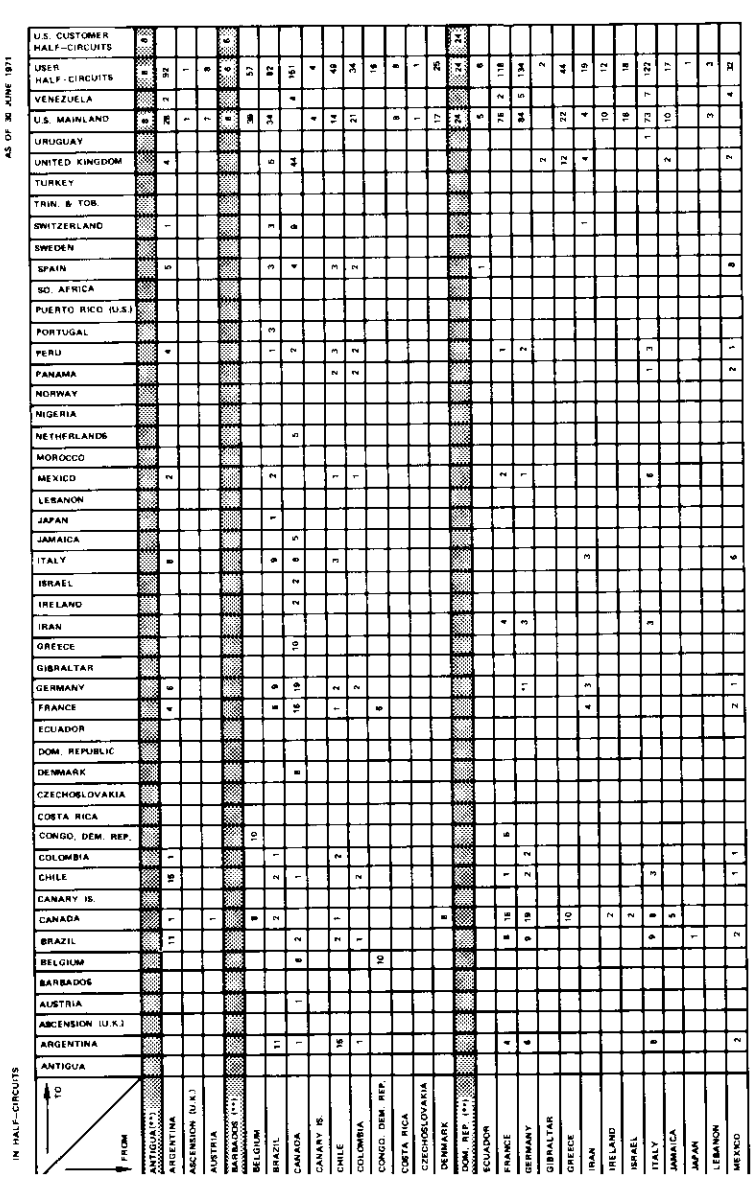
$$A_{PAS} = \frac{1}{2} \sum_{i=1}^{46} \left(\sum_{i=1}^{2.3} 0.75 \right) \simeq 39.7 \text{ erlangs.} \qquad \text{(A3)}$$

If this offered traffic load were serviced via SPADE, then a total of 52 satellite circuits would be required. In this case, the improvement* provided by SPADE

TABLE A-1. LOW-DENSITY TRAFFIC TRUNKS—ATLANTIC OCEAN REGION, 1971

Trunk Size, n	Total Two-Way Trunks, n_t, per Size	Total Two-Way Circuits, n_c, per Size
1	25	25
2	26	52
3	10	30
4	10	40
5	8	40
6	8	48
7	4	28
8	7	56
9	3	27
10	4	40
TOTAL	105	386

* The improvement factor for the 1971 Atlantic Region falls just below the predicted value (5 to 15). This is due to the following reasons: (1) some small trunks are now via-routed through transit exchanges due to the relatively high cost of downlink chains required for FDM/FM operation; (2) the unused channels in the preassigned carriers were not included in the analysis due to their variability.



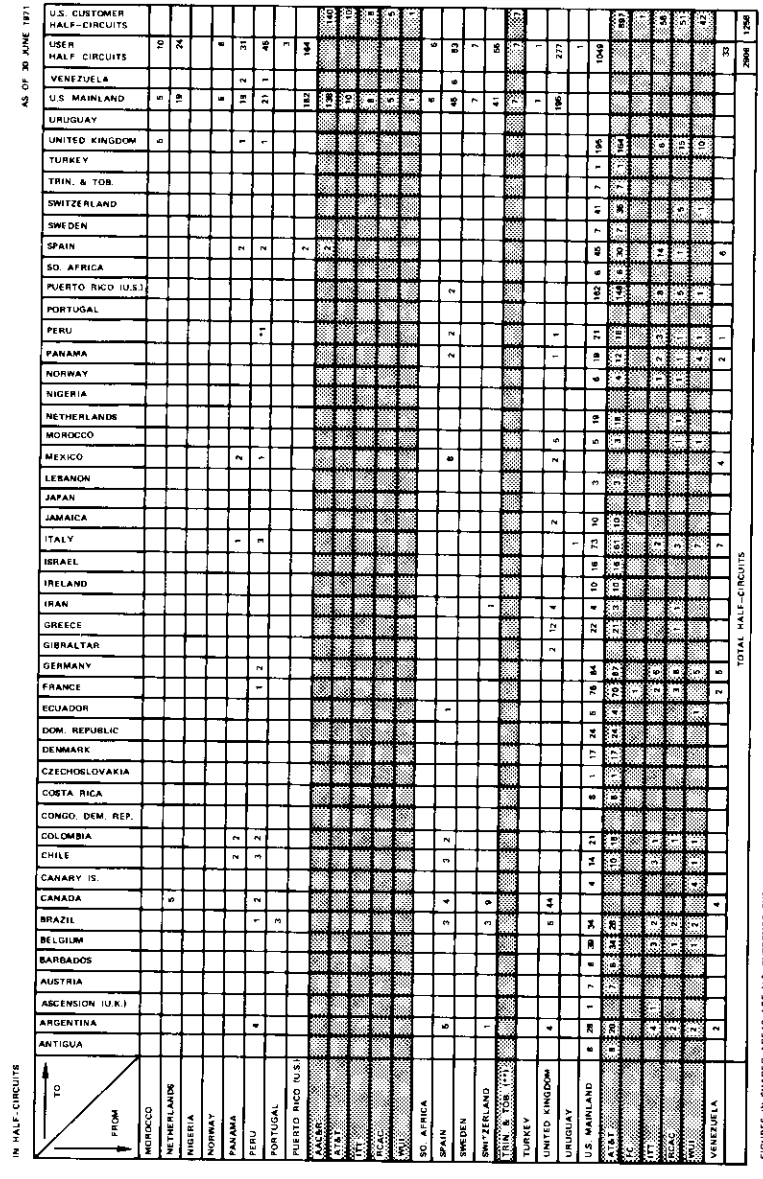


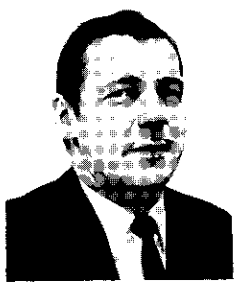
Figure A-1. Full-Time Service Over INTELSAT IV, F-2 and INTELSAT III, F-7 by User in the Atlantic Region (excluding NASA)

would be

$$I = \frac{n_{PA}}{n_{DA}} = \frac{193}{52} \cong 4. \qquad \text{(A4)}$$

REFERENCE

[A1] Communications Satellite Corporation, Second Quarterly 1971 Operational Report.



George D. Dill received a B.S. in applied physics from UCLA in 1957. For seven years he worked on satellite data processing and control systems at the Philco-Ford Western Development Laboratories. He joined COMSAT in September 1964 and is now Manager of the Telemetry, Tracking, and Command Branch in the Communications Processing Laboratory of COMSAT Laboratories.

ANALYSE D'UN COMPENSATEUR D'ECHO ADAPTABLE A LA REPONSE IMPULSIONNELLE DU CIRCUIT D'ECHO

S. J. CAMPANELLA, H. G. SUYDERHOUD ET M. ONUFRY

Sommaire

Sur les circuits téléphoniques à grande distance l'écho aux fréquences vocales est un phénomène indésirable bien connu. Pour la protection contre les échos, on utilise dans la technique classique des dispositifs commandés par la parole appelés suppresseurs d'écho. Leur fonctionnement présente une limitation fondamentale en raison d'une dichotomie de conception; cette limitation apparaît lorsque les deux interlocuteurs parlent simultanément.

Des techniques récentes ont permis d'envisager une conception radicalement différente, qui éliminera la dichotomie et, en particulier, améliorera la qualité des télécommunications par satellites. Le présent article décrit la conception et la performance d'un appareil appelé compensateur d'écho. Les résultats indiqués sont suffisants pour encourager la poursuite des études vers une réalisation économique.

UN SYSTEME DE CONVERSION A-N ET N-A FONCTIONNANT A LA VITESSE DE 150 MB/S

O. A. HORNA

Sommaire

Un système de conversion analogique-numérique et numérique-analogique a été mis au point pour la transmission numérique des signaux de télévision en couleur au moyen des satellites de télécommunications. Le système emploie cinq "blocs" fonctionnels différents sur des cartes de circuits imprimés classiques. Il permet d'obtenir une précision de 8 bits et un temps de conversion total inférieur à 50 ns.

La boucle de conversion A-N n'emploie aucun amplificateur opérationnel; toutes les opérations d'addition et de soustraction de tension se font directement a l'entrée des comparateurs. La boucle A-N comporte un circuit de réaction numérique pouvant corriger les erreurs de quantification dans l'étage, ou les étages d'entrée, réduisant ainsi les