THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 52 January 1973 Number 1

Measurements of Transfer Inefficiency of 250-Element Undercut-Isolated Charge Coupled Devices

By M. F. TOMPSETT, B. B. KOSICKI, and D. KAHNG

(Manuscript received August 10, 1972)

A 250-element charge coupled device is described in which the transfer electrodes are delineated and isolated using an undercut-etch technique. The device has metal electrodes on two thicknesses of oxide and is primarily intended to be operated in a two-phase manner. Measurements of transfer inefficiency as a function of frequency have been made on both n- and p-channel devices. Below 1 MHz, values of 4×10^{-4} per transfer independent of transfer frequency have been obtained. Above 1 MHz the transfer inefficiency progressively rises as the dynamics of charge motion limit the transfer of charge.

I. INTRODUCTION

A new method of fabricating charge coupled devices[1] (CCD) using the technique of undercut isolation has been reported recently.[2] A schematic cross section of a device made using this technique is shown in Fig. 1. The essential feature is a method of forming electrically isolated but self-aligned metal electrodes on two thicknesses of oxide. By connecting the electrodes in pairs, which may be done externally or using electrochemically plated regions on the device, as shown in Fig. 1, a two-level oxide structure[3,4] that may be operated in a two-

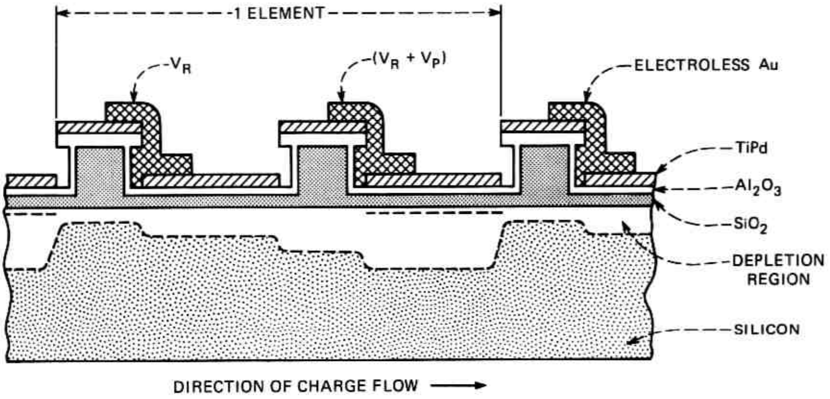


Fig. 1—Schematic longitudinal cross section of an undercut-isolated stepped-oxide CCD.

phase mode is obtained. This structure has several advantages over other structures. For example, compared to the three-phase structures, the geometrical constraints of having three phases, the requirement to fabricate 2- to 3-μm gaps, and the instabilities associated with the exposed oxide in these gaps are removed. Compared to other two-phase structures,[3-5] there is no need for a refractory metal technology or ion implantation, and the packing density of elements can be higher.

Test devices using undercut isolation and 250 elements long have been fabricated, and their transfer inefficiencies measured. The increase in number of elements from an earlier device[2] has allowed the small values of transfer inefficiency inherent in this structure to be measured accurately.

II. DEVICE FABRICATION

The 250-element CCD, which is the subject of this paper, was fabricated using the same photolithographic masks, except for two, as an earlier 500-element three-phase device[6] so that the undercut-isolated structure could be quickly evaluated. A photograph of one end of a finished device is shown in Fig. 2. The transfer region with the alternate thin and thick oxide levels is seen in the center of the photograph. The transfer electrodes are connected alternately on either side directly to two metal buses and via diffused cross-unders to two other buses. These cross-unders are not necessary for a two-phase CCD but were retained from the earlier three-phase device design to enable four-phase operation to be carried out for experimental purposes.

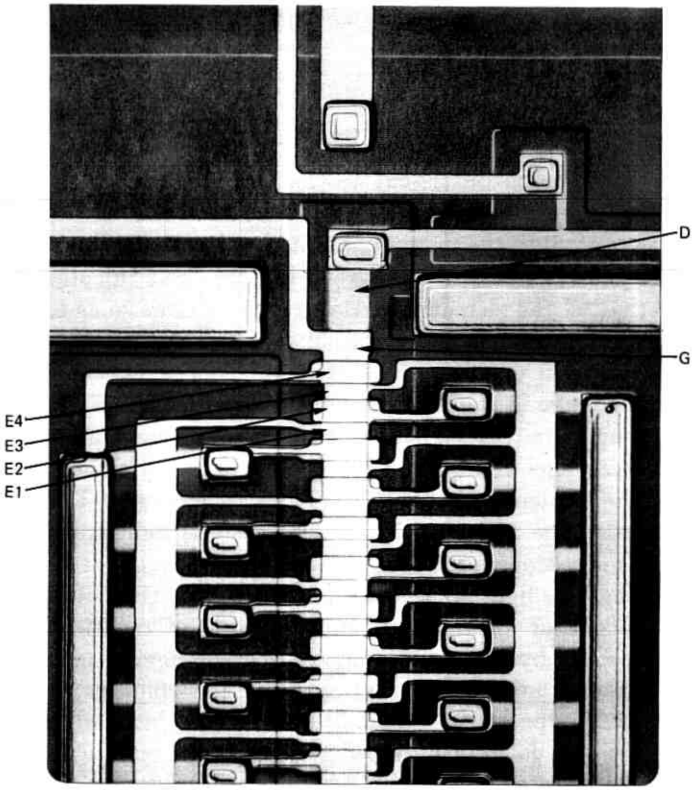


Fig. 2—View of the output section of a finished 250-element device, showing output diode D, output gate G, electrodes E1 and E3 over thick oxide, and electrodes E2 and E4 over thin oxide.

The transfer region is defined laterally by a "channel-stopping" diffusion that enhances the substrate doping. A device made exactly to the mask dimensions would have transfer electrodes that were 11 μm long over the thin oxide and 7 μm long over the thick oxide, with an 18-μm-wide channel. The devices were fabricated as described in an earlier paper[2] on both n- and p-type substrates.

III. MEASUREMENTS OF TRANSFER INEFFICIENCY

In order to measure the performance of the devices, voltages and pulses appropriate for either p- or n-channel devices, and two- or four-phase modes of operation were provided. Owing to circuit limitations, negative square pulses for driving p-channel devices up to frequencies of 10 MHz, and positive pulses up to 2 MHz, for testing n-channel devices were available. However, the n-channel devices could also be driven at up to 7 MHz using sinusoidal drive. As has already been mentioned, the device was made by modifying the design of an existing device, which had a very narrow transfer channel, so that the size of the output signal was not as large as is really desirable for easy and accurate measurements of transfer inefficiency. With a pulse voltage of 20 V, the maximum size of a charge packet was 0.5 pC. In all the measurements, a background charge was injected into the device so that all the elements carried a small charge so as to keep the interface states filled. Varying the amount of background charge in the range from 20 to 80 percent of a full charge packet caused no appreciable change in the measured transfer inefficiency.

At frequencies up to 2 MHz, the transfer inefficiency ϵ was measured by periodically injecting a single packet of charge into the device and observing the sequence of charge packets that emerged. The injection of charge was done either optically with a small light spot projected through a microscope or electrically. An advantage of the optical method is that the light spot could be moved near the output and the form of the output signal for a small number of transfers could be established. Also, by moving the spot along the device and observing the output signal, any discontinuities in transfer efficiency at a region of poor transfer, possibly caused by a partially blocked channel or an open electrode, could be detected and the device rejected. Obtaining a numerical value for ϵ from the observed sequence of output charge packets is based on comparison with the expected sequences[7,8] for different values of transfer inefficiency product $n\epsilon$, where n is the number of transfers.

Particularly for measuring values of $n\epsilon > 1$, it is more accurate to use another technique in which a sinusoidal input at different frequencies is fed to the device and the amplitude of the output measured. The frequency response of the device corrected for the response of the output amplifier is then plotted. Comparison with the theoretical response curves[7] enables values of $n\epsilon$ to be obtained. The advantage of this method is that values of transfer inefficiency for high values of drive frequency f_o could be obtained using input signals and amplifiers with bandwidths much lower than the drive frequency f_o.

IV. MEASURED VALUES OF TRANSFER INEFFICIENCY

A plot of transfer inefficiency versus frequency for both n- and p-channel devices operated in the two-phase mode is presented in

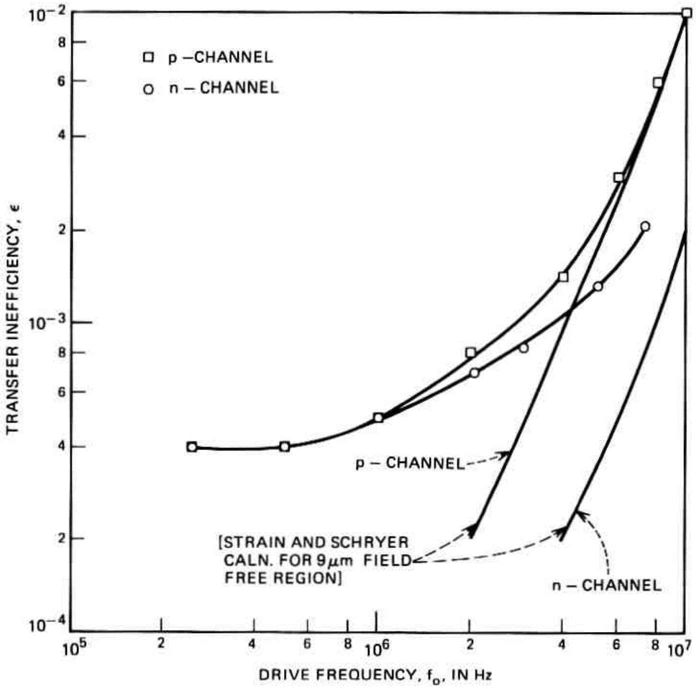
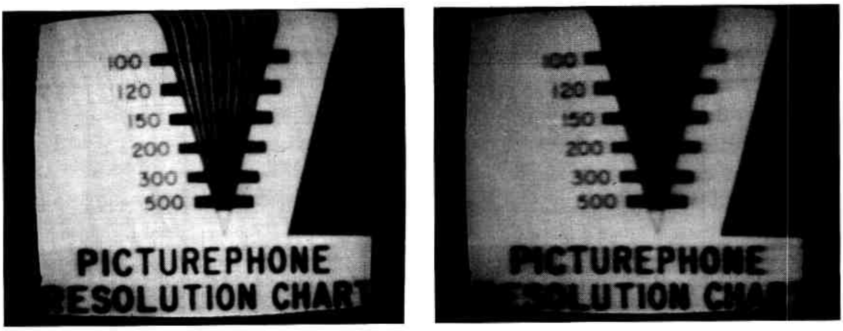


Fig. 3—Measurements of transfer inefficiency (per transfer) for both n- and p-channel undercut-isolated CCDs. The theoretical values for p- and n-channel devices assuming a 9-μm field free region under the transfer electrodes on the thin oxide have also been plotted.

Fig. 3. The transfer inefficiencies for both the p- and the n-channel devices, as predicted by calculations of charge motion[9] for devices with mobilities of 200 and 400 cm^2 V^{-1}s^{-1} respectively, are also shown on the figure. A 9-μm-long field free region is assumed under each electrode on the thin oxide, since the electric fields from the neighboring electrodes will penetrate at each end of the electrode.

Referring to Fig. 3, the transfer inefficiency of the devices appears to be flat below 0.5 MHz, perhaps due to limitations caused by interface states.[10] Above 0.5 MHz, the transfer inefficiency progressively degrades until, for the p-channel device, it rises exponentially following the theoretical curve. The rounding of the experimental curve is due to the joint contributions of the interface states and the dynamics of charge transfer. The greater carrier mobility in the n-channel devices is reflected in the lower transfer inefficiencies of these devices at frequencies in excess of 1 MHz.



DIRECT

DELAYED

Fig. 4—Showing the use of a 250-element n-channel undercut-isolated CCD to delay a *Picturephone*® video signal by 121 μs. The direct (undelayed) display is seen on the left and the delayed display on the right.

The transfer inefficiency per transfer obtained on the p-channel device operated in the four-phase mode at 1 MHz was 5×10^{-4}, which is about the same as that obtained in two-phase mode. However, in the two-phase mode, only half the number of transfers are required for a CCD with the same number of elements so that a device operated in this way has half the transfer inefficiency product $n\epsilon$ of one operated in four-phase mode. This is an important observation, not only because of the improved performance, but because of the additional advantage that two-phase interconnection gives to the design of functional devices.

The p-channel device was used, as described in more detail elsewhere,[11] to delay a *Picturephone*® video signal by 121 μs with barely noticeable degradation in the display as shown in Fig. 4.

V. CONCLUSIONS

The structure described has given transfer inefficiencies which are more than adequate to permit the design of devices for many applications. The two-level oxide structure with electroplated interconnections leads to some relatively simple designs of devices for various applications. The active region of the device is fully protected with a double-layer oxide and there are no exposed regions of oxide which can charge up and degrade the performance of the devices. There is no need for refractory metal electrodes and high-temperature processing to obtain a good second-level dielectric layer for insulation, or for fine features to be etched in the metallization as required in other structures. In addition, there is no critical reregistration required in the cell, so that compared to other structures, a smaller cell may be fabricated given the same fabrication tolerances. This would lead to higher packing densities and a capability of operating at higher frequencies. An encapsulant may be required to protect the undercut regions from dirt, damage, and electrical breakdown.

VI. ACKNOWLEDGMENTS

We acknowledge the technical assistance of D. A. Brooks, P. M. Ryan, and E. J. Zimany, Jr., in making the measurements.

REFERENCES

1. Boyle, W. S., and Smith, G. E., "Charge Coupled Semiconductor Devices," B.S.T.J., *49*, No. 4 (April 1970), pp. 587–593.
2. Powell, R. J., Berglund, C. N., Clemens, J. T., and Nicollian, E. H., "A Two-Phase Stepped Oxide CCD Shift Register Using Undercut Isolation," Appl. Phys. Ltrs., *20*, 1972, pp. 413–414.
3. Kahng, D., and Nicollian, E. H., U. S. Patent Number 3651349.
4. Kosonocky, W. F., and Carnes, J. E., "Charge Coupled Digital Circuits," IEEE J. Solid-State Circuits, *SC-6*, 1971, pp. 314–322.
5. Krambeck, R. H., Walden, R. H., and Pickar, K. A., "Implanted Barrier Two-Phase Charge Coupled Device," Appl. Phys. Ltrs., *19*, 1971, pp. 520–522.
6. Bertram, W. J., Sealer, D. A., Séquin, C. H., and Tompsett, M. F., "Recent Advances in Charge Coupled Imaging Devices," IEEE INTERCON Digest of Papers, 1972, pp. 292–293.
7. Joyce, W. B., and Bertram, W. J., "Linearized Dispersion Relation and Green's Function for Discrete Charge Transfer Devices with Incomplete Transfer," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1741–1759.
8. Tompsett, M. F., "Charge Transfer Devices," J. Vac. Sci. Tech., *9*, 1972, pp. 1166–1181.
9. Strain, R. J., and Schryer, N. L., "A Nonlinear Diffusion Analysis of Charge Coupled Device Transfer," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1721–1740.
10. Tompsett, M. F., "The Quantitative Effects of Interface States on the Performance of Charge Coupled Devices," International Electron Devices Meeting, Washington, October 1971. To be published, IEEE Trans. Electron Devices, January 1973.
11. Tompsett, M. F., and Zimany, E. J., Jr., "Use of Charge Coupled Devices for Analog Delay," ISSCC Digest of Technical Papers, 1972, pp. 136–137. To be published, IEEE J. Solid-State Circuits, April 1973.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

Formulas on Queues in Burst Processes–I

By B. GOPINATH, DEBASIS MITRA, and M. M. SONDHI*

(Manuscript received July 11, 1972)

Queues arising in buffers due to either random interruptions of the channel or variable source rates are analyzed in the framework of a single switched system. Examples of systems to which the results of the paper may be applied are: multiplexing of speech with data in telephone channels and, in certain instances, buffering of data generated by the coding of moving images in the Picturephone® *system. The switched system consists of a uniform source, buffer, switch and channel. The source feeds data to the buffer at a uniform rate. The buffer's access to the channel is controlled by the switch; if the switch is closed, the buffer empties to the extent of the channel's transmission rate. The on-off pattern of the switch is indicated by a* $0 - 1$ *burst process* $\{E_j\}$, $j = 0, 1, 2, \cdots$; *if* $E_j = 0$, *the switch is closed for the duration* $[j, j + 1)$. *The burst phenomenon is introduced to account for two different processes responsible for the event* $E_j = 0$. *There are relatively long periods during which* $E_j = 0$ *uniformly, and the activity separated by such periods is defined to be a burst. During a burst,* $E_j = 0$ *only infrequently. The duration of a burst is an independently distributed random variable with a geometric or weighted sum of geometric distributions. The inter-burst periods are assumed to be sufficiently long for the buffer to empty at some point during these periods of inactivity. During a burst* $\{E_j\}$ *is a Bernoulli sequence of independent random variables.*

Exact expressions for a variety of performance functionals related to the system described above are obtained, together with qualitative results. Recursive formulas are obtained for the following: (i) steady-state distribution of buffer content for a finite buffer of size N; *(ii) mean time for first passage across a level* N; *(iii) the probability of overflow, for a given level* N, *during a burst; (iv) mean time for first passage across a level* N *during a burst. The recursion in each case is with respect to* N. *The asymptotic behavior of the main recursions is determined.*

[*] The sequence of names was determined by coin tossing.

I. INTRODUCTION

A convenient framework for an unified analysis of a variety of digital communication systems involving buffering–some are discussed later–is provided by the system in Fig. 1. The source emits data uniformly at the rate of one symbol per unit time. The transmission rate of the channel is $(k + 1)$ symbols per unit time where k is some positive integer. The buffer has access to the channel only when the switch is closed. The switch is controlled by a burst process $\{E_j\}$, $j = 0, 1, 2, \cdots$. E_j, for every j, is either 0 or 1. If $E_j = 0$, the switch is closed for the duration $[j, j + 1)$; otherwise the switch is open. The burst process is introduced to account for cases where two basically different types of phenomena are responsible for the event $E_j = 0$. There are relatively long periods during which $E_j = 0$ uniformly; the activity separated by such periods is defined to be a burst. On the other hand, during a burst, $E_j = 0$ only infrequently. The duration or length of a burst is a random variable. It is assumed that the burst length is independently distributed with a geometric or a weighted sum of geometric distributions. The interburst periods are assumed to be sufficiently long for the buffer to empty during these periods. The statistical assumption made in the paper about the controlling sequence $\{E_j\}$ within a burst is that it is a Bernoulli sequence of independent random variables and $\Pr\{E_j = 1\} = \pi$ where $0 < \pi < 1$. In a companion paper, the case where $\{E_j\}$ is first-order Markov will be considered.

Important aspects of various digital communication systems are subsumed within the framework of the system described above. Diverse schemes for multiplexing data with speech on telephone channels[1,2] are representative of one class of such systems. A summary of the main features of the system which has been described in some detail in Ref. 1 follows. The central idea is to utilize the telephone channel during the periods of silence in speech which amount to as much as half of the total conversation period to transmit digital data. The speaker needs to have priority for the use of the channel since otherwise the quality of speech is impaired. $E_j = 0(1)$ corresponds to the decision that silence (speech) exists during the interval $[j, j + 1)$

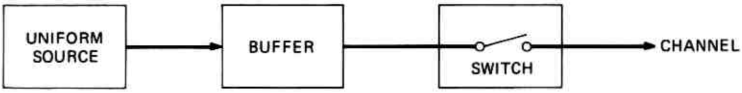


Fig. 1—Switched communication system.

so that only after it is decided that the speaker is silent does the buffer have access to the channel. An excellent example of the burst phenomenon may be found in speech monologues. Due to the presence of phrases in speech, two types of silences, interphrase and intraphrase silences, exist;[3,4] the former type consists of silences no less than 250 ms long and this is substantially longer than the mean duration of (uninterrupted) intraphrase silence.

There exists another class of digital communication systems composed of systems with only one source which transmits at a nonuniform rate. Most of the time the source rate is less than, say, r_0 bits per unit time and r_0 is less than the channel rate r. Occasionally, for short periods of time, the source rate spurts to a level r_1 which exceeds r. During such periods buffering becomes necessary. These occasional bursts of overloading of the channel are indicated by the $\{E_j\}$ process. The relation to the switched communication system of Fig. 1 is clear if $(r_1 - r)$ is normalized to unity, and $(r - r_0)$ corresponds to k.

An example of such a system for which the analysis of this paper is relevant arises in buffering of data generated by the coding of moving images in the *Picturephone*® system.[5] In this case, of course, r_0 and r_1 should be interpreted as average rates[6] in the two regimes, or, when the worst case is of interest, as the extreme rates. The results of this paper appear to be relevant[7] for variable rate in-frame coding, since during bursts of high detail, the correlation of the data rates for successive picture elements is not high. For frame-to-frame coding the first-order Markov model, to be treated in a companion paper, is of interest.

Exact expressions for diverse performance functionals related to the system in Fig. 1 are obtained, together with qualitative results. As a whole they provide a rather comprehensive set of criteria for the design of the important parameters of the system, such as the buffer size and the transmission rate of the channel. A summary of the main contributions follows.*

(*i*) A recursive formula is obtained for the steady-state distribution of buffer content for finite buffers. The recursion is with respect to N where N is the size of the buffer.

(*ii*) It is proved that F_N, the mean time for first passage through a level N, is given by

$$F_N = \frac{1}{\pi}F_{N-1} - \frac{1-\pi}{\pi}F_{N-k-1} + \frac{1}{\pi}\cdot$$

* N is used to denote both the buffer size [as in (*i*), (*iii*) and (*vi*)] and a level [as in (*ii*) and (*iv*)]. In what follows, the definition of N should be clear from the context.

(*iii*) g_N, the probability of overflow of a buffer of size N during a burst, the duration of which is distributed geometrically with a parameter ρ is given by

$$g_N = \left(\frac{1}{\rho\pi}\cdot\frac{1}{g_{N-1}} - \frac{1-\pi}{\pi}\cdot\frac{1}{g_{N-k-1}}\right)^{-1}.$$

(*iv*) A closed expression and a recursive formula are obtained for the mean time for first passage through a level N *during a burst;* the recursion is with respect to N.

(*v*) The asymptotic behavior of all formulas in items (*i*) through (*iii*), as N becomes large, is given. The derivation is dependent on the following: of the roots of the polynomials associated with the recursions, either one or two roots, depending on which recursion is being considered, lie outside the unit circle.

(*vi*) It is proved that under certain conditions on the initial probability distribution of the contents of the buffer, the probability of a buffer being full is a monotonic, nondecreasing sequence with respect to time; if the buffer is initially empty, the above-mentioned conditions are satisfied. One of the main implications of the result is that the steady-state probability of the buffer being full is an upper bound on the probabilities of the buffer being full at any instant. Furthermore, a particularly simple recursion is obtained for P_N, the steady-state probability of a buffer of size N being full:

$$P_N = \left(\frac{1}{\pi}\cdot\frac{1}{P_{N-1}} - \frac{1-\pi}{\pi}\cdot\frac{1}{P_{N-k-1}}\right)^{-1}.$$

(Observe that $1/P_N$ is also the mean time for recurrence of the state corresponding to a full buffer.)

The closed expressions obtained are for all k and N, and the recursions hold for all $N \geqq 2k + 1$. Wherever applicable, the buffer is assumed to be initially empty. An important feature of the given formulas is that they are also given in the form of recursions. The advantages of recursive formulas over the alternate versions need to be emphasized. For a given N, typically, a closed expression for a recursion involves inverting a matrix of order N. For large N, the effort is substantial. If, in addition, it is borne in mind that a designer is interested in functionals associated with a range of possible buffer sizes, the advantages of recursive formulas of the form given in this paper become overwhelming. This is only to be expected since the recursions are obtained by taking into full account the structure of the matrices involved.

II. EQUATIONS OF PROCESSES

Let B_j be the number of symbols in the buffer at the jth instant. For a finite buffer of size N,

$$\begin{aligned} B_{j+1} &= \mathrm{Max}\{B_j - k, 0\} \quad \text{if} \quad E_j = 0 \\ &= \mathrm{Min}\{B_j + 1, N\} \quad \text{if} \quad E_j = 1. \end{aligned}$$

Since B_j depends only on B_{j-1} and E_{j-1}, the state of the Markov chain of interest at the jth instant, S_j, is determined by the value of B_j where $B_j \in \{0, 1, 2, \cdots, N\}$. Let $P_m(n)$ denote the probability of the state $S_m = n$. Then

$$P_m(0) = (1 - \pi) \sum_{j=0}^{k} P_{m-1}(j) \tag{1}$$

$$P_m(i) = \pi P_{m-1}(i - 1) + (1 - \pi) P_{m-1}(i + k) \qquad i = 1, 2, \cdots, N - k \tag{2}$$

$$P_m(i) = \pi P_{m-1}(i - 1) \qquad i = N - k + 1, N - k + 2, \cdots, N - 1 \tag{3}$$

$$P_m(N) = \pi[P_{m-1}(N - 1) + P_{m-1}(N)]. \tag{4}$$

It is well known from the theory of Markov chains[8] that the limiting distribution of the states $P(i)$ is obtained from (1) through (4) by equating $P_m(i)$ and $P_{m-1}(i)$ to $P(i)$.

2.1 *Equations for Some New Probabilities*

Central to most of what follows are the probabilities $Q_m(i)$, where

$$Q_m(i) = \Pr\{(S_m = i) \cap (B_j \leqq N, j \leqq m)\}$$

and the buffer size exceeds N. For convenience, let X_m denote the event $S_j \in \{0, 1, 2, \cdots, N\}$ for all j, $0 \leqq j \leqq m$, so that

$$Q_m(i) = \Pr\{(S_m = i) \cap X_m\}. \tag{5}$$

The equation governing the transitions of $\{Q_i\}$ is derived. It is shown that there exists a matrix A which relates $\{Q_i\}$ to $\{Q_{i-1}\}$, i.e.,

$$Q_i(j) = \sum_{l=0}^{N} A_{jl} Q_{i-1}(l) \tag{6}$$

or, in matrix notation, $Q_i = AQ_{i-1}$.

In (5) $i \in \{0, 1, \cdots, N\}$, so that

$$\begin{aligned} Q_m(i) &= \Pr\{(S_m = i) \cap X_{m-1}\} \\ &= \sum_{j=0}^{N} \Pr\{(S_m = i) \cap X_{m-1} \cap (S_{m-1} = j)\}. \end{aligned}$$

Hence,

$$\begin{aligned} Q_m(i) &= \sum_{j=0}^{N} \Pr\{(S_m = i) \mid (S_{m-1} = j) \cap X_{m-1}\} \Pr\{(S_{m-1} = j) \cap X_{m-1}\} \\ &= \sum_{j=0}^{N} \Pr\{(S_m = i) \mid (S_{m-1} = j) \cap X_{m-1}\} Q_{m-1}(j) \\ &= \begin{cases} (1-\pi) \sum_{j=0}^{k} Q_{m-1}(j) \quad \text{if} \quad i = 0 & \text{(7a)} \\ \pi Q_{m-1}(i-1) + (1-\pi) Q_{m-1}(i+k) & \\ \qquad\qquad \text{if} \quad i = 1, 2, \cdots, N-k & \text{(7b)} \\ \pi Q_{m-1}(i-1) & \\ \qquad\qquad \text{if} \quad i = N-k+1, N-k+2, \cdots, N. & \text{(7c)} \end{cases} \end{aligned}$$

(7) defines the $(N+1)$ by $(N+1)$ matrix A. Sometimes when the need arises, the $(N+1)$ by $(N+1)$ matrix A associated with a given N will be specified by $A(N)$.

$$A = \begin{array}{cccccccccc} 1 & 2 & & k+1 & k+2 & & & & N+1 & \\ (1-\pi) & (1-\pi) & \cdots & (1-\pi) & & & & & & 1 \\ \pi & 0 & & & (1-\pi) & & & & & 2 \\ & \ddots & & & & \ddots & & & & \\ & & \ddots & & & & & & (1-\pi) & N-k+1 \\ & & & \ddots & & & & & & \\ & & & & \cdot\pi & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & \cdot\pi & 0 & 0 & N \\ & & & & & & & \pi & 0 & N+1 \end{array} \tag{8)*}$$

It will be observed that the only difference between (7) and the transition equations (1) through (4) for a finite buffer of size N, is that eq. (4) is modified since a transition from state N to state N is not possible in the present context. For the same reason, the matrix A is not a Markov matrix since the sum of the elements of the last, i.e., $(N+1)$th column is $(1-\pi)$, the remaining columns sum to unity as is the case for all columns of Markov matrices.

* The dots indicate continuation of the values of the adjoining elements; remaining elements are assumed to be 0.

If the transition matrix of the basic Markov process, i.e., the matrix defined by eqs. (1) through (4), is irreducible, then $(\mathrm{I} - \rho A)$ is nonsingular for $|\rho| \leqq 1$. The proof follows from a well-known result in matrix theory[9] which in this case states that if all the columns of $(\mathrm{I} - \rho A)$ are weakly column-sum dominant and at least one column of $(\mathrm{I} - \rho A)$ is strongly column-sum dominant, then the matrix is nonsingular.

III. STEADY-STATE PROBABILITIES FOR FINITE BUFFERS

In this section, a formula is given for recursively generating the steady-state probabilities $P(i)$ where the recursion is, with respect to N, the size of the buffer. To distinguish the steady-state probabilities for different buffer sizes, the symbol $P^N(i)$ is introduced to denote $P(i)$ for a buffer of size N.

If $N \geqq k + 1$, as is almost always the case, an equation of the type given in (2), namely,

$$P^N(i-1) - \frac{1}{\pi}P^N(i) + \frac{1-\pi}{\pi}P^N(i+k) = 0 \tag{9}$$

occurs at least once and since $N \gg k$ usually, the main body of equations defining the steady-state probabilities is of that form. It is proved in Ref. 1 what may reasonably be expected, namely, every solution of the homogenous set of equations that define the steady-state probabilities is of the form

$$P^N(j) = \sum_{i=1}^{k+1} b_i \mu_i^{N-j} \qquad j = 0, 1, \cdots, N \tag{10}$$

where μ_i are the simple roots of the polynomial

$$\mu^{k+1} - \frac{1}{\pi}\mu^k + \frac{1-\pi}{\pi}. \tag{11}$$

If the polynomial has multiple roots the obvious modifications must be made. [Note: Since $0 < \pi < 1$, the polynomial in (11) has distinct roots whenever $\pi \neq k/(k+1)$; when $\pi = k/(k+1)$, the only repeated root is 1.]

The complete recursive formula for $P^N(j)$ is obtained in two parts. First, a recursive formula for a set of solutions $q_N(j)$ to the steady-state equations is obtained and, second, a recursive formula for the normalizing constant Σ_N is obtained. Finally,

$$P^N(j) = \frac{1}{\Sigma_N} q_N(j) \quad j = 0, 1, \cdots, N. \tag{12}$$

3.1 *Recursions for* $\{q_N(j)\}$

Let

$$q_N(N) = 1 \tag{13}$$

and suppose $\{q_N(j)\}$ satisfies the steady-state equations of a finite buffer of size N. Hence, $q^N(j)$ has the form given in (10).* For fixed N and $i = 1, 2, \cdots, k + 1$, let

$$d_i \triangleq \sum_{j=1}^{k+1} a_j \mu_j^{i-1}.$$

The transformation $\{a_j\} \rightarrow \{d_i\}$ is invertible since the Vandermonde matrix is nonsingular. Now,

$$\begin{aligned} d_i &= \sum_{j=1}^{k+1} a_j \mu_j^{N-(N-i+1)} \\ &= q_N(N - i + 1) \qquad i = 1, 2, \cdots, k + 1. \end{aligned} \tag{14}$$

Also, from the steady-state equations themselves,

$$d_1 = q_N(N) = 1 \tag{15}$$

$$d_i = q_N(N - i + 1) = \frac{1 - \pi}{\pi^{i-1}} \qquad i = 2, 3, \cdots, k + 1. \tag{16}$$

Hence, significantly, $\{d_i\}$ is independent of N from which it follows that $\{a_i\}$ is also independent of N.

$$\begin{aligned} q_{N+k+1}(j) &= \sum_{i=1}^{k+1} a_i \mu_i^{N+k+1-j} \\ &= \sum_{i=1}^{k+1} a_i \left\{ \frac{1}{\pi} \mu_i^{(N+k)-j} - \frac{1 - \pi}{\pi} \mu_i^{N-j} \right\} \qquad \text{from (11)} \\ &= \frac{1}{\pi} q_{N+k}(j) - \frac{1 - \pi}{\pi} q_N(j) \qquad j = 0, 1, \cdots, N. \end{aligned} \tag{17}$$

The formula for $\{q_{N+k+1}(j)\}$ is complete if (15) and (16) are appended,

[*] To distinguish between $\{P^N(j)\}$ and $\{q_N(j)\}$, denote the coefficients in the form for the latter by $\{a_i\}$, i.e., $b_i = (1/\Sigma_N) a_i$.

i.e.,

$$q_{N+k+1}(N + i) = \frac{1 - \pi}{\pi^{k-i+1}} \qquad i = 1, 2, \cdots, k \tag{16}$$

$$q_{N+k+1}(N + k + 1) = 1. \tag{15}$$

3.2 *Recursion for the Normalizing Constant*

Let

$$\Sigma_N \triangleq \sum_{j=0}^{N} q_N(j). \tag{18}$$

Now

$$\begin{aligned}\sum_{j=N+1}^{N+k+1} q_{N+k+1}(j) &= 1 + (1 - \pi) \sum_{i=1}^{k} \left(\frac{1}{\pi}\right)^i \\ &= \frac{1}{\pi^k}.\end{aligned} \tag{19}$$

Summing both sides of (17),

$$\Sigma_{N+k+1} - \frac{1}{\pi^k} = \frac{1}{\pi}\left[\Sigma_{N+k} - \frac{1}{\pi^{k+1}}\right] - \frac{1 - \pi}{\pi}\Sigma_N$$

i.e.,

$$\Sigma_{N+k+1} = \frac{1}{\pi}\Sigma_{N+k} - \frac{1 - \pi}{\pi}\Sigma_N. \tag{20}$$

(20) is the recursion for the normalizing constant. The derivation of the recursive formula for $\{P^N(j)\}$ is now complete.

Observe that in the course of the above analysis, a simple recursive formula for the rather important steady-state probability of the buffer being full, i.e., $P^N(N)$, has been obtained.

$$P^N(N) = \frac{q_N(N)}{\Sigma_N} = \frac{1}{\Sigma_N} \tag{21}$$

and Σ_N, of course, is obtained from (20).

IV. MEAN FIRST PASSAGE TIME

Suppose N is a fixed positive integer and the buffer capacity is greater than N. A functional that provides substantial insight into the problem of designing a buffer for which the probability of overflow is small is F_N, the mean time required for the buffer contents to first exceed N. It is particularly useful in the context of burst processes where only incomplete data are available concerning the burst length distribution–provided that the length of bursts is bounded, a simple comparison of the bound with F_N provides an useful guide. In this section a recursive formula for F_N, the recursion being with respect to N, is obtained. To correspond with the practical situation, the buffer is initially assumed to be empty; the same recursive formula holds for the other interesting initial condition, namely, the buffer initially contains an unit symbol.

X_m is the event that $S_j \in \{1, 2, \cdots, N\}$ for all j, $0 \leqq j \leqq m$.

$$O_i \triangleq \Pr\{\text{overflow occurs for the first time at } i\} \tag{22}$$

$$\begin{aligned} &= \Pr\{(E_{i-1} = 1) \cap (S_{i-1} = N) \cap X_{i-1}\} \\ &= \pi \Pr\{(S_{i-1} = N \cap X_{i-1})\}, \text{ from the independence of } \{E_i\} \\ &= \pi Q_{i-1}(N) \end{aligned} \tag{23}$$

where $\{Q_i\}$ is as defined in eq. (5). It has been shown in Section 2.1 that

$$Q_i = AQ_{i-1}. \tag{6}$$

Hence,

$$\begin{aligned} O_i &= \pi Q_{i-1}(N) \\ &= \pi(0, \cdots, 0, 1)Q_{i-1} \\ &= \pi(0, \cdots, 0, 1)A^{i-1}Q_0 \\ &= \pi e_r^t A^{i-1} Q_0 \end{aligned} \tag{24}$$

where e_i denotes the vector* with a single element equal to unity at the ith location and all remaining elements 0; $r = N + 1$. Finally,

$$F_N = \text{Mean time for first passage through level } N$$

$$= \sum_{i=1}^{\infty} iO_i \tag{25}$$

$$= \pi \sum_{i=1}^{\infty} ie_r^t A^{i-1} Q_0$$

$$= \pi e_r^t \left(\sum_{i=1}^{\infty} iA^{i-1} \right) Q_0$$

$$= \pi e_r^t (I - A)^{-1}(I - A)^{-1} Q_0. \tag{26}$$

Let

$$x^t \triangleq e_r^t (I - A)^{-1}$$

* The superscript t denotes the transpose of a vector.

so that

$$x^t(I - A) = e_r^t. \tag{27}$$

But

$$x^t = \frac{1}{\pi}(1, 1, \cdots, 1) \tag{28}$$

is a solution of (27) since the elements of the last, i.e., $(N + 1)$th column of A sum to $(1 - \pi)$ and the remaining columns sum to 1. Moreover, (28) is the unique solution of (27) since $(I - A)$ is non-singular. Hence,

$$\begin{aligned} F_N &= (1, 1 \cdots, 1)(I - A)^{-1}Q_0 \\ &= 1^t(I - A)^{-1}Q_0 \end{aligned} \tag{29}$$

where 1 denotes the vector with all elements equal to unity. In the following section, the above formula with $Q_0 = e_1$ is analyzed further to yield a recursive formula.

4.1 *Recursive Formula for F_N*

Henceforth, it is necessary to be specific about the dimensions of A–the matrix A associated with a given N is denoted by $A(N)$. The buffer is assumed to be initially empty, i.e., $S_0 = 0$ or, equivalently, $Q_0 = e_1$.

Since* $|I - A(N)|[I - A(N)]^{-1}e_1$ is the vector of (signed) cofactors of the 1st row of $[I - A(N)]$

$$|I - A(N)|1^t[I - A(N)]^{-1}e_1 = |D(N)| \tag{30}$$

where $D(N)$ is the $(N + 1)$ by $(N+1)$ matrix obtained from $A(N)$ by replacing all elements of the first row of $A(N)$ by unity. Then, from (29),

$$F_N = \frac{|D(N)|}{|I - A(N)|}. \tag{31}$$

Adding rows 2, 3, $\cdots$, $(N + 1)$ of $[I - A(N)]$ to the first row, it can be verified that

$$|I - A(N)| = \pi^{N+1}. \tag{32}$$

In Appendix A it is shown that

$$|D(N)| = |D(N - 1)| - (1 - \pi)\pi^k|D(N - k - 1)| + \pi^N. \tag{33}$$

* $|X|$ denotes the determinant of the matrix X.

Hence,

$$\frac{|D(N)|}{\pi^{N+1}} = \frac{1}{\pi}\cdot\frac{|D(N-1)|}{\pi^{N}} - \frac{(1-\pi)\pi^{k}}{\pi^{k+1}}\cdot\frac{|D(N-k-1)|}{\pi^{N-k}} + \frac{1}{\pi}, \quad (34)$$

i.e.,

$$F_N = \frac{1}{\pi}F_{N-1} - \frac{1-\pi}{\pi}F_{N-k-1} + \frac{1}{\pi}\cdot \quad (35)$$

The above relation is the desired recursive formula for the mean first passage time. It was obtained under the assumption that the buffer is initially empty. An alternative assumption about the initial distribution, which is also of interest, is that the buffer contains an unit symbol, i.e., $S_0 = 1$. It may be shown that even for this case the mean first passage time satisfies the formula (35) though, of course, the initial conditions to the recursion in the formula are different.

V. PROBABILITY OF NO OVERFLOW IN A BURST

The results of this section are useful when information concerning the length of bursts is available. It is assumed that the distribution of burst length may be expressed as a weighted sum of geometric distributions. Given below are formulas which yield the probability that the contents of the buffer during bursts do not exceed N, a given positive integer.

At this stage, assume that the distribution of burst length is geometric; the generalization to distributions that are weighted sums of geometric distributions will be taken up later. If the burst length is denoted by l, then

$$\Pr\{l = i\} = (1-\rho)\rho^{i-1} \quad i = 1, 2, \cdots. \quad (36)$$

for some ρ, $0 < \rho < 1$. Let $G_N \triangleq \Pr$ {buffer contents do not exceed N during a burst}. The usual decomposition into mutually exclusive events yields

$$\begin{aligned} G_N &= \sum_{m\geqq 1} \Pr\{S_j \in (0, 1, \cdots, N), \quad j = 0, 1, \cdots, m; \\ &\qquad\qquad\qquad\qquad \text{and burst length} = m\} \\ &= \sum_{m\geqq 1} \Pr\{X_m \cap l = m\} \\ &= \sum_{m\geqq 1} \Pr\{X_m\}\Pr\{l = m\}. \end{aligned} \quad (37)$$

The last relation holds since, $\Pr\{X_m|l = m\} = \Pr\{X_m\}$. Now,

$$\begin{aligned}\Pr\{X_m\} &= \sum_{i=0}^{N} \Pr\{S_m = i \cap X_m\} \\ &= \sum_{i=0}^{N} Q_m(i) \qquad\qquad \text{from (5),} \\ &= 1^t Q_m \\ &= 1^t A^m Q_0 \qquad\qquad (38)\end{aligned}$$

where A is the $(N + 1)$ by $(N + 1)$ transition matrix defined in Section 2.1 and Q_0 is the vector given by the initial distribution–it may be assumed that $S_0 \in (0, 1, \cdots, N)$. Hence

$$\begin{aligned}G_N &= \sum_{m \geqq 1} 1^t A^m Q_0 (1 - \rho)\rho^{m-1} \\ &= \frac{(1-\rho)}{\rho} 1^t \left\{ \sum_{m \geqq 1} (\rho A)^m \right\} Q_0 \\ &= \frac{1-\rho}{\rho} 1^t \{ (I - \rho A)^{-1} - I \} Q_0 \\ &= \frac{(1-\rho)}{\rho} \{ 1^t (I - \rho A)^{-1} Q_0 - 1 \} \cdot \qquad (39)\end{aligned}$$

In the sequel, a recursive formula for G_N is developed for the case where $S_0 = 0$ or, equivalently, $Q_0 = e_1$.

5.1 *Recursive Formula for G_N*

The matrix A associated with a given N is denoted by $A(N)$. $|I - \rho A(N)| \{I - \rho A(N)\}^{-1} e_1$ is the vector of (signed) cofactors of the first row of $\{I - \rho A(N)\}$. Therefore, $|I - \rho A(N)| 1^t \{1 - \rho A(n)\}^{-1} e_1$ is the determinant of the matrix $B(N)$ obtained by replacing every element of the first row of $\{I - \rho A(N)\}$ by unity.

$$1^t \{1 - \rho A(N)\}^{-1} e_1 = \frac{|B(N)|}{|I - \rho A(N)|} \cdot \qquad (40)$$

Let the (signed) cofactor of the element $\{I - \rho A(N)\}_{1i}$ be denoted by C^{1i}, $i = 1, 2, \cdots, N + 1$. From the definition of $B(N)$,

$$|B(N)| = \sum_{i=1}^{N+1} C^{1i}. \qquad (41)$$

The elements of the $(N+1)$th column of $\{I - \rho A(N)\}$ sum to $\{1 - \rho(1-\pi)\}$ and the elements of the remaining columns sum to $(1-\rho)$. Hence, by adding the rows $2, 3, \cdots, N+1$ to row 1, it follows that

$$\begin{aligned} |I - \rho A(N)| &= (1-\rho) \sum_{i=1}^{N} C^{1i} + \{1 - \rho(1-\pi)\} C^{1,N+1} \\ &= (1-\rho) \sum_{i=1}^{N+1} C^{1i} + \{1 - \rho(1-\pi) - (1-\rho)\} C^{1,N+1} \\ &= (1-\rho)|B(N)| + \rho\pi C^{1,N+1} \qquad \text{from (41)}, \end{aligned}$$

i.e.,

$$\frac{|B(N)|}{|I - \rho A(N)|} = \frac{1}{1-\rho} - \frac{\rho\pi}{1-\rho} \frac{C^{1,N+1}}{|I - \rho A(N)|}. \tag{42}$$

Recapitulating,

$$\begin{aligned} G_N &= \frac{1-\rho}{\rho} [1^t\{1 - \rho A(N)\}^{-1} e_1 - 1] && \text{from (39)} \\ &= \frac{1-\rho}{\rho} \left[\frac{|B(N)|}{|I - \rho A(N)|} - 1 \right] && \text{from (40)} \\ &= 1 - \pi \frac{C^{1,N+1}}{|I - \rho A(N)|} && \text{from (42).} \end{aligned} \tag{43}$$

The remainder of the derivation is in two parts. First, a closed form expression for $C^{1,N+1}$ is obtained. The second part is on the recursive formula for $|I - \rho A(N)|$ and this formula is derived in Appendix B.

$$I - \rho A(N) = \begin{array}{cccccccccc|l}
1 & 2 & 3 & & k+1 & k+2 & & & & N+1 & \\
\hline
(1+\lambda) & \lambda & \lambda & \ldots & \lambda & & & & & & 1 \\
\mu & 1 & 0 & & & \lambda & & & & & 2 \\
 & \ddots & \ddots & \ddots & & & \ddots & & & & \\
 & & & \mu & 1 & 0 & & & & \lambda & N-k+1 \\
 & & & & \ddots & \ddots & \ddots & & & & \\
 & & & & & & & \mu & 1 & 0 & N \\
 & & & & & & & & \mu & 1 & N+1
\end{array} \tag{44}$$

where $-\rho(1-\pi) = \lambda$ and $-\rho\pi = \mu$. $C^{1,N+1}$, the (signed) cofactor to $\{I - \rho A(N)\}_{1,N+1}$, is the signed determinant of an upper triangular matrix;

$$\begin{aligned} C^{1,N+1} &= (-1)^{N+2}\mu^N = (-1)^{N+2}(-\rho\pi)^N \\ &= \rho^N\pi^N. \end{aligned} \tag{45}$$

In Appendix B it is shown that if x_N, a scalar, is used to denote $|I - \rho A(N)|$, then the following recursive formula holds:

$$x_N = x_{N-1} - \rho^{k+1}\pi^k(1-\pi)x_{N-k-1}. \tag{46}$$

Hence,

$$\frac{x_N}{\pi^N\rho^N} = \frac{1}{\pi\rho}\left(\frac{x_{N-1}}{\pi^{N-1}\rho^{N-1}}\right) - \frac{1-\pi}{\pi}\left(\frac{x_{N-k-1}}{\pi^{N-k-1}\rho^{N-k-1}}\right). \tag{47}$$

Let

$$y_N \triangleq \frac{x_N}{\pi^N\rho^N} = \frac{|I - \rho A(N)|}{\pi^N\rho^N} \tag{48}$$

so that,

$$y_N = \frac{1}{\pi\rho}\,y_{N-1} - \frac{1-\pi}{\pi}\,y_{N-k-1}. \tag{49}$$

From (43) and (45),

$$G_N = 1 - \frac{\pi^{N+1}\rho^N}{|I - \rho A(N)|}$$

i.e.,

$$G_N = 1 - \frac{\pi}{y_N}. \tag{50}$$

(49) and (50) together provide the desired recursive formula for the probability that the contents of the buffer does not exceed a given level N during bursts if the buffer is initially empty and the distribution of burst lengths is geometric.

Suppose the distribution of burst lengths is the weighted sum of geometric distributions; i.e.,

$$\Pr\{\text{burst length} = i\} = \sum_{j=1}^{J} \alpha_j(1-\rho_j)(\rho_j)^{i-1}. \tag{51}$$

It may then be shown that $G_N = \sum_{j=1}^{J}\alpha_j G_{j,N}$ where $G_{j,N}$ is obtained from (50) and (49) with ρ replaced by ρ_j in the latter equation and $G_{j,N}$ identified with G_N.

VI. MEAN TIME FOR FIRST PASSAGE IN A BURST

In Section IV certain formulas for the mean time for first passage across prespecified levels are given. In burst processes where data regarding the length of bursts is available, a more meaningful functional is one in which a level is defined to be crossed only if this event occurs during a burst. Bursts, then, may be visualized as a period of observation of the buffer. First passage across N, a positive integer, is defined to occur at i if

$$\{S_j \leqq N,\ j = 0, 1, 2, \cdots, i-1 \text{ and } S_i = N+1 \text{ and, burst length } \geqq i\}.$$

Let R_i denote the probability of this event. The functional of interest is $H_N = \sum_{i=1}^{\infty} iR_i$. The burst length distribution is assumed to be geometric; generalization to larger classes of distributions may be undertaken as indicated in the preceding section. Hence, if l denotes burst length,

$$\Pr\{l = i\} = (1-\rho)\rho^{i-1} \quad i = 1, 2, \cdots \tag{52}$$

for some ρ, $0 < \rho < 1$.

In the notation of Section 2.1,

$$\begin{aligned} R_i &= \Pr\{S_{i-1} = N \cap X_{i-1} \cap E_{i-1} = 1 \cap l \geqq i\} \\ &= \Pr\{E_{i-1} = 1\}\Pr\{S_{i-1} = N \cap X_{i-1} | l \geqq i\}\Pr\{l \geqq i\} \\ &= \pi \Pr\{S_{i-1} = N \cap X_{i-1}\}\Pr\{l \geqq i\} \\ &= \pi Q_{i-1}(N)\rho^{i-1} \\ &= \pi \overset{t}{e_r}(\rho A)^{i-1}Q_0. \end{aligned} \tag{53}$$

A is, of course, the $(N+1)$ by $(N+1)$ matrix defined in Section 2.1 and Q_0 is the initial condition vector.

$$H_N = \pi \overset{t}{e_r}\Big(\sum_{i \geqq 1} i(\rho A)^{i-1}\Big)Q_0$$

i.e.,

$$H_N = \pi \overset{t}{e_r}(I - \rho A)^{-1}(I - \rho A)^{-1}Q_0. \tag{54}$$

The above concludes the derivation of the closed formula for H_N–the rest of the section is concerned with recursive versions of the formula for the case where the buffer is initially empty, i.e., $Q_0 = e_1$. Once again, it is necessary to revert to the use of the symbol $A(N)$ to denote the matrix A associated with N.

For fixed N,

$$z(\rho) \triangleq e_r^t \sum_{i \geqq 0} \rho^{i+1} A^i(N) e_1$$

$$= \rho e_r^t \sum_{i \geqq 0} \{\rho A(N)\}^i e_1$$

$$= \rho e_r^t \{I - \rho A(N)\}^{-1} e_1. \tag{55}$$

Hence,

$$z'(\rho) = \frac{d}{d\rho} z(\rho) = e_r^t \sum_{i \geqq 0} (i+1) \rho^i A^i(N) e_1$$

$$= e_r^t \sum_{i \geqq 1} i \{\rho A(N)\}^{i-1} e_1$$

$$= \frac{1}{\pi} H_N \cdot \tag{56}$$

Returning to $z(\rho)$ and (55), observe that

$$z(\rho) = \frac{\rho C^{1,N+1}}{|I - \rho A(N)|} \tag{57}$$

$$C^{1,N+1} = \rho^N \pi^N. \tag{45}$$

Hence,

$$z(\rho) = \frac{\rho^{N+1} \pi^N}{|I - \rho A(N)|} \tag{58}$$

and, from (56),

$$H_N = \frac{d}{d\rho} \left\{ \frac{\rho^{N+1} \pi^{N+1}}{|I - \rho A(N)|} \right\}.$$

Let

$$\frac{1}{v_N} \triangleq \frac{\rho^{N+1} \pi^{N+1}}{|I - \rho A(N)|}. \tag{59}$$

Since $v_N = (1/\rho\pi) y_N$ where y_N has been defined previously in (48) and the recursion in (49) for y_N is linear, v_N satisfies the same recursion. Hence, with $u_N \triangleq (d/d\rho) v_N(\rho)$, the following formula is obtained:

$$H_N = -\frac{u_N}{v_N^2} \tag{60}$$

and,

$$\begin{cases} v_N = \dfrac{1}{\rho\pi} v_{N-1} - \dfrac{1-\pi}{\pi} v_{N-k-1} & (61) \\ u_N = \dfrac{1}{\rho\pi} u_{N-1} - \dfrac{1-\pi}{\pi} u_{N-k-1} - \dfrac{1}{\rho^2\pi} v_{N-1}. & (62) \end{cases}$$

VII. ASYMPTOTICS OF RECURSIONS

The main recursions occurring in the paper are of the following forms:

$$x_N = \frac{1}{\pi} x_{N-1} - \frac{1-\pi}{\pi} x_{N-k-1} \tag{63}$$

$$y_N = \frac{1}{\pi} y_{N-1} - \frac{1-\pi}{\pi} y_{N-k-1} + \frac{1}{\pi} \tag{64}$$

$$z_N = \frac{1}{\rho\pi} z_{N-1} - \frac{1-\pi}{\pi} z_{N-k-1} \qquad \text{where } 0 < \rho < 1. \tag{65}$$

Equation (63) occurs in the formula for the (unnormalized) steady-state probabilities and in the formula for the normalization constant; (64) occurs in the formula for the mean first passage time; (65) occurs in the formula for the probability of no overflow during bursts. The fundamental solutions of these recursions are obtained from the roots of the following polynomials.

$$F(\mu) \triangleq \mu^{k+1} - \frac{1}{\pi} \mu^k + \frac{1-\pi}{\pi} \tag{66}$$

$$G(\mu) = \mu^{k+1} - \frac{1}{\rho\pi} \mu^k + \frac{1-\pi}{\pi}. \tag{67}$$

Equation (66) is associated with (63) and (64); (67) with (65). The two results given below enumerate and estimate the roots of $F(\mu)$ and $G(\mu)$ outside the unit circle.

Lemma 1[1]*: Except for one positive real root $1/\theta$, and 1, all other roots of $F(\mu)$ lie inside the unit circle $|\mu| \leqq 1$. The root $1/\theta$ lies outside the unit circle if and only if $k > \pi/(1-\pi)$.*

Lemma 1 is a specialization of a result proved in Ref. 1. Bounds on θ are also given there.

Lemma 2: $G(\mu)$ has k roots inside the unit circle $|\mu| \leqq 1$, no roots in the annular ring $1 \leqq |\mu| \leqq 1/\rho$, and one real, positive root outside the circle $|\mu| \leqq 1/\rho$.

Proof:

$$G(0) = \frac{1 - \pi}{\pi} > 0$$

$$G(1) = \frac{1}{\pi}(1 - 1/\rho) < 0$$

$$G(1/\rho) = \frac{1 - \pi}{\pi}\left[1 - \frac{1}{\rho^{k+1}}\right] < 0.$$

Since $G(0) > 0$ and $G(1) < 0$, there exists a real positive root of $G(\mu)$, r, where $r < 1$. Since $G(1/\rho) < 0$ and $G(\mu) \to \infty$ as $\mu \to \infty$, there exists a real positive root of $G(\mu)$, R, where $R > 1/\rho$. The following theorem which is stated without proof may now be applied.

Pellet's Theorem:[10] *Given the polynomial*

$$f(z) = a_0 + a_1 z + \cdots + a_p z^p + \cdots + a_n z^n, \quad a_p \neq 0.$$

If the polynomial

$$F_p(z) = |a_0| + |a_1|z + \cdots + |a_{p-1}|z^{p-1} - |a_p|z^p + |a_{p+1}|z^{p+1} + \cdots |a_n|z^n$$

has two positive zeros r and R, $r < R$, then $f(z)$ has exactly p zeros in or on the circle $|z| < r$ and no zeros in the annular ring $r < |z| < R$.

Identifying p with k, n with $k + 1$ and $f(z)$ with $G(\mu)$ the rest of the proof follows.

The reader may now verify that, for large N,

$$\begin{aligned} x_N &\cong C_1 && \text{if} \quad k < \frac{\pi}{1 - \pi} \\ &\cong C_1 + C_2 N && \text{if} \quad k = \pi/1 - \pi \\ &\cong C_1 + C_2\left(\frac{1}{\theta}\right)^N && \text{if} \quad k > \frac{\pi}{1 - \pi} \end{aligned}$$

$$\begin{aligned} y_N &\cong C_1 + NC_2 && \text{if} \quad k < \frac{\pi}{1 - \pi} \\ &\cong C_1 + NC_2 + N^2 C_3 && \text{if} \quad k = \pi/1 - \pi \\ &\cong C_1 + NC_2 + C_3\left(\frac{1}{\theta}\right)^N && \text{if} \quad k > \frac{\pi}{1 - \pi} \end{aligned}$$

$$z_N \cong C_1(R)^N$$

where, R and $1/\theta$ are roots previously defined and the C's are constants. The constants may be obtained by fairly straightforward computations.

The qualitative difference between the forms of the expressions corresponding to $k < \pi/(1 - \pi)$ and $k > \pi/(1 - \pi)$ are noteworthy. This is not unexpected, since it may be recalled that in Ref. 1 it was proved in a more general context that the Markov chain associated with the infinite buffer is positive recurrent if and only if $k > \pi/(1-\pi)$.

VIII. A MONOTONICITY PROPERTY OF THE PROBABILITY OF A FINITE BUFFER BEING FULL

The steady-state probability of a buffer being full, i.e., $P(N)$ where N is the size of the buffer [see Section III and, in particular, eq. (21)] may be expected to be an important factor in the practical design of buffers. This is so not only because of the immediate implications of the definition but also because $1/P(N)$ is the average recurrence time of state N. However, this approach would appear to overlook the possibility that the probability of the buffer being full in the transient, i.e., in the approach to steady state, is seriously underestimated by $P(N)$. Such an event is not easy to rule out because, after all, $P(N)$ is an element of only one (normalized) eigenvector of the transition matrix while all the modes or eigenvectors and eigenvalues of the matrix contribute to yield $P_m(N)$ when m is finite. However, one of the implications of the result in this section is that, under certain conditions on the initial probability distribution of the contents of the buffer, $P(N)$ is indeed an upper bound on $P_m(N)$, i.e., $P_m(N) \leqq P(N)$, $m = 0, 1, \cdots$; furthermore, the important case of the buffer being initially empty satisfies the conditions just mentioned.

For a buffer of size N, the result states the following. Suppose at the mth instant the state probabilities satisfy the inequalities:

$$\pi P_m(i) - (1 - \pi) \sum_{j=i+1}^{i+k} P_m(j) \geqq 0 \quad i = 0, 1, \cdots, N - k \tag{68}$$

$$\pi P_m(i) - (1 - \pi) \sum_{j=i+1}^{N} P_m(j) \geqq 0$$
$$i = N - k + 1, N - k + 2, \cdots, N - 1. \tag{69}$$

Then (a) $P_m(N) \leqq P_{m+1}(N)$, and, as shown below, (b) the inequalities in (68) and (69) are satisfied with $P_m(l)$ replaced by $P_{m+1}(l)$ for $l = 0, 1, 2, \cdots, N$. Therefore, if (68) and (69) hold, $P_i(N) \leqq P_{i+1}(N)$ for all i, $i \geqq m$; i.e., the probability of the buffer being full is a monotonic, non-decreasing sequence. (a) may be trivially verified. The proof of (b) is as follows.

(*i*) $i = 0$.

$$\begin{aligned}
&\pi P_{m+1}(i) - (1-\pi)\{P_{m+1}(i+1) + P_{m+1}(i+2)\cdots + P_{m+1}(i+k)\}\\
&\quad = \pi(1-\pi)[P_m(0) + P_m(1) + \cdots + P_m(k)] - (1-\pi)\pi\\
&\qquad \times [P_m(0) + P_m(1) + \cdots P_m(k-1)] - (1-\pi)^2[P_m(k+1)\\
&\qquad\qquad + P_m(k+2) + \cdots + P_m(2k)]\\
&\quad = (1-\pi)[\pi P_m(k) - (1-\pi)\{P_m(k+1)\\
&\qquad\qquad + P_m(k+2)\cdots + P_m(2k)\}] \geqq 0
\end{aligned}$$

(*ii*) $1 \leqq i \leqq N - 2k$.

$$\begin{aligned}
&\pi P_{m+1}(i) - (1-\pi)\{P_{m+1}(i+1) + P_{m+1}(i+2)\cdots + P_{m+1}(i+k)\}\\
&\quad = \pi(1-\pi)[P_m(0) + P_m(1) + \cdots + P_m(k)] - (1-\pi)\pi[P_m(0)\\
&\qquad + P_m(1) + \cdots + P_m(k-1)] - (1-\pi)^2[P_m(k+1)\\
&\qquad\qquad + P_m(k+2) + \cdots + P_m(2k)]\\
&\quad = (1-\pi)[\pi P_m(k) - (1-\pi)\{P_m(k+1)\\
&\qquad\qquad + P_m(k+2) + \cdots + P_m(2k)\}] \geqq 0
\end{aligned}$$

(*iii*) $N - 2k + 1 \leqq i \leqq N - k - 1$.

$$\begin{aligned}
&\pi P_{m+1}(i) - (1-\pi)\{P_{m+1}(i+1)\\
&\qquad\qquad + P_{m+1}(i+2) + \cdots + P_{m+1}(i+k)\}\\
&\quad = \pi[P_m(i-1) + (1-\pi)P_m(i+k)] - (1-\pi)\\
&\qquad \times \pi[P_m(i) + P_m(i+1) + \cdots + P_m(i+k-1)]\\
&\qquad - (1-\pi)^2[P_m(i+k+1) + P_m(i+k+2) + \cdots + P_m(N)]\\
&\quad = \pi[\pi P_m(i-1) - (1-\pi)\{P_m(i) + P_m(i+1)\\
&\qquad + \cdots + P_m(i+k-1)\}] + (1-\pi)[\pi P_m(i+k) - (1-\pi)\\
&\qquad\quad \times \{P_m(i+k+1) + P_m(i+k+2) + \cdots + P_m(N)\}] \geqq 0
\end{aligned}$$

(*iv*) $i = N - k$.

$$\begin{aligned}
&\pi P_{m+1}(i) - (1-\pi)\{P_{m+1}(i+1)\\
&\qquad\qquad + P_{m+1}(i+2) + \cdots + P_{m+1}(i+k)\}\\
&\quad = \pi[\pi P_m(i-1) + (1-\pi)P_m(i+k)] - (1-\pi)\pi[P_m(i)\\
&\qquad\quad + P_m(i+1) + \cdots + P_m(i+k-1)] - (1-\pi)\pi P_m(N)\\
&\quad = \pi[\pi P_m(i-1) - (1-\pi)\{P_m(i)\\
&\qquad\qquad + P_m(i+1) + \cdots + P_m(i+k-1)\}] \geqq 0
\end{aligned}$$

(v) $N - k + 1 \leqq i \leqq N - 1$.

$$\begin{aligned}
\pi P_{m+1}(i) - (1-\pi)\{P_{m+1}(i+1) + P_{m+1}(i+2) + \cdots + P_{m+1}(N)\} \\
= \pi^2 P_m(i-1) - (1-\pi)\pi[P_m(i) + P_m(i+1) + \cdots + P_m(N)] \\
= \pi[\pi P_m(i-1) - (1-\pi)\{P_m(i) + P_m(i+1) \\
+ \cdots + P_m(N)\}] \geqq 0
\end{aligned}$$

(b) is proved.

Observe that $P_m(0) = 1$, $P_m(i) = 0$, $i = 1, 2, \cdots, N$ satisfies the inequalities (68) and (69). However, the other initial distribution of interest, namely, $P_m(0) = 0$, $P_m(1) = 1$, $P_m(i) = 0$, $i = 2, 3, \cdots, N$ does not satisfy the inequalities. Also, it may be verified that for the latter set of initial conditions, the monotonicity property does not hold.

It is interesting to note that if (68) and (69) hold, then $P_m(0) \geqq P_{m+1}(0)$, so that together with (b), $P_i(0) \geqq P_{i+1}(0)$ for all i, $i \geqq m$, i.e., the probability of the buffer being empty is a monotonic, non-increasing sequence.

APPENDIX A

Recursive Formula for $|D(N)|$

$$D(N) = \begin{array}{cccccccc|l}
1 & 2 & & k+1 & & & N & N+1 & \\
1 & 1 & \cdots & \cdots & \cdots & \cdots & 1 & 1 & 1 \\
-\pi & 1 & 0 & (\pi-1) & 0 & & & & 2 \\
 & \ddots & \ddots & & \ddots & \ddots & & & \\
 & & -\pi & 1 & 0 & & \cdots & (\pi-1) & N-k+1 \\
 & & & -\pi & 1 & 0 & & 0 & N-k+2 \\
 & & & & \ddots & \ddots & \ddots & & \\
 & & & & & \cdot & 1 & 0 & N \\
 & & & & & & -\pi & 1 & N+1
\end{array}$$

Expanding $|D(N)|$ along the $(N+1)$st row yields

$$|D(N)| = |D(N-1)| + \pi|X| \tag{70}$$

where

$$X \triangleq \begin{array}{c|cccccccccc|c}
 & 1 & 2 & & k+1 & & & & & N-1 & N & \\
\hline
 & 1 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1 & 1 & 1 \\
 & -\pi & 1 & 0 & (\pi - 1) & & & & & & & 2 \\
 & & \ddots & \ddots & & \ddots & & & & & & \\
 & & & -\pi & 1 & 0 & & & & (\pi - 1) & 0 & N-k-1 \\
 & & & & -\pi & 1 & 0 & & & 0 & 0 & N-k \\
 & & & & & -\pi & 1 & 0 & & 0 & (\pi - 1) & N-k+1 \\
 & & & & & & \ddots & \ddots & \ddots & & & \\
 & & & & & & & & -\pi & 1 & 0 & \\
 & & & & & & & & & -\pi & 0 & N
\end{array}$$

Expanding $|X|$ along last column:

$$|X| = (-1)^{N+1}(-\pi)^{N-1} + (\pi - 1)(-1)^{N+(N-k+1)}|Y| \qquad (71)$$

where

$$Y \triangleq \begin{array}{c|ccccccccc|c}
 & 1 & 2 & & k+1 & & & & & N-1 & \\
\hline
 & 1 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1 & 1 \\
 & -\pi & 1 & 0 & \pi - 1 & & & & & & 2 \\
 & & \ddots & \ddots & & \ddots & & & & & \\
 & & & -\pi & & 1 & 0 & & & \pi - 1 & N-k-1 \\
 & & & & & -\pi & 1 & & & 0 & N-k \\
 & & & & & & \ddots & \ddots & & & \\
 & & & & & & & -\pi & & 1 & \\
 & & & & & & & & & -\pi & N-1
\end{array}$$

Expanding $|Y|$ along the last $(k - 1)$ row.

$$|Y| = (-\pi)^{k-1}|D(N - k - 1)|. \qquad (72)$$

Combining (70) through (72):

$$|D(N)| = |D(N-1)| - (1-\pi)\pi^k |D(N-k-1)|. \tag{73}$$

APPENDIX B

Recursive Formula for $|I - \rho A(N)|$

$I - \rho A(N)$ is given in (44). Let x_n denote $|I - \rho A(N)|$. Also,

$$\lambda = -\rho(1-\pi) \quad \text{and} \quad \mu = -\rho\pi.$$

(*i*) Expand $|I - \rho A(N)|$ along the last, i.e., $(N+1)$th row, of $[I - \rho A(N)]$.

$$x_N = x_{N-1} - \mu |X| \tag{74}$$

where

$$X \triangleq \begin{array}{cccccccc|l}
1 & 2 & & k+1 & & & & N & \\
1+\lambda & \lambda & \cdots & \lambda & 0 & & & 0 & 1 \\
\mu & 1 & 0 & & \lambda & & & & 2 \\
 & \ddots & \ddots & & & \ddots & & & \\
 & & & & & & \lambda & 0 & N-k-1 \\
 & & & & & & 0 & 0 & N-k \\
 & & & & & & 0 & \lambda & N-k+1 \\
 & & & \ddots & \ddots & & & & \\
 & & & & & \mu & 1 & 0 & \\
 & & & & & & \mu & 0 & N
\end{array}$$

(*ii*) Expand $|X|$ along the last, i.e., Nth column of X.

$$\begin{aligned} |X| &= (-1)^{(N-k+1)+N}\lambda |Y| \\ &= (-1)^{k-1}\lambda |Y| \end{aligned} \tag{75}$$

where

$$Y \triangleq \begin{array}{cccccccccc|l}
1 & 2 & & k+1 & & & & & & & \\
1+\lambda & \lambda & \cdots & \lambda & & & & & & & 1 \\
\mu & 1 & & & \lambda & & & & & & 2 \\
 & \ddots & \ddots & & & \ddots & & & & & \\
 & & & & \mu & 1 & 0 & & & \lambda & N-k-1 \\
 & & & & & \mu & 1 & 0 & & 0 & N-k \\
 & & & & & 0 & \mu & 1 & 0 & 0 & N-k+1 \\
 & & & & & & & \ddots & \ddots & & \\
 & & & & & & & & \mu & 1 & \\
 & & & & & & & & & \mu & N-1
\end{array}$$

(*iii*) Expand $|Y|$ along the last $k - 1$ rows.

$$|Y| = \mu^{k-1}x_{N-k-1}. \tag{76}$$

Hence,

$$\begin{aligned} x_N &= x_{N-1} - \mu|X| && \text{from (74)} \\ &= x_{N-1} - \mu(-1)^{k-1}\lambda|Y| && \text{from (75)} \\ &= x_{N-1} - \mu(-1)^{k-1}\lambda\mu^{k-1}x_{N-k-1} && \text{from (76)} \\ &= x_{N-1} - \rho^{k+1}\pi^k(1 - \pi)x_{N-k-1}. && (77) \end{aligned}$$

REFERENCES

1. Mitra, D., and Gopinath, B., "Buffering of Data Interrupted by a Source with Priority," Proc. Fourth Asilomar Conf. Circuits Syst., 1970.
2. Sherman, D. N., "Data Buffer Occupancy Statistics for Asynchronous Multiplexing of Data in Speech," Proc. Intl. Conf. Commun., IEEE, San Francisco, 1970.
3. Brady, P. T., "A Technique for Investigating On-Off Patterns of Speech," B.S.T.J., *44*, No. 1 (January 1965), pp. 1–22.
4. Brady, P. T., "A Model for Generating On-Off Speech Patterns in Two-Way Conversations," B. S. T. J., *48*, No. 7 (September 1969), pp. 2445–2472.
5. Limb, J. O., "Buffering of Data Generated by the Coding of Moving Images," B.S.T.J., *51*, No. 1 (January 1972), pp. 239–259.
6. Limb, J. O., private communication.
7. Haskell, B., private communication.
8. Karlin, S., *A First Course in Stochastic Processes*, New York: Academic Press, 1966.
9. Taussky, O., "A Recurring Theorem on Determinants," Amer. Math. Monthly, *56*, 1949, pp. 672–676.
10. Marden, M., "Geometry of Polynomials," Mathematical Surveys, *3*, Amer. Math. Soc., Providence, Rhode Island, 1966.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

A Frame-to-Frame *Picturephone*® Coder For Signals Containing Differential Quantizing Noise

By D. J. CONNOR, B. G. HASKELL, and F. W. MOUNTS

(Manuscript received July 18, 1972)

The frame-to-frame coder described in Ref. 1 used an 8-bit PCM signal for input. If, instead, the signal is obtained by digitally integrating the output of an element difference coder, the quantization noise may be misinterpreted as motion, and cause unnecessary transmission. In the particular example of the Phase I coder,[2] *the quantization noise loads the frame codec to the extent that it produces an unacceptable picture.*

In this paper, a frame-to-frame coder for Picturephone® *signals is described which is capable of coding the digital output of a Phase I codec for transmission over a 2-megabit/second channel. Improved methods are used to segment the noisy picture into moving areas and background areas. The moving areas are then transmitted using a number of data reduction techniques. During periods of slow movement, clusters of frame-to-frame differences in the moving area are transmitted. For moderate movement, frame differences are sent only in every other field, the moving areas of intervening fields being transmitted by a conditional field interpolation technique. For rapid movement, 2:1 horizontal subsampling is used, and, finally, during violent motion when the buffer fills, frame repeating is used.*

The picture quality obtained from a laboratory simulation of this system is believed to be satisfactory even for a very active subject. With small amounts of motion the subjective quality is actually improved because the visibility of the quantizing noise from the Phase I codec is reduced by the inherent frame repeating action of the coder.

I. INTRODUCTION AND SUMMARY

In Ref. 1 an 8-bit-per-picture element (pel) *Picturephone*-type signal is coded using only 2 megabits/second (Mb/s). Clusters of significant frame differences are transmitted using a double-length code (four-bit and six-bit) for the frame differences and eight-bit addresses for the clusters. During periods of moderate movement, every other significant frame difference along the line is transmitted, the intervening elements being obtained by linear interpolation. If, during violent motion, the buffer fills, frame repeating is used.

In this system a frame difference was deemed significant if its magnitude exceeded some threshold value ($T = 4, 5, 6, 7$) which depended on the buffer fullness. Two exceptions to this criterion were made, however: (*i*) if a significant change was surrounded on both sides by two insignificant changes, then the change was deemed insignificant, and (*ii*) if two clusters of significant changes were separated by three or less insignificant changes, then the clusters were joined by relabeling the intervening changes as significant.

For maximum flexibility of the *Picturephone* transmission system, it is desirable that an interframe coder be able to accept as an input a signal that has previously been coded by an intraframe coder, such as an element difference coder. Such a signal will have a significantly higher level of quantization noise than an 8-bit PCM signal. The Phase I codec[2] is an example of an element difference coder. Since the quantization noise from this coder has been carefully shaped for minimum visibility, the signal it produces probably contains the highest noise level of any signal likely to be encountered by an interframe coder. Designing an interframe coder to work with such a signal thus reveals many of the problems involved in working with signals having realistic noise levels.

If the input signal contains element differential quantizing noise, the system in Ref. 1 does not perform well at all. An inordinate number of sizable frame-to-frame differences arise due to the quantizing noise, and in the case of the Phase I codec, acceptable video transmission at 2Mb/s is impossible. Raising the threshold of significance reduces the number of background frame differences which are transmitted, but it also reduces the number of subjectively important frame differences in the moving area which are sent. Unacceptable picture quality results.

Averaged over a small region in space and time, the frame differences due to quantizing noise differ in many ways from the subjectively important frame differences due to movement. For example, frame differences due to movement are correlated spatially, whereas frame differences due to quantizing noise are not.

These properties have been exploited to give a method for segmenting the picture into moving areas and stationary areas.[3] The moving area as defined by the segmenter tends to be slightly larger than the actual moving area, but it has been found that this is necessary if a subjectively acceptable picture is to be obtained.

The number of picture elements which must be transmitted using the noisy input and this segmenter is much larger than with the 8-bit input and the segmenting criterion of Ref. 1. Thus, even with a good segmenter, the data rate is larger than 2 Mb/s using only the data reduction techniques of Ref. 1. Other means of data compression are required if a 2-Mb/s rate is to be obtained.

Two techniques are proposed. First, since the segmenting criterion used here requires that all picture elements in the moving area be transmitted, a large number of zero frame differences are sent, i.e., the average transmitted frame difference is much smaller than in Ref. 1. Under these circumstances, variable word length codes can be used to good advantage. Using a variable word length code optimized for moderate motion, only about two bits per frame difference are required on the average. Using this same code during periods of active motion requires about three bits per frame difference on the average.

Using the new segmenter and variable word length coding of frame differences, transmission below 2 Mb/s is easily accomplished during periods of slow movement. When motion becomes a little more rapid, however, the 2-Mb/s rate is surpassed, and another data compression technique must be used. Two-to-one horizontal subsampling generally results in subjectively unacceptable picture quality because the movement is too slow to hide the resolution loss. Thus, a conditional field interpolation technique[4,5] is used as the second method of data rate reduction.

With this technique, frame differences in the moving area are transmitted only during every other field. Each pel in the moving area of the intervening fields is obtained at the receiver by a four-way average of vertically adjacent picture elements in the two fields adjacent to the one being coded. However, if the four-way average is in error by an amount larger than some prescribed threshold, then a quantized correction value must be sent to maintain acceptable picture quality.[4]

The receiver as described above would still have to be told which picture elements in the intervening field are in the moving area, and which are in the background. However, since movement is so highly correlated from field to field, we believe that this information can be extracted from the two fields adjacent to the one being interpolated.

With rapid motion, 2:1 horizontal subsampling can be employed. This is brought in under buffer control. When motion becomes violent and the buffer fills, then transmission ceases for one frame period and the previous frame is repeated.

Using the data compression techniques described above, a laboratory simulation was constructed to test the important aspects of a 2-Mb/s frame-to-frame codec that is capable of coding the digital output of a Phase I codec. A simplified block diagram of the simulation is shown in Fig. 1. A digital signal identical to the output of the frame-to-frame codec was passed through another digital Phase I codec without degrading the picture noticeably. The system described is capable of accommodating about the same amount of movement as that in Ref. 1, with a picture quality comparable to that of the Phase I codec.

The Phase I codec was designed, of course, without any thought of frame-to-frame coding. It is not surprising, therefore, that many difficulties arise when frame-to-frame coding techniques are applied to the output of a Phase I codec. Changes in the Phase I coder to reduce the quantization noise would not only result in a simpler interframe coder, but could also lead to a data rate less than the 2 Mb/s obtained here. How much less will have to await further study.

The next four sections describe in more detail the operation of the frame-to-frame coder. The last section describes the simulation.

II. SEGMENTING THE PICTURE INTO "MOVING" AND "STATIONARY" AREAS

An essential preliminary to the development of the coder described in this paper was the development of methods for detecting or segmenting the moving area in a video signal which has already been corrupted by noise due to an in-frame coding operation. A full description of the work done on this problem will be given in subsequent papers. In this

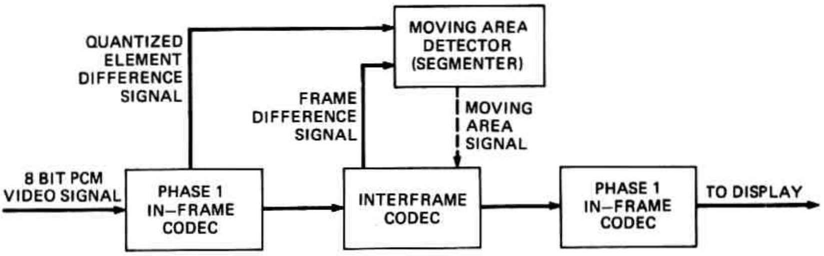


Fig. 1—A simplified block diagram of the simulation showing the signals used and produced by the segmenter.

section, we will simply state the various properties of the video signal and the coding noise which can be exploited in detecting the moving area. Following that, we give a description of the actual segmenter that was developed for use in the system described in this paper.

In order to separate the frame-to-frame brightness changes caused by movement from those caused by noise from an element difference quantizer, advantage can be taken of certain distinguishing properties. The most important property of the movement-generated frame differences is that they are spatially correlated. Two properties of the noise are important:

(*i*) It is almost entirely uncorrelated spatially;
(*ii*) The magnitudes of individual noise spikes are equal to the spacing of the representative levels used in the element difference quantizer.

The second property of the noise results from the fact that in stationary areas a small noise perturbation from one frame to the next can cause a change in the representative level used to encode a particular element difference. This change will be to an adjacent representative level in the quantizing scale, and, consequently, the resultant frame difference will be equal to the spacing of those levels. The more widely spaced outer levels of the companded quantizing scale are used to encode detailed areas and contrasty vertical edges. Thus, the frame difference noise is greatest in these regions.

Finally, a useful property of moving areas is that they are spatially and temporally contiguous. In other words, if a pel is in the moving area, it is highly probable that the spatially adjacent pels and the same pel in the next frame are in the moving area.

The signals employed by the segmenter in detecting the moving area are indicated in Fig. 1. A block diagram of the processing of the quantized element difference signal and the frame difference signal is given in Fig. 2. The frame difference signal undergoes two separate spatial filtering operations which increase the signal-to-noise ratio for the spatially correlated frame differences caused by movement. Filter A is designed to enhance the frame difference signal associated with moving edges and particularly with vertical edges moving horizontally. This signal is characterized by high horizontal spatial frequencies and lower vertical spatial frequencies. By averaging the frame difference signal from adjacent lines, these low vertical frequencies are enhanced relative to the spectrally flat frame difference noise.

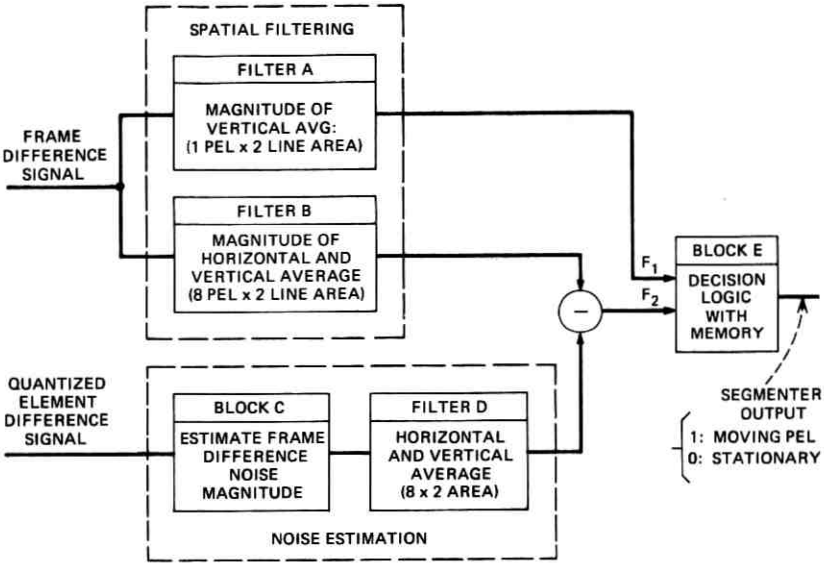

SPATIAL FILTERING
FILTER A
MAGNITUDE OF VERTICAL AVG: (1 PEL x 2 LINE AREA)
FRAME DIFFERENCE SIGNAL
FILTER B
MAGNITUDE OF HORIZONTAL AND VERTICAL AVERAGE (8 PEL x 2 LINE AREA)
F_1
F_2
−
BLOCK E
DECISION LOGIC WITH MEMORY
QUANTIZED ELEMENT DIFFERENCE SIGNAL
BLOCK C
ESTIMATE FRAME DIFFERENCE NOISE MAGNITUDE
FILTER D
HORIZONTAL AND VERTICAL AVERAGE (8 x 2 AREA)
NOISE ESTIMATION
SEGMENTER OUTPUT
1: MOVING PEL
0: STATIONARY

Fig. 2—A simplified block diagram of the segmenter showing the spatial filtering and noise estimation processes.

Filter B is designed to enhance the frame difference signal associated with the movement of relatively flat areas. This signal has most of its energy at low spatial frequencies. By averaging the frame difference signal in an 8-pel-by-2-line area, an increased S/N ratio is obtained. After the averaging operation, the signals from both filters are rectified since the frame difference signal can be of either sign.

Although filter B enhances the movement-generated frame differences in relatively flat areas, it is found that in highly detailed, stationary areas its output commonly exceeds the output arising in slowly moving, flat areas, such as hair. Thus, simple threshold detection is no good. However, it is possible to compensate the output of filter B for these detail-dependent variations in the frame difference noise by subtracting a filtered estimate of the magnitude of the noise signal.

As mentioned above, individual frame differences caused by quantizing noise are equal to the spacing between representative levels of the element-difference quantizing scale. Thus, in blocks C and D in Fig. 2, the filtered estimate of the noise signal is derived from the quantized element-difference signal by generating at the output of block C a non-negative signal that is proportional to the spacing between the input representative level and the adjacent smaller level in the element-difference quantizing scale. (Because the probability distribution of element differences is monotonic and peaks at zero, the most probable transition between representative levels due to a noise perturbation is from an outer level to the adjacent smaller level.) The estimated frame-difference, noise-magnitude assignment for the representative levels is modified for the four inner levels of the 16-level quantizing scale of the Phase I codec as shown in Table I, which gives the output versus input for block C. This modification reflects the fact that noise frame-differences are relatively small in flat areas of the picture. Experimentally it was found that flat, stationary areas could be more easily distinguished from flat, moving areas if no noise compensation was used in these regions. Thus, the estimated frame-difference noise magnitude for the four inner levels is set to zero.

The filtered and noise-compensated frame-difference signals serve as inputs to the decision logic of block E. This logic takes advantage of the fact that moving areas tend to be contiguous both spatially and temporally. Thus, if movement is occurring at a particular pel, there is a high probability that movement is occurring at pels that are spatially and temporally adjacent. Consequently, the philosophy for the design of the decision logic was to use a high decision threshold for the detection of movement in regions of the picture which were previously stationary, and a lower threshold in regions where movement had recently been detected.

A block diagram of the decision logic is given in Fig. 3. (For simplicity, a number of delays required to keep the binary signals in register have not been shown.) The filtered and noise-compensated frame-difference signals serve as inputs to this logic. They are first converted to binary signals by threshold operations having the following transfer characteristics,

$$B_i = 1 \quad \text{if} \quad F \geqq T_i$$

$$B_i = 0 \quad \text{if} \quad F < T_i$$

where F is the input, T_i the threshold, and B_i the corresponding binary output signal. A control signal from the interframe coder that indicates the amount of movement by measuring the buffer fullness is used to raise the thresholds T_1 and T_3 during periods of fast motion.[1] Movement detection is easy in this situation, and the segmenting accuracy can be increased.

In order to best describe the operation of the decision logic, we will start with the block labeled "Binary Threshold Logic with Hysteresis." This block will be referred to as an N out of M (N/M) device after

Table I—Representative Levels of Phase I Codec Quantizer and Corresponding Estimates of Frame Difference Noise Magnitude

Quantized Element Difference	Estimate of Frame Difference Noise Magnitude
±2/256	0/256
±6	0
±14	4
±30	8
±46	8
±62	8
±78	8
±94	8

Limb and Pease.[6] A block diagram of this device is given in Fig. 4. The accumulator in the N/M device keeps a count of the number of ones in the 8-by-3 block of 24 pels adjacent to the pel of interest as shown in Fig. 5. (Thus, M is 24.) If the output of the accumulator is greater than or equal to the threshold $N_1 = 9$, the output flip-flop is set; and segmenter output function B_5 becomes a one to indicate moving area. In keeping with the design philosophy mentioned above, the flip-flop can only be reset by having the output of the accumulator drop below the lower threshold $N_2 = 4$. Note that by setting N_1 equal to nine, the signal B_3, which indicates the occurrence of flat area movement on the present line can never by itself cause the flip-flop to

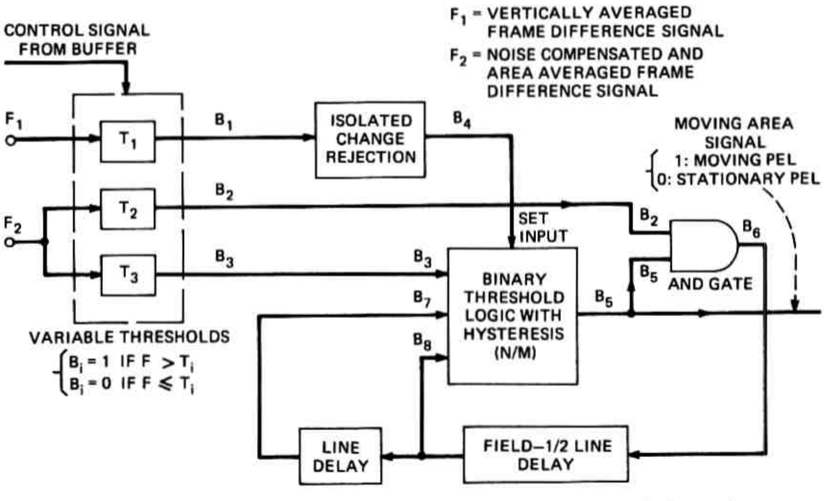

CONTROL SIGNAL FROM BUFFER
F1 = VERTICALLY AVERAGED FRAME DIFFERENCE SIGNAL
F2 = NOISE COMPENSATED AND AREA AVERAGED FRAME DIFFERENCE SIGNAL
F1
F2
T1
T2
T3
B1
B2
B3
B4
ISOLATED CHANGE REJECTION
MOVING AREA SIGNAL
1: MOVING PEL
0: STATIONARY PEL
SET INPUT
B2
B6
B5
AND GATE
B3
B7
B8
BINARY THRESHOLD LOGIC WITH HYSTERESIS (N/M)
B5
VARIABLE THRESHOLDS
Bi = 1 IF F > Ti
Bi = 0 IF F ≤ Ti
LINE DELAY
FIELD–1/2 LINE DELAY

Fig. 3—Decision logic. The N/M device processes binary signals from the present and previous fields to produce the moving area signal.

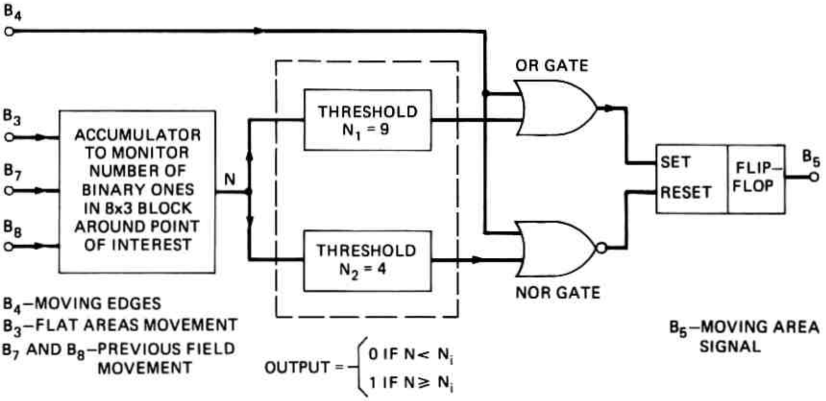

B₄
B₃
B₇
B₈
ACCUMULATOR TO MONITOR NUMBER OF BINARY ONES IN 8x3 BLOCK AROUND POINT OF INTEREST
N
THRESHOLD N₁ = 9
THRESHOLD N₂ = 4
OR GATE
NOR GATE
SET
RESET
FLIP-FLOP
B₅
B₄–MOVING EDGES
B₃–FLAT AREAS MOVEMENT
B₇ AND B₈–PREVIOUS FIELD MOVEMENT
OUTPUT = 0 IF N< Nᵢ ; 1 IF N≥ Nᵢ
B₅–MOVING AREA SIGNAL

Fig. 4—*N*/*M* device. Two thresholds are applied to the output of an accumulator that keeps a count of the number of binary ones in an 8-pel-by-3-line area around the point of interest. These thresholds control the state of the output flip-flop along with the moving edge signal B_4.

be set. Initially, the only way the flip–flop can be set is for a one to occur in the signal B_4. Since this function indicates movement of edges, edge movement must be detected before flat area movement. However, once edge movement is detected, the flip-flop is set and the lower threshold N_2 determines whether adjacent pels on the same line will be designated as moving. In addition, referring to Fig. 3, if B_2, which is a more sensitive but noisier indicator of flat area movement than B_3, is a one when the flip-flop is set, B_6 will be a one. Hence, in keeping with the design philosophy, the value of N_1 for the spatially and temporally adjacent pels in the next field will be effectively lowered by the appearance of these ones in B_7 and B_6. As a result of the interactions described above, the *N*/*M* device tends to fill in moving areas, and to designate areas as moving for a short while after they become stationary.

Given the above description of the *N*/*M* device, the functions and choice of design variables for the various other blocks in Fig. 3 become

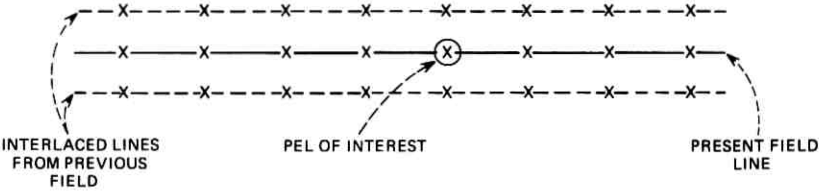


Fig. 5—Arrangement of the 8-by-3 block of pels monitored by the *N*/*M* device.

evident. The threshold T_1 is set relatively high (~10 on an 8-bit PCM scale of 256 levels) to insure that a binary one in the function B_1 is indeed caused by the movement of an edge. This function undergoes further processing so that isolated ones (no other ones within two pels horizontally in either direction) arising from noise spikes are set to zero.[1] Similarly, the threshold T_3 is set relatively high (~4/256) to insure that the condition $B_3 = 1$ corresponds to movement in flat areas. The threshold T_2 on the other hand can be set lower (~2/256), since it causes ones to occur in B_6 only if the segmenter output, B_5, is a one. However, it eliminates from B_6 most of the "fill-in" pels generated by the N/M device. This process stabilizes the feedback loop around this device.

If the thresholds T_1 to T_3 are fixed, they must be set quite low in order to detect very slow motion. Given the level of quantization noise from a Phase I coder, such low thresholds inevitably lead to the inclusion of some background points in the moving area. By using the control signal from the buffer, the thresholds can be made speed dependent. For even moderate motion, the segmenting is then virtually ideal.

III. VARIABLE WORD-LENGTH CODING OF FRAME DIFFERENCES

In Ref. 1, the 9-bit frame differences ($-255 \cdots 0 \cdots +255$) were quantized into 64 levels. Since the Phase I coder gives an effective 6-bit signal (6 bits with the seventh bit alternately 0 and 1 along the line), only frame differences that are multiples of 4/256 can occur. This set of frame differences is sufficiently coarsely quantized for efficient transmission.

Also, in Ref. 1 it was very much easier to separate the subjectively important frame differences from those few due to camera and system noise. In the system described here, where a Phase I signal is used as an input, once the moving area has been identified, all frame differences in it must be transmitted since it is not possible to tell which are due to movement and which are due to quantizing noise. Within the moving area, as defined by the segmenter, many zero frame differences do occur. However, since they are randomly interspersed among the nonzero frame differences, it is much more efficient to transmit them than it would be to delete them and address the remaining nonzero frame differences.[1]

This causes the average magnitude of transmitted frame differences to be considerably smaller than in Ref. 1 where an 8-bit input is used.

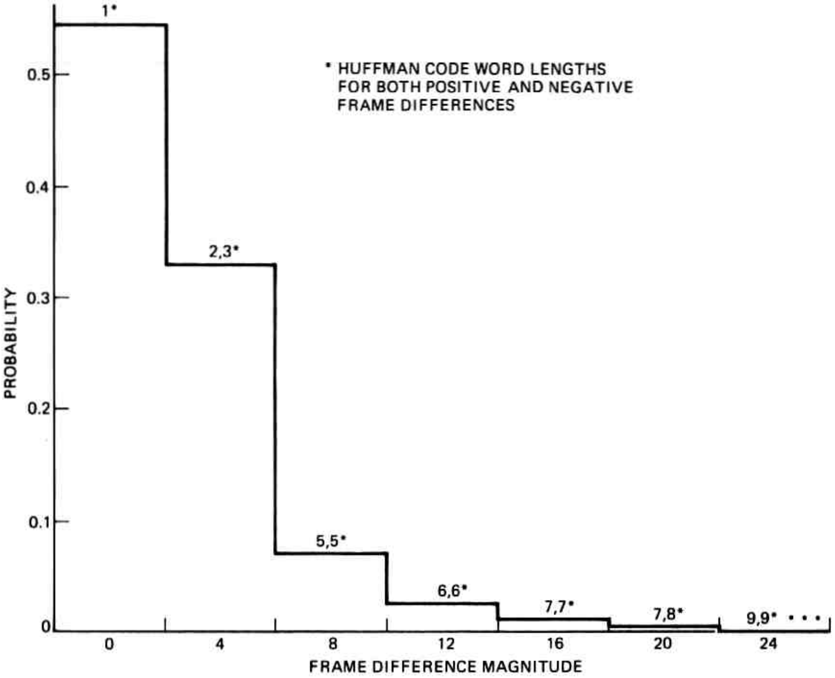


Fig. 6—Typical histogram of moving area frame differences during moderate motion. Huffman code word lengths are shown for each level.

Thus, a more complex variable word-length code can be used to good advantage in reducing the average number of bits required to transmit a frame difference. Preliminary measurements indicate that with a good variable word-length code, less than two bits per frame difference are required on the average during periods of slow movement. During moderate movement, a little more than two bits per frame difference are required; and during rapid movement, about three bits are needed.

Figure 6 shows a typical histogram of the magnitude of the frame differences in the moving area during moderate motion. Also shown are the Huffman code word lengths corresponding to this distribution. The average word length per frame difference is 2.05 bits.

IV. CONDITIONAL FIELD INTERPOLATION

During very low-speed movement, variable word-length coding of frame differences in the moving area is sufficient to code at a rate below 2 Mb/s. Unfortunately, the speed at which the bit rate rises

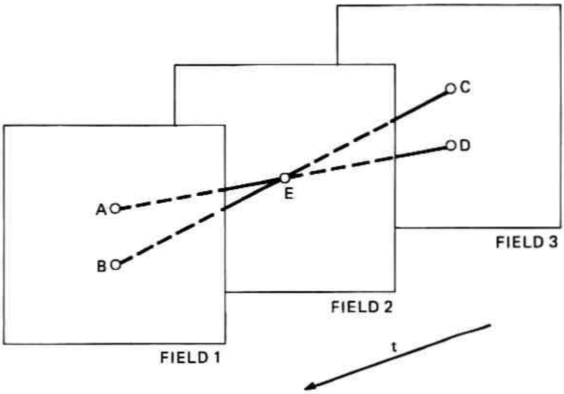


Fig. 7.—Four-way vertical averaging. Fields 1 and 3 are sent via frame differences in the moving area. Information about moving area pels (E) in field 2 is sent only if the interpolation error $|E-(A+B+C+D)/4|$ exceeds a threshold.

above 2 Mb/s is still too slow to hide the resolution loss incurred by 2:1 horizontal subsampling. Thus, another data compression technique is used.

With conditional field interpolation (called conditional vertical subsampling in Ref. 4) only every other field is transmitted by sending frame differences in the moving area. The moving area pels in the intervening fields are obtained from a 4-way average of vertically adjacent pels in the two adjacent fields. In Fig. 7, fields 1 and 3 have been transmitted via frame differences in the moving area, and pel E is to be sent via conditional field interpolation. Pels A and C are directly above E, and pels B and D are directly below E. The 4-way average $(A + B + C + D)/4$ is computed and used as a prediction of E. If the interpolation error does not exceed some prescribed threshold value, then nothing is sent, and the 4-way average is used in place of E. If the interpolation error does exceed the threshold, then a quantized correction value is transmitted.

Since the receiver treats background area in the interpolated fields differently than it does moving area, it must be told which picture elements are in the moving area and vice-versa. Preliminary measurements indicate that addresses for the moving area of the interpolated fields could probably be transmitted using less than 0.1 Mb/s. Alternatively, the moving area of the interpolated fields might be satisfactorily obtained from the union of the moving areas in the two adjacent uninterpolated fields. This would not require any additional information to be transmitted.

In order to determine whether or not the field interpolation error was acceptable, threshold values between 7 and 15 out of 255 were used. These values gave acceptable to marginally acceptable picture quality, and a data rate which was drastically reduced compared with sending frame differences.

V. BLOCK DIAGRAM

Figure 8 shows a block diagram of the system. (The segmenting operation is shown in detail in Figs. 2 to 4.) During very slow movement, every field is transmitted by sending frame differences (B′ − D) in

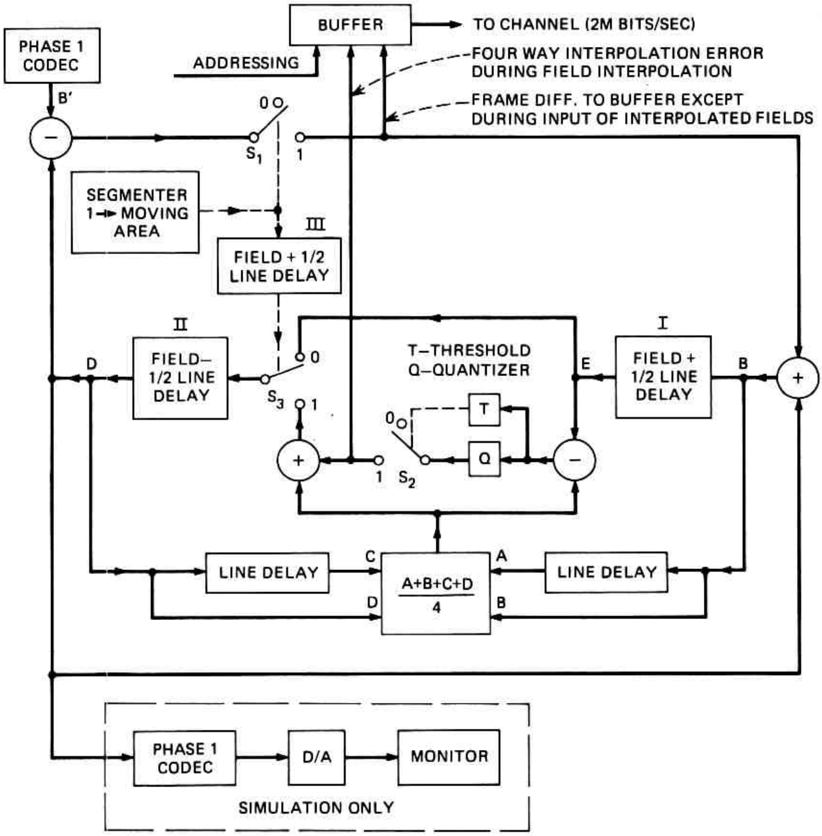


FIG. 8.—Frame-to-frame coder for *Picturephone*® signals with Phase I quantizing noise. During field interpolation, information from two fields is fed to the buffer simultaneously.

the moving area as defined by the segmenter. S_3 is in the 0 position to give an uninterrupted frame memory, and S_2 is in the 0 position so that no interpolation error information reaches the buffer. S_1 is controlled by the segmenter. When in the 0 position, the previous frame value D (see Fig. 7) is fed to delay I, and no frame difference is fed to the buffer. When in the 1 position, the new pel $B' = D + (B' - D)$ is fed to the delay, and a frame difference is fed to the buffer for coding, addressing, and transmission.

When movement becomes more rapid and the buffer fills beyond some prescribed threshold, only every other field is sent via frame differences in the moving area as outlined above. Mode switching occurs only at the end of a field. During input of a field which is to be interpolated, S_2 and S_3 are in the 0 positions allowing uninterpolated fields to enter delay II unchanged. S_1 is controlled by the segmenter as before; however, *no frame differences are fed to the buffer for transmission.* Coding and transmission of this field takes place at a later time. Thus, during input of interpolated fields no amplitude information is fed to the buffer. Addressing information needed to specify the moving area at the receiver could be sent at this time if it is found to be more efficient; however, this information could just as well be obtained from the output of delay III and sent later during the actual coding and transmission of the interpolated fields.

During input of uninterpolated fields, coding and transmission of frame differences in the moving area are carried out as usual by means of switch S_1. However, at the same time, coding and transmission of interpolated fields are also performed. When pel E in an interpolated field (see Fig. 7) emerges from delay I, pels A, B, C, and D are emerging from their respective delays as shown in Fig. 8. The output of delay III identifies E as either a background or a moving area pel.

If E is a background pel, S_2 and S_3 are switched to the 0 positions. E enters delay II and no information is fed to the buffer. If E is a moving area pel, then S_3 is switched to position 1, and S_2 is controlled by the threshold logic T. The threshold logic compares the magnitude of the interpolation error $[E - (A + B + C + D)/4]$ with a prescribed threshold. If the error is smaller than the threshold value T, then S_2 is opened (0 position), nothing is fed to the buffer for transmission, and the 4-way average enters delay II in place of E. If the interpolation error is too large, S_2 is closed (1 position), a quantized interpolation error generated by the quantizer Q is fed to the buffer for transmission, and the corrected interpolation value is fed to delay II in place of E.

A number of implementation aspects have not been discussed.

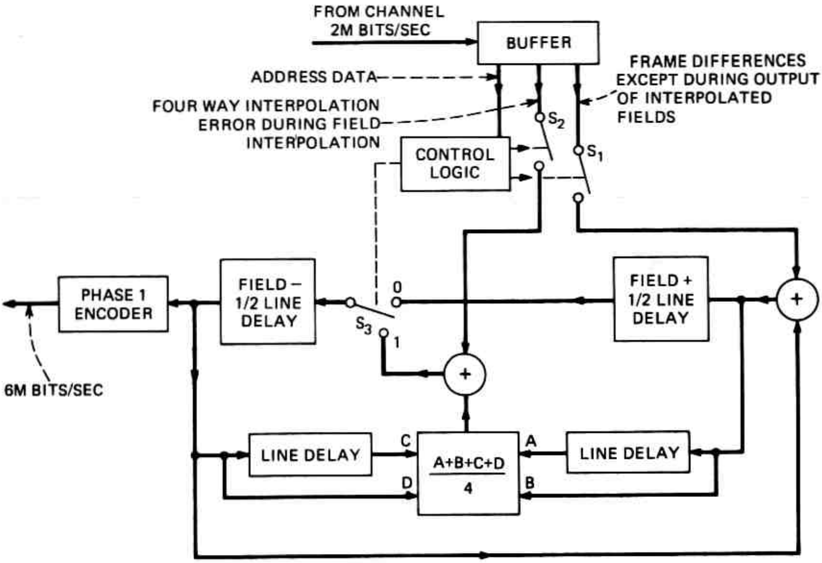


Fig. 9—Receiver configuration. Information is read from the buffer two fields at a time during field interpolation.

For example:

(*i*) All moving area picture element information fed to the buffer must, of course, be accompanied by addressing information, and efficient addressing may require that some of the switch control functions be modified, e.g., isolated point rejection, gap bridging (see Ref. 1).

(*ii*) During field interpolation, information from two fields is fed to the buffer simultaneously. Thus, some multiplexing arrangement must be devised in order to implement the system as described. For example, a buffer might be provided for each field and the outputs switched.

(*iii*) The receiver configuration is very similar to that of the transmitter (see Fig. 9).

(*iv*) Two-to-one horizontal subsampling, and frame repeating have not been discussed here since they are covered elsewhere.[1,6]

VI. SIMULATION OF THE SYSTEM

A number of short cuts were taken to simulate the system described above. First, no coding, buffering or transmission of the data was undertaken. In the simulation, only the picture processing performed at the transmitter was undertaken. The picture which would have appeared at a receiver in the absence of transmission errors was equivalent to the output of field memory II in Fig. 8. As in Ref. 1, buffer control of the picture processing was obtained by using an analog integrator to keep track of the number of bits that would have been in a real buffer had one been built. Also, as in Ref. 1, a buffer size of 67,000 bits was chosen so that it would completely empty if the input of data were stopped for one frame period.

Second, the effect of the variable word-length coding was only partially simulated. Recall from Section III that with a good variable word-length code, pels in the moving area could be transmitted using, on the average, less than two bits per frame difference during periods of slow movement, approximately two bits during moderate movement, and about three bits during rapid movement. This was simulated by counting two bits per frame difference during periods of slow and moderate movement and four bits during rapid movement when 2:1 horizontal subsampling was employed.

During conditional field interpolation, the same bit assignment scheme was used to account for the transmission of interpolation errors. Although transmitted interpolation errors were not quantized in the simulation, preliminary results indicate that they can be quantized quite coarsely. Thus, a 2-bit, 4-bit assignment is not unreasonable.

Transmission of moving area addresses for the interpolated fields was not simulated. Preliminary measurements indicate that with rapid motion, the number of clusters requiring addressing is, on the average, about two per line. If 16 bits are used to address each cluster, then about 0.1 Mb/s would be required to transmit them. If, as was conjectured in Section IV, this moving area can be obtained adequately from the uninterpolated fields, then no extra information need be transmitted.

Finally, transmitted information from interpolated fields was delayed by a field period before being fed to the buffer simulator purely for reasons of expedience. This means that during most of conditional field interpolation, information from two fields does not enter the buffer simulator at the same time as is described in Section V. This should not affect the results very much since much less data is generated during interpolated fields than during uninterpolated fields. However, frame repeating due to buffer filling may occur slightly more often in the actual system than it did in the simulation if the same buffer size is used.

The acceptability of the pictures obtained using the simulation described above was determined mainly by comparison with pictures from the Phase I codec alone. This codec gives pictures that have moderate amounts of both granular noise and edge busyness throughout the picture. The frame-to-frame codec described above transmits information only about the moving area. Consequently, the Phase I codec noise in the background becomes stationary and, hence, much less noticeable. In this sense, the pictures are improved.

Some loss of quality is caused, however, by the use of subsampling. Under some conditions, a slight jerkiness in the movement being depicted is noticeable as the codec enters the vertical subsampling mode. Also, for very high-speed movement, a slight checkered pattern at contrasty edges is detectable. This is caused by the use of both horizontal and vertical subsampling.

On an overall basis, the picture quality produced by this 2-Mb/s codec is felt to be equal to the quality of the input Phase I codec signal.

VII. ACKNOWLEDGMENTS

Many helpful discussions were held with J. C. Candy, J. O. Limb, and R. F. W. Pease. We also acknowledge the technical support of Mrs. M. A. Franke and W. G. Scholes.

REFERENCES

1. Candy, J. C., Franke, Mrs. M. A., Haskell, B. G., and Mounts, F. W., "Transmitting Television as Clusters of Frame-to-Frame Differences," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1889–1917.
2. Abbott, R. P., "A Differential Pulse-Code-Modulation Coder for Videotelephony Using Four Bits Per Sample," IEEE Trans. Commun. Tech., *COM-19*, No. 6 (December 1971), pp. 907–912.
3. Connor, D. J., Limb, J. O., Pease, R. F. W., and Scholes, W. G., "Segmenting Noisy Television Pictures into Moving and Stationary Areas," unpublished work.
4. Pease, R. F. W., "Conditional Vertical Subsampling—A Technique to Assist in the Coding of Television Signals," B.S.T.J., *51*, No. 4 (April 1972), pp. 787–802.
5. Limb, J. O., and Pease, R. F. W., "A Simple Interframe Coder for Video Telephony," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1877–1888.
6. Pease, R. F. W., and Limb, J. O., "Exchange of Spatial and Temporal Resolution in Television Coding," B.S.T.J., *50*, No. 1 (January 1971), pp. 191–200.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

A DC-to-2.3-GHz Amplifier Using an "Embedding" Scheme

By GERARD WHITE and GEN M. CHIN

(Manuscript received July 26, 1972)

A novel circuit technique is described for embedding a high-frequency amplifier in a low-frequency circuit to achieve a defined, flat gain from dc to the cutoff frequency of the hf amplifier. The technique provides this low-frequency gain without compromising the hf design optimization. An embodiment of this technique is described which has provided, experimentally, amplifier gain from dc to a half-power point of 2.3 GHz.

I. INTRODUCTION

This paper describes a novel circuit technique for providing broadband amplifier gain response from dc to extremely high frequencies. The technique provides this wide spectrum response without compromising either the hf response or the dc stability. This performance is obtained by "embedding" a parameter-optimized hf amplifier within a dc gain-defining circuit, and providing means for ensuring a smooth transition between the low-frequency to high-frequency operating modes. This technique avoids the compromise of hf performance which is frequently present in direct coupled amplifiers.[1,2]

Amplifiers with response to dc are frequently required in communication systems employing a baseband Pulse Code Modulation (PCM) type encoding scheme where the entropy of the information signal is unknown. Because of the simplicity afforded by binary PCM, its use has been adopted in many optical systems.[3,4] In such cases, the channel information rate is restricted mainly by the bandwidth of the electronic driving circuits. The economics of noncoherent optical systems again dictate the use of extremely broadband amplifiers and, if the system simplicity is not to suffer,[5] a gain response to dc is required. Such systems should benefit from an hf optimized amplifier providing gain to dc. A particular realization of the embedding technique is described in this paper which may find use in such systems. This realization provides essentially flat gain response up to approximately 0.5 of the constituent device common base cutoff frequencies. The technique results in an amplifier voltage gain of 8 dB over the range dc to 2.3 GHz with a step response rise time of 200 ps. At frequencies below the pre-cutoff resonance, inband ripple is typically less than 1.5 dB.

II. THE EMBEDDING TECHNIQUE

Amplifier embedding is applicable to a number of circuit realizations but is best described in terms of the simple common emitter stage shown in Fig. 1. At dc, this stage exhibits a gain of

$$G_{\mathrm{dc}} \cong \frac{R_{C_1} + R_{C_2}}{R_{E_1} + R_{E_2}},$$

and at high frequencies the gain is simply $G_{\mathrm{hf}} = R_{C_1}/R_{E_1}$, so that, by making the equality

$$\frac{R_{C_1} + R_{C_2}}{R_{E_1} + R_{E_2}} = R_{C_1}/R_{E_2},$$

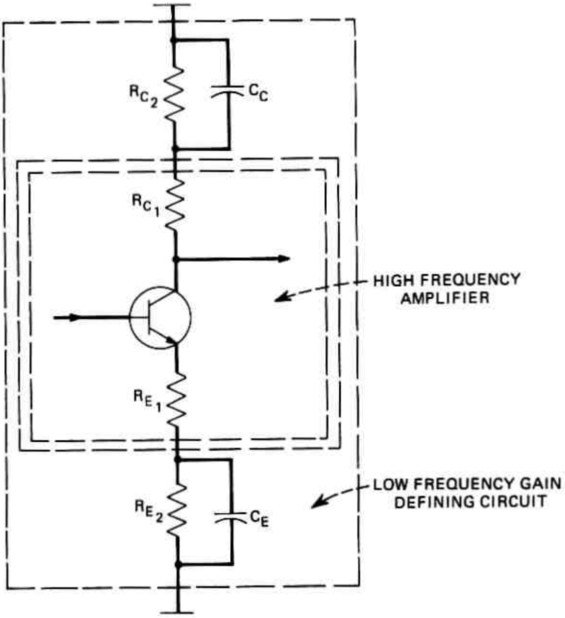


Fig. 1—Embedded common emitter stage.

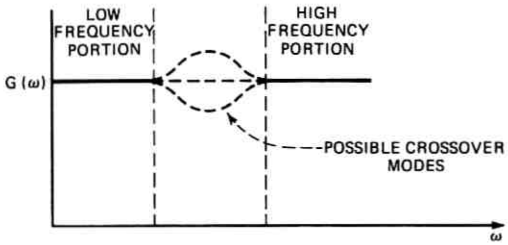


Fig. 2—Amplifier gain spectrum.

the two portions of the gain spectrum, shown in Fig. 2, will be equal. In addition to this requirement, a flat transitional crossover is required. This critical crossover point has always been the major difficulty in split-band additive amplifiers.[6] To analyze the requirements for a flat crossover, it is more meaningful to consider the practical circuit arrangement shown in Fig. 3. The circuit also incorporates emitter-follower buffering stages at the input and output. The additional

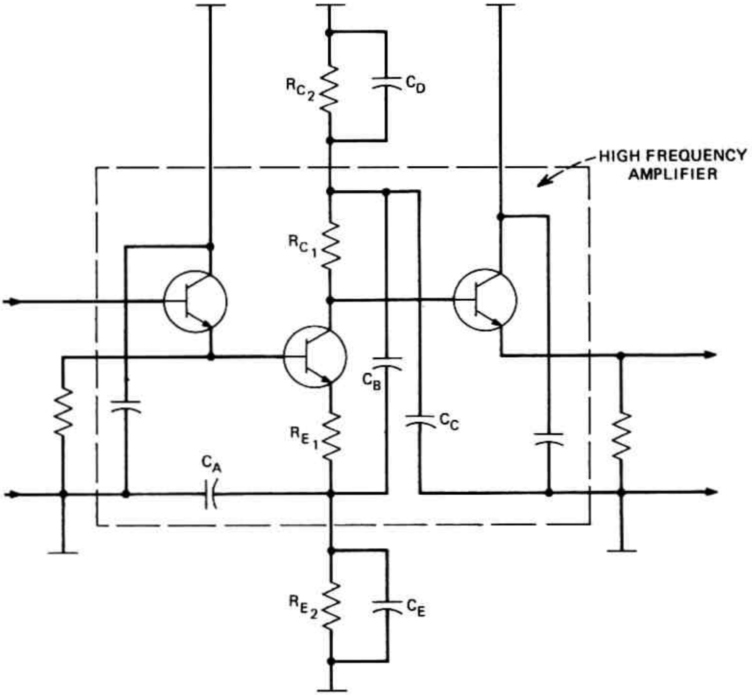


Fig. 3—Practical hf amplifier embedded in a low-frequency gain-defining circuit.

capacitances are for minimization of the lengths of the constituent current loops in the circuit for good high-frequency gain performance and the improvement of stability by effecting a reduction in the points of interaction of these loops. The condition for a flat gain spectrum can be analyzed using the simple equivalent circuit shown in Fig. 4. The simplicity of this circuit is afforded by the frequency of crossover (approximately 3kHz) being much less than the cutoff frequency of the transistor so that a low-frequency model is permissible.

The transfer function of this circuit is readily shown to be

$$\frac{V_0}{V_{\text{in}}} = \frac{R_{C1}}{R_{E1}}\left[\frac{1 + s\tau_C}{1 + s\tau_E}\right]$$
$$\cdot\left[\frac{[R_{C_1}/R_{E_1}(1+s\tau_C)+1]D-[sC_BR_{C_2}(1+s\tau_E)+sC_BR_{E_2}(1+s\tau_C)]}{[R_{C_1}/R_{E_1}(1+s\tau_E)+1]D-[sC_BR_{C_2}(1+s\tau_E)+sC_BR_{E_2}(1+s\tau_C)]}\right]$$

where

$$D = (1 + s\tau_C)(1 + s\tau_E) + sR_{C_2}C_B(1 + s\tau_E) + sR_{E_2}C_B(1 + s\tau_C)$$

$$\tau_C = C_C'R_{C_2}, \qquad C_C' = C_C + C_D$$

$$\tau_E = C_E'R_{E_2}, \qquad C_E' = C_A + C_E$$

for

$$\frac{R_{E1}}{R_{C1}} = \frac{R_{E1} + R_{E2}}{R_{C1} + R_{C2}} \qquad \text{and} \qquad \alpha \cong 1\,.$$

It is clearly seen that by making $\tau_E = \tau_C$, the poles and zeros of the system cancel, thus producing a flat crossover. The value of C_B is seen to be unimportant. Large values of C_B will produce a dominant pole and zero which will tend to mask the required equality of τ_C and τ_E.

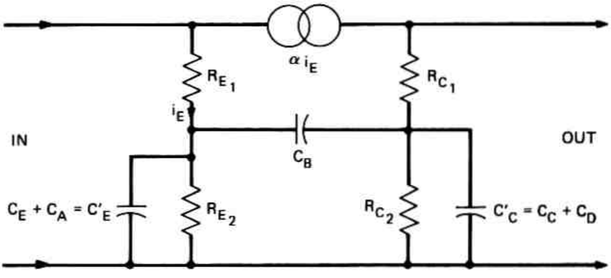


Fig. 4—Equivalent circuit of embedded gain cell.

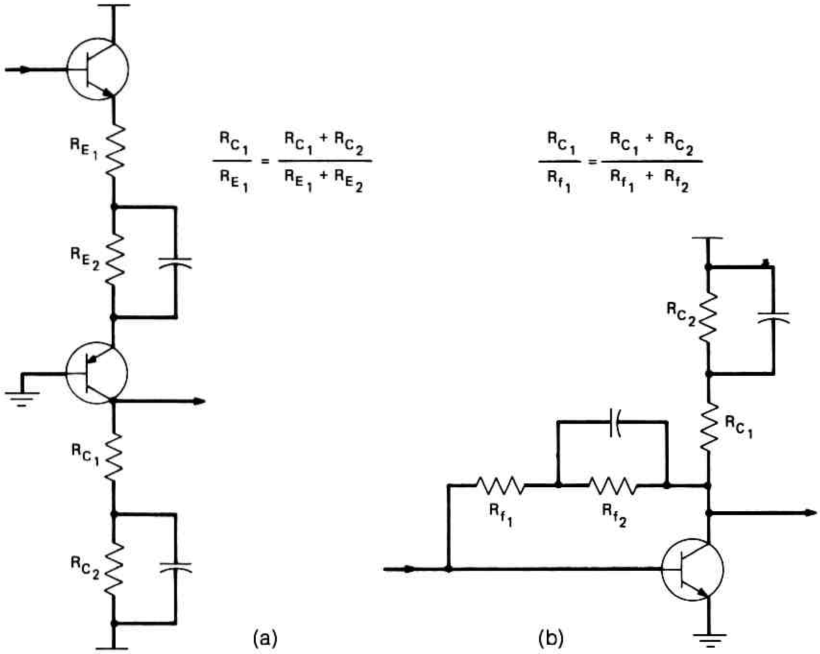

R_{E_1}
$\frac{R_{C_1}}{R_{E_1}} = \frac{R_{C_1}+R_{C_2}}{R_{E_1}+R_{E_2}}$
$\frac{R_{C_1}}{R_{f_1}} = \frac{R_{C_1}+R_{C_2}}{R_{f_1}+R_{f_2}}$
R_{E_2}
R_{C_2}
R_{C_1}
R_{f_1}
R_{f_2}
R_{C_1}
R_{C_2}
(a)
(b)

Fig. 5—Embedding technique applied to (a) modified cascode stage, (b) shunt feedback stage. (Gain equality requirement indicated for lf and hf cases.)

The consequence of separating the lf and hf gain-determining components is that the lf portion may be designed to ensure a high dc stability (e.g., R_{E_2} large) without impairment of the hf optimization.

The technique has more general applications to any gain stage where the stage gain is defined by the ratio of two real impedances; these further applications of embedding are shown for the modified cascode and shunt feedback stages of Figs. 5a and b. The realization of Fig. 5a is particularly important since the optimization of the high-frequency circuit parameters leads to frequencies of operation where the elimination of the Miller effect afforded by this circuit is important. Also, the low external emitter impedances presented provide an enhanced stability, important at these high frequencies.

These circuits constitute a class of split-band additive amplifiers. Split-band amplifiers were first described by Wheeler[7] many years ago, but their successful realization has been retarded by the difficulties of achieving a satisfactory crossover mode in extremely high-frequency amplifiers. The embedding scheme provides a crossover transition at relatively low frequencies so that its physical synthesis is straightforward. Although the internal gain is provided in a product sense, the embedding technique is essentially additive, and thus related to the distributed amplifier of Percival[8] and the Gilbert[9] gain cell.

III. A DC-TO-2.3-GHZ EMBEDDED AMPLIFIER

The embedding principle has been embodied in a broadband amplifier of the type shown in Fig. 3. This amplifier has been designed for optimum hf performance. To demonstrate the way in which this optimization is achieved without compromising the low-frequency stability, it is pertinent to indicate briefly the hf design optimization. The design procedure is based on the simplifying assumptions that the major frequency-limiting mechanisms are the frequency fall-off of the transistor current gain α (modeled as a single pole fall-off), and the collector-base capacitance of the gain transistor (this being assumed to be the dominant parasitic reactance). For simplicity, both effects are evaluated separately. The consequence of the former effect is evaluated by substituting the single-pole approximation for α in the gain transfer function for the stage, assuming negligible loading at the output of the gain transistor (this latter approximation is justifiable since the practical realization employs double emitter follower buffering to the output). The gain transfer function is

$$\frac{V_0}{V_i} = \frac{\alpha R_{C_1}}{R_{E_1}}\left[\frac{R_{E_1}/(1-\alpha)^2}{R_{E_1}/(1-\alpha)^2 + R_s}\right],$$

R_s being the source resistance. Upon substitution of the single-pole approximation to α, and making the further simplification that α_0, the low-frequency, common-base, short-circuit gain, is equal to unity, the following equation results,

$$\frac{V_0}{V_i} = \frac{R_{C_1}}{R_{E_1}}\left[\frac{(1+s\tau_\alpha)}{1+2s\tau_\alpha+(1+R_s/R_{E_1})s^2\tau_\alpha}\right]$$

where τ_α is the time constant associated with the single-pole approximation to the frequency-dependent current gain ($\tau_\alpha \cong 1/\omega_T$). The dominant time constants here are a conjugate pair of poles at

$$p_1, p_2 = \frac{-1}{1+R_s/R_{E_1}} \pm j\frac{\sqrt{R_s/R_{E_1}}}{1+R_s/R_{E_1}}.$$

The position of these poles, normalized to τ_α, is shown in Fig. 6.

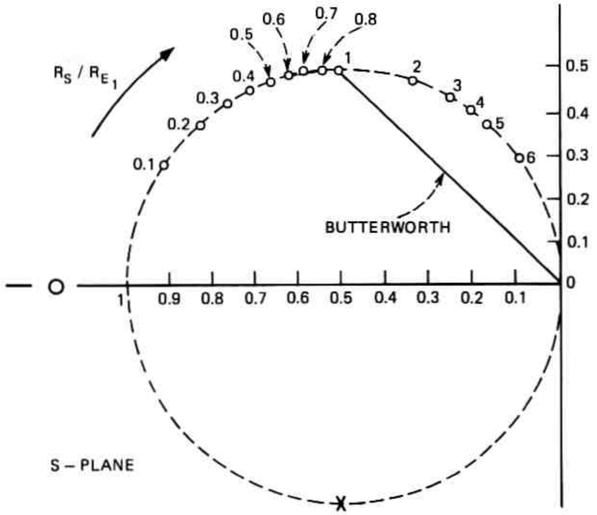


Fig. 6—Pole-zero loci for R_{E_1} variations (normalized to τ_α).

Selecting a normal Butterworth response ($Q = 0.7$) gives

$$\frac{1}{1 + R_s/R_{E_1}} = \frac{\sqrt{R_s/R_{E_1}}}{1 + R_s/R_{E_1}} \qquad \text{or} \qquad R_{E_1} = R_s.$$

This may be interpreted as an upper limit on R_{E_1} which, of course, prohibits the achievement of a high dc stability factor. Practically, the upper value of R_{E_1} is restricted by considerations of providing high-stage gain while maintaining a corner frequency, due to the collector base capacitance, of not less than the corner frequency due to high-frequency fall-off of the current gain. The devices used exhibited an f_T of 4 GHz and a C_{ob} of 0.25 pF. This output capacitance (plus other additive capacitance due to loading) restricts the value of R_{C_1} to 100 Ω for a stage gain of 8 dB, with the concomitant effect that R_{E_1} (including the dynamic internal emitter resistance) be approximately 30 ohms. For efficient quiescent point definition, and hence dc stability, the value of external emitter resistance should be of the order of 1000 ohms. These parameter values allow a complete description of all other parameter values used. The embedded amplifier was designed with a band transition frequency of approximately 3 kHz; this allows easily realizable synthesis. The actual circuit was fabricated using a hybrid technology consisting of tantalum nitride and gold thin films on alumina substrates with beam leaded

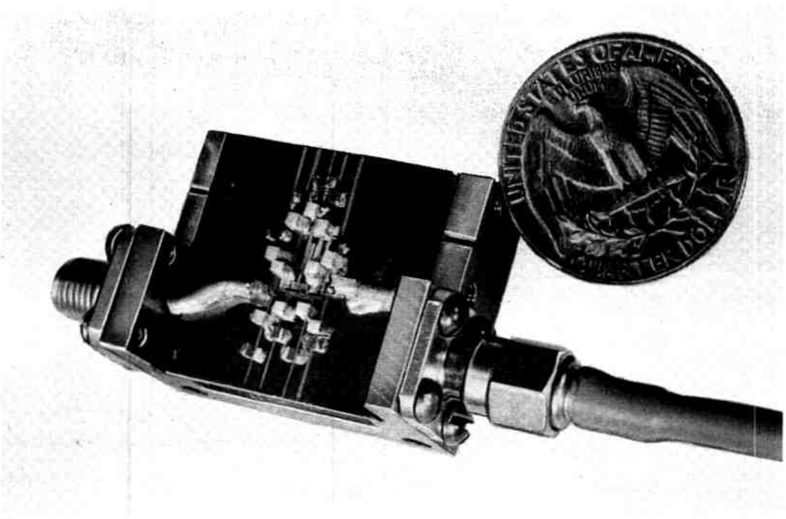


Fig. 7—Embedded amplifier in thin-film, beam lead technology.

device chips. A photograph of this realization is shown in Fig. 7. The pulse response (see Fig. 8) exhibits a 200-ps rise time with a gain of 8 to 10 dB; the frequency response (Fig. 9) shows a flat gain curve up to

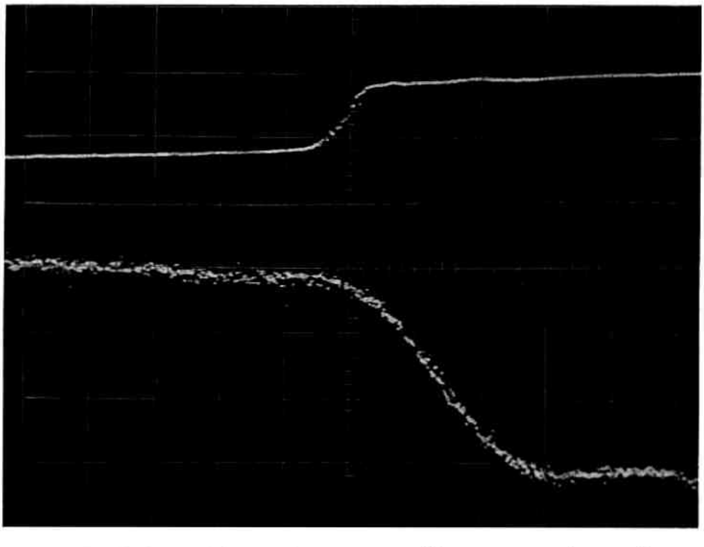


Fig. 8—Embedded amplifier pulse response. Upper trace, input. Lower trace, output. (200 mV/cm, 100 ps/cm.)

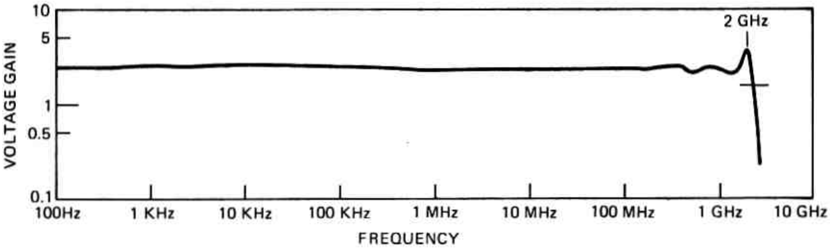


Fig. 9—Eight-decade gain response of broadband embedded amplifier.

2 GHz with the exception of some parasitic resonance-induced deviations just below 2 GHz. The conjugate pole resonance at 2 GHz is also apparent at the cutoff. This should easily be eliminated by choosing a suitable ratio of R_s/R_{E_1} to provide a subcritical Q factor.

The effect of mismatch of the emitter and collector circuit time constants, τ_C and τ_E, is shown in Fig. 10. These waveforms illustrate the effects of gross mismatches where the relaxation times observed are commensurate with a 3-KHz transition frequency.

IV. CONCLUSIONS

A technique has been described for providing gain to dc in hf amplifiers without compromising the hf circuit optimization. This

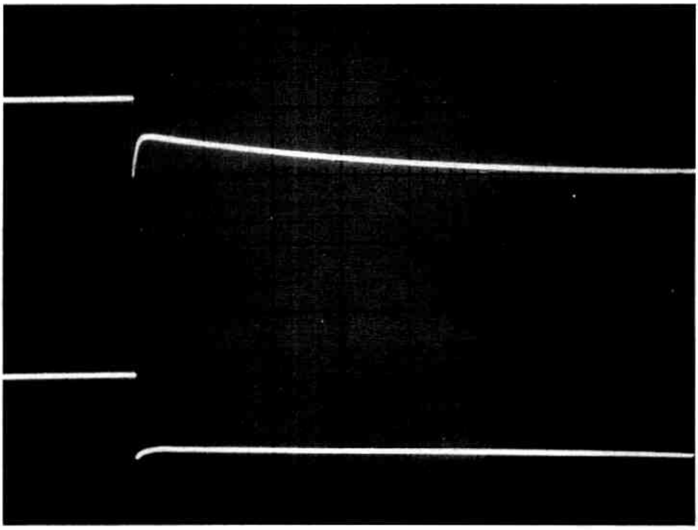


Fig. 10—Mismatch of emitter and collector time constants. Upper, $\tau_E \ll \tau_C$. Lower, $\tau_E \cong \tau_C$. (5 μs/cm.)

embedding technique can be applied to a number of types of gain stage. The common emitter stage, from which results have been obtained, exhibited a dc-to-2.3-GHz bandwidth with a voltage insertion gain of approximately 2.6. The gain was restricted to this value to minimize the effects of collector-to-base junction capacitance. With other stages, such as the modified cascode, this restriction is not as severe, and much higher gains are to be expected. The high-to-low-frequency transition is easily realized, in contrast to other split-level amplifier configurations.

The enhanced dc stability available from the embedding technique favors a cascading of individual gain cells to obtain larger amplifier gains. It is worth noting that the freedom of choice of device operating point inherent in the embedding technique allows an optimization of biasing conditions for low-noise operation, appropriate to post-detection amplification in optical PCM terminals.

V. ACKNOWLEDGMENTS

The authors thank Mr. W. E. Whitacre for some early experimental work, and Mr. D. Babusci and Mr. J. R. Perucca for their assistance in the hybrid thin-film realization.

REFERENCES

1. Solomon, J. E., and Wilson, G. R., "A Highly Densitized, Wideband Monolithic Amplifier," IEEE J. of Solid State Circuits, *SC-1*, No. 1 (September 1966), pp. 19–28.
2. Chaplin, G. B. B., Candy, C. J. N., and Coles, A. J., "Transistor Stages for Broadband Amplifiers," IEEE Paper No. 2892E (May 1959).
3. Denton, R. T., and Kinsel, T. S., "Terminals for a High Speed Optical Pulse Code Modulation Communications System: I. 224 Mbit/s Single Channel," Proc. IEEE, *56*, No. 2 (February 1968), pp. 140–145.
4. White, G., "Optical Modulation at High Information Rates," B.S.T.J., *50*, No. 8 (October 1971), pp. 2607–2645.
5. King, B. G., Ortel, W. C. G., and Schulte, H. J., "Outdoor Optical Transmission Experiments," IEEE International Convention, N. Y., 1971, pp. 82–83.
6. Pettit, J. M., and McWhorter, M. M., *Electronic Amplifier Circuits*, New York: McGraw-Hill Book Co., 1961.
7. Wheeler, H. A., "Maximum Speed of Amplification in a Wideband Amplifier," Wheeler Monograph II, Wheeler Labs., Inc., Great Neck, N. Y., July 1949.
8. Percival, W. S., "Improvements in and Relating to Thermionic Value Circuits," British Patent 460,562, 1937.
9. Gilbert, B., "A DC-500 MHz Amplifier/Multiplier Principle," IEEE, Solid State Circuit Conference, Philadelphia, 1968, pp. 114–115.

Copyright © 1973 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

Coupling Coefficients For Imperfect Asymmetric Slab Waveguides

By D. MARCUSE

(Manuscript received July 6, 1972)

This paper presents a collection of formulas that are necessary for the treatment of radiation and mode conversion phenomena of imperfect asymmetric slab waveguides. The coupled mode theory of dielectric waveguides is briefly reviewed, and general expressions for the coupling coefficients are given. The field expression of the guided and the radiation TE and TM modes of the asymmetric slab waveguide are stated, and are used to derive formulas for the coupling coefficient for slight core boundary irregularities.

I. INTRODUCTION

Mode coupling phenomena and radiation losses caused by core-cladding interface irregularities have been studied extensively for symmetric slab waveguides and for round optical fibers.[1–7] These results are not immediately applicable to the asymmetric slab waveguides used in integrated optics circuits. It is the purpose of this paper to collect the formulas for the normal modes of the asymmetric slab waveguides, and for the coupling coefficients between guided modes and guided and radiation modes caused by core boundary irregularities of these waveguides.

The coupling coefficients between guided modes are useful for the design of distributed feedback sections for lasers and for an evaluation of unintentional mode coupling caused by core boundary roughness. The results collected in this paper are further necessary for the evaluation of radiation losses caused by core boundary roughness.

Because of the many parameters that enter into the theory, it is impossible to evaluate the formulas in graphical form for all cases of practical interest. This paper is thus a collection of the required formulas which the reader can use to evaluate his particular problems.

II. SUMMARY OF THE COUPLED MODE THEORY

The coupling theory is based on an expansion of the solution of Maxwell's equations in terms of normal modes. The general theory of the mode expansion and the derivation of the coupling coefficients have been published by A. W. Snyder.[2,3] His theory is based on local normal modes. Local normal modes resemble the modes of the ideal asymmetric slab waveguide with perfect core boundary. However, the boundary of the perfect guide is allowed to change in such a way that it coincides with the actual deformed core boundary at the particular point z along the waveguide axis at which the coupling coefficients are to be evaluated. Since the waveguide width parameter is no longer a constant, the local normal modes are not themselves solutions of Maxwell's equations. They must be superimposed with z-dependent expansion coefficients to form such a solution. The fact that these modes form a complete orthogonal set and coincide with the modes of a fictitious waveguide, the width of which is locally (at the point z under consideration) the same as that of the deformed waveguide, explains the name "local normal modes." It is also possible to express the general field in terms of the modes of the ideal waveguide, the constant width of which differs slightly from that of the actual waveguide. This expansion suffers from convergence difficulties that are caused by the fact that the normal components of the electric field are discontinuous at the core boundary. The modes of the ideal guide are discontinuous at the dielectric interface of the ideal guide which does not coincide with that of the actual guide. The expansion in terms of ideal modes of the waveguide is thus discontinuous term by term at a point where the entire series must be continuous, and furthermore, it must describe a discontinuous field at the interface of the actual waveguide at a point where each individual term of the expansion remains continuous. The expansion in terms of local normal modes, on the other hand, describes the field discontinuity with a series, the individual terms of which are discontinuous in just the right way at the point of discontinuity of the entire series. The convergence behavior of this latter expansion can thus be expected to be superior to the expansion in terms of ideal modes.

The electric and magnetic fields of the imperfect asymmetric slab waveguide are expressed by the series expansions

$$\mathbf{E} = \sum_{\nu} \left(c_\nu^{(+)} \boldsymbol{\mathcal{E}}_\nu^{(+)} + c_\nu^{(-)} \boldsymbol{\mathcal{E}}_\nu^{(-)}\right) \tag{1}$$

$$\mathbf{H} = \sum_{\nu} \left(c_\nu^{(+)} \boldsymbol{\mathcal{H}}_\nu^{(+)} + c_\nu^{(-)} \boldsymbol{\mathcal{H}}_\nu^{(-)}\right). \tag{2}$$

The expansion coefficients $c_\nu^{(+)}$ and $c_\nu^{(-)}$ are functions of the length coordinate z. The superscripts $(+)$ and $(-)$ indicate waves traveling in positive and negative z-direction. The sums in (1) and (2) are symbolic representations of a summation over guided modes plus an integration over the radiation modes of the continuum.[4,5] In order to simplify the notation, both sum and integral are indicated by the same symbol. In the integral, the summation index ν is replaced by the continuous variable ν, and the sum must be understood as the integral

$$\sum_\nu \rightarrow \int_0^\infty d\nu. \tag{3}$$

The local normal modes E_ν and H_ν are solutions of the equations

$$\mp i\beta_\nu(\mathbf{e}_z \times \mathcal{H}_\nu^{(\pm)}) + \nabla_t \times \mathcal{H}_\nu^{(\pm)} = i\omega\epsilon_0 n^2 \mathcal{E}_\nu^{(\pm)} \tag{4}$$

$$\mp i\beta_\nu(\mathbf{e}_z \times \mathcal{E}_\nu^{(\pm)}) + \nabla_t \times \mathcal{E}_\nu^{(\pm)} = - i\omega\mu_0 \mathcal{H}_\nu^{(\pm)}. \tag{5}$$

The upper and lower signs and superscripts belong together. The symbols appearing in these equations have the following meaning.

β_ν = propagation constant of mode ν
$\mathbf{e}_z$ = unit vector in z-direction
∇_t = transverse part of the operator ∇
ω = radian frequency
ϵ_0 = dielectric permittivity of the vacuum
μ_0 = magnetic susceptibility of the vacuum
n = dielectric constant of the waveguide $[n = n(x, y, z)]$.

Substitution of the field expansions (1) and (2) into Maxwell's equations and use of the orthogonality relations [see (9)] lead to the set

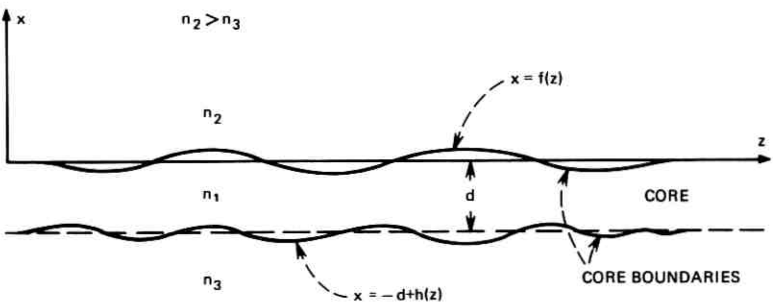


Fig. 1—Sketch of the asymmetric slab waveguide with distorted core boundaries.

of coupled wave equations[2,3,4]

$$\frac{dc_\mu^{(+)}}{dz} = -i\beta_\mu c_\mu^{(+)} + \sum_\nu \left(K_{\mu\nu}^{(+,+)} c_\nu^{(+)} + K_{\mu\nu}^{(+,-)} c_\nu^{(-)}\right) \tag{6}$$

$$\frac{dc_\mu^{(-)}}{dz} = i\beta_\mu c_\mu^{(-)} - \sum_\nu \left(K_{\mu\nu}^{(-,+)} c_\nu^{(+)} + K_{\mu\nu}^{(-,-)} c_\nu^{(-)}\right). \tag{7}$$

The coupling coefficients have the form[2,3]

$$K_{\mu\nu}^{(\pm,p)} = \frac{1}{4P}\int_{-\infty}^{\infty}\left\{\pm\left(\frac{\partial\,\mathcal{E}_\nu^{(+)}}{\partial z} \times \mathcal{H}_\mu^{(+)*}\right)\cdot \mathbf{e}_z - p\left(\mathcal{E}_\mu^{(+)*} \times \frac{\partial \mathcal{H}_\nu^{(+)}}{\partial z}\right)\cdot \mathbf{e}_z\right\} dx. \tag{8}$$

The asterisk indicates complex conjugation. The superscript p stands for $(+)$ or $(-)$ while the factor p assumes the values $+1$ and -1. P is a normalization parameter which is related to the power carried by the modes via the relation

$$\frac{1}{2}\int_{-\infty}^{\infty}(\mathcal{E}_\nu^{(+)} \times \mathcal{H}_{\nu'}^{(+)*})\cdot \mathbf{e}_z dx = P\delta_{\nu\nu'}. \tag{9}$$

The symbol $\delta_{\nu\nu'}$ indicates the Dirac delta function if both ν and ν' represent continuous variables, it represents the Kronecker delta symbol if both ν and ν' are discrete labels, and it is zero if one subscript belongs to discrete modes while the other indicates a mode of the continuum.

The coupling coefficients (8) are not very easy to evaluate since they are expressed in terms of derivatives of the mode functions. A. W. Snyder[3] has shown that the coupling coefficients can be transformed to the following more useful form. $(\beta_\nu{}^{(-)} = -\beta_\nu{}^{(+)})$

$$K_{\mu\nu}^{(\pm,p)} = -\frac{\omega\epsilon_0}{4P(\beta_\mu^{(\pm)} - \beta_\nu^{(p)})}\int_{-\infty}^{\infty}\frac{\partial n^2}{\partial z}\mathcal{E}_\mu^{(\pm)*}\cdot \mathcal{E}_\nu^{(p)} dx. \tag{10}$$

The coupled wave equations (6) and (7), with the coupling coefficients (10), provide an exact description of imperfect dielectric waveguides. The one-dimensional integral in (10) constitutes a specialization to a two-dimensional problem in view of our interest in the asymmetric slab waveguide. By extending the integration over the cross-sectional x, y plane, the general coupling coefficients are obtained.

For our purpose, it is advantageous to derive approximate coupling coefficients for asymmetric slab waveguides with discontinuous index distributions. We use the fact that the normal component E_x, of the electric field obeys the relation

$$n_1^2 E_{x'1} = n_2^2 E_{x'2}. \tag{11}$$

It is shown in Fig. 1 that n_1 and n_2 are the values of the refractive index at either side of the interface. In order to derive the desired expression for the coupling coefficient, we assume that the discontinuous index distribution is smoothed out (in an arbitrary way) into a continuous distribution. We assume that the wavelength of the radiation is very much larger than the region over which the refractive index varies continuously and write

$$E_{x'} = \frac{n_1^2}{n^2} E_{x'1}. \tag{12}$$

We show in the appendix how the integral in (10) can be evaluated and obtain the result

$$\begin{aligned} K_{\mu\nu}^{(\pm,p)} = -\frac{\omega\epsilon_0}{4P(\beta_\mu^{(\pm)} - \beta_\nu^{(p)})} \Bigg\{ & (n_1^2 - n_2^2)\frac{df}{dz}\left[\frac{n_1^2}{n_2^2}\mathcal{E}_{\mu x}^{(\pm)*}\mathcal{E}_{\nu x}^{(p)} + \mathcal{E}_{\mu y}^{(\pm)*}\mathcal{E}_{\nu y}^{(p)}\right. \\ & \left. + \mathcal{E}_{\mu z}^{(\pm)*}\mathcal{E}_{\nu z}^{(p)}\right]_{x=f(z)} - (n_1^2 - n_3^2)\frac{dh}{dz}\left[\frac{n_1^2}{n_3^2}\mathcal{E}_{\mu x}^{(\pm)*}\mathcal{E}_{\nu x}^{(p)}\right. \\ & \left. + \mathcal{E}_{\mu y}^{(\pm)*}\mathcal{E}_{\nu y}^{(p)} + \mathcal{E}_{\mu z}^{(\pm)*}\mathcal{E}_{\nu z}^{(p)}\right]_{x=-d+h(z)} \Bigg\}. \end{aligned} \tag{13}$$

The index distribution can now again be considered as discontinuous. The functions $f(z)$ and $h(z)$ describe the deformation of the upper and lower side of the core boundary (see Fig. 1). The field components are taken inside the core region at the core boundary. The refractive index of the core is n_1 while n_2 and n_3 are the indices above and below the core region. The electric field components are related in the following way.

$$\left.\begin{aligned} \mathcal{E}_{\nu x}^{(-)} &= \mathcal{E}_{\nu x}^{(+)} \\ \mathcal{E}_{\nu y}^{(-)} &= \mathcal{E}_{\nu y}^{(+)} \\ \mathcal{E}_{\nu z}^{(-)} &= -\mathcal{E}_{\nu z}^{(+)} \end{aligned}\right\}. \tag{14}$$

The approximation involved in the coupling coefficient (13) consists in using the x component $E_{\nu x}$ and z component $E_{\nu z}$ of the local normal modes instead of their normal and tangential components with respect to the interface. The approximation is valid provided that

$$\frac{df}{dz} \ll 1 \qquad \text{and} \qquad \frac{dh}{dz} \ll 1. \tag{15}$$

For many practical applications it is sufficient to use an approximate solution of the coupled wave eqs. (6) and (7). In particular, for the calculation of the radiation loss coefficient, we use the approximate solution of (6) ($\mu = \rho$)

$$c_\rho^{(\pm)}(z) = \pm\, c_\nu^{(p)} e^{-i\beta_\rho^{(\pm)} z} \int_0^z K_{\rho\nu}^{(\pm,p)}(u) \cdot \exp\left[-i \int_0^u (\beta_\nu^{(p)}(v) - \beta_\rho^{(\pm)}) dv \right] du. \tag{16}$$

The coefficient $c_\nu^{(p)}$ is the amplitude of the guided mode, the losses of which we want to calculate, taken at $z = 0$. The propagation constant β_ν is a function of z since it belongs to a guided mode in a non-uniform waveguide, β_ρ is independent of z since it belongs to a radiation mode. The relative power loss $\Delta P_\nu / P_\nu$ that mode ν suffers in traveling from $z = 0$ to $z = L$ is given in Refs. 5 and 6.

$$\frac{\Delta P_\nu}{P_\nu} = \frac{1}{|c_\nu^{(p)}|^2} \sum \int_0^{n_2 k} |c_\rho^{(\pm)}(L)|^2 d\rho. \tag{17}$$

The sum in front of the integral sign indicates that we must add up the contributions of forward and backward traveling modes as well as the contributions from the various kinds of radiation modes that will be discussed in the next section. The integral extends over the range of ρ values that belongs to propagating (non-evanescent) radiation modes. We show below that the functional form of the radiation modes is not the same over the entire integration range.

III. MODES OF THE ASYMMETRIC SLAB WAVEGUIDE

We consider only the special case in which there is no field variation and no waveguide distortion in y direction. This fact is symbolically expressed by the equation

$$\frac{\partial}{\partial y} = 0. \tag{18}$$

With the restriction (18), the fields of the slab waveguide can be classified as either TE or TM fields.[8] The TE fields have only the following non-vanishing field components

$$E_y,\ H_x,\ H_z. \tag{19}$$

The TM fields have the non-vanishing field components

$$H_y,\ E_x,\ E_z. \tag{20}$$

It is assumed that the refractive indices of the waveguide are ordered in the following way

$$n_1 > n_2 \geqq n_3. \tag{21}$$

IV. GUIDED TE MODES

The x and z components of the magnetic field follow from the E_y component by differentiation

$$H_x = -\frac{i}{\omega\mu_0}\frac{\partial E_y}{\partial z} \tag{22}$$

$$H_z = \frac{i}{\omega\mu_0}\frac{\partial E_y}{\partial x}. \tag{23}$$

The guided TE modes of the asymmetric slab waveguide are obtained as follows (the factor $\exp[i(\omega t - \beta z)]$ is always suppressed):

$$\mathcal{E}_y = A_g e^{-\gamma x} \qquad \text{for} \qquad 0 \leqq x < \infty \tag{24}$$

$$\mathcal{E}_y = A_g\left(\cos\kappa x - \frac{\gamma}{\kappa}\sin\kappa x\right) \qquad \text{for} \qquad -d \leqq x \leqq 0 \tag{25}$$

$$\mathcal{E}_y = A_g\left(\cos\kappa d + \frac{\gamma}{\kappa}\sin\kappa d\right)e^{\theta(x+d)} \qquad \text{for} \qquad -\infty < x \leqq -d. \tag{26}$$

The parameters appearing in these field expressions are defined by the equations:

$$\gamma^2 = \beta^2 - n_2^2 k^2 \tag{27}$$

$$\theta^2 = \beta^2 - n_3^2 k^2 \tag{28}$$

$$\kappa^2 = n_1^2 k^2 - \beta^2 \tag{29}$$

$$k^2 = \omega^2 \epsilon_0 \mu_0. \tag{30}$$

The propagation constant β is determined from the eigenvalue equation

$$\tan \kappa d = \frac{\kappa(\gamma + \theta)}{\kappa^2 - \gamma\theta}. \tag{31}$$

The normalization of the mode is obtained by expressing the amplitude coefficient A_g in terms of the power P carried by the mode.

$$A_g^2 = \frac{4\kappa^2\omega\mu_0 P}{|\beta|\left(d + \frac{1}{\gamma} + \frac{1}{\theta}\right)(\kappa^2 + \gamma^2)}. \tag{32}$$

The mode labels ν, that were used in the coupled wave equations and the field expansions, are suppressed. The modes are labeled according to the solutions of the eigenvalue equation (31).

V. TE RADIATION MODES

The propagation constants of the radiation modes do not have a discrete set of values. However, the asymmetric slab waveguide has different types of radiation modes. In the range

$$n_3 k \leqq \beta \leqq n_2 k \tag{33}$$

we find only one type of radiation mode, the fields of which decay exponentially into the region three with refractive index n_3, but are standing waves in the space above the waveguide with refractive index n_2. We can visualize the physical mechanism for exciting these modes by assuming that a source at infinity sends a plane wave that impinges on the core of the slab waveguide under an angle that is given by

$$\cos \alpha = \frac{\beta}{n_2 k}. \tag{34}$$

The incident plane wave penetrates into the core region but is totally internally reflected if the angle α stays in the range given by (33). This total internal reflection occurs because we assumed that $n_3 < n_2$. In the space above the core we find a reflected wave added to the incident wave supplied by the external source. This explains the occurrence of standing waves in this region. It is not possible to find solutions of Maxwell's equations satisfying the boundary conditions which have only traveling waves outside of the core region.

In the range of propagation constants given by (33) we find the following expression for the field of the radiation modes

$$\mathcal{E}_y = A_r \cos \rho x + \frac{\sigma}{\rho} B_r \sin \rho x \qquad \text{for} \qquad 0 \leqq x < \infty \tag{35}$$

$$\mathcal{E}_y = A_r \cos \sigma x + B_r \sin \sigma x \qquad \text{for} \qquad -d \leqq x \leqq 0 \tag{36}$$

$$\mathcal{E}_y = (A_r \cos \sigma d - B_r \sin \sigma d) e^{-i\Delta(x+d)} \qquad \text{for} \qquad -\infty < x \leqq -d. \tag{37}$$

H_x and H_z are obtained from E_y with the help of (22) and (23). The constant B is related to the constant A by the equation

$$B_r = \frac{\Delta - i\sigma \tan \sigma d}{\Delta \tan \sigma d + i\sigma} A_r \tag{38}$$

and the parameters appearing in these equations are defined as follows

$$\sigma^2 = n_1^2 k^2 - \beta^2 \tag{39}$$

$$\rho^2 = n_2^2 k^2 - \beta^2 \tag{40}$$

$$\Delta^2 = n_3^2 k^2 - \beta^2. \tag{41}$$

Note that Δ is imaginary for β values in the range given by (33).

It is convenient to identify the parameter ρ with the mode label ν to label the radiation modes. We thus use the identity

$$\delta_{\nu\nu'} = \delta(\rho - \rho') \tag{42}$$

in (9) and find for the amplitude coefficient A the relation

$$A_r^2 = \frac{4\omega\mu_0 P}{\pi|\beta|} \frac{\rho^2\left(\sigma \cos \sigma d + \dfrac{\Delta}{i} \sin \sigma d\right)^2}{\rho^2\left(\sigma \cos \sigma d + \dfrac{\Delta}{i} \sin \sigma d\right)^2 + \sigma^2\left(\sigma \sin \sigma d - \dfrac{\Delta}{i} \cos \sigma d\right)^2}. \tag{43}$$

With β in the range (33) we find that ρ is confined to the region

$$0 \leqq \rho \leqq (n_2^2 - n_3^2)^{\frac{1}{2}} k. \tag{44}$$

Next we proceed to list the radiation modes that belong to propagation constants in the range

$$\text{and} \qquad \left.\begin{array}{ll} 0 \leqq \beta \leqq n_3 k & \beta \ \text{real} \\ 0 \leqq |\beta| < \infty & \beta \ \text{imaginary} \end{array}\right\}. \tag{45}$$

The corresponding range for ρ is given by

$$(n_2^2 - n_3^2)^{\frac{1}{2}}k < \rho < \infty. \tag{46}$$

For real values of β, these modes propagate along the z axis while they are evanescent waves in z direction for imaginary values of β. It is again possible to visualize the modes in the range (46) as being excited by a source outside of the waveguide core located at infinity. This source sends a plane wave toward the slab whose angle of propagation with respect to the z axis is given by (34). However, there is now no longer total internal reflection at the lower boundary of the core so that we obtain an incident and reflected (in x-direction) wave in the space above as well as inside the core. Below the core there is a transmitted propagating wave. However, we may now assume with equal justification that a second source sends a plane wave in the direction of the core from below. If both sources are turned on simultaneously, we obtain standing waves (in x-direction) below as well as above the waveguide core. The exact form of the radiation field depends on the relative phases between the two sources. It is thus not surprising that there should be an infinite number of ways in which orthogonal sets of radiation modes can be constructed.

We list only the E_y components of the modes and obtain the H_x and H_z components by differentiation from (22) and (23). ($i = 1, 2$)

$$\mathcal{E}_y = C_r\left(\cos \rho x + \frac{\sigma}{\rho} F_i \sin \rho x\right) \qquad 0 \leqq x < \infty \tag{47}$$

$$\mathcal{E}_y = C_r(\cos \sigma x + F_i \sin \sigma x) \qquad -d \leqq x \leqq 0 \tag{48}$$

$$\mathcal{E}_y = C_r\left\{(\cos \sigma d - F_i \sin \sigma d) \cos \Delta(x + d) + \frac{\sigma}{\Delta}(\sin \sigma d + F_i \cos \sigma d) \sin \Delta(x + d)\right\} \qquad -\infty < x \leqq -d. \tag{49}$$

The parameters σ, ρ and Δ are given by (39), (40) and (41), Δ is now a real constant. Whereas the amplitude coefficient B_r in (35) through (37) was related to A_r by the boundary conditions, we now face the situation where the amplitude coefficient F_i remains arbitrary. Equations (47) through (49) satisfy Maxwells' equations as well as the boundary conditions without any further restriction having to be imposed on the coefficient F_i. This freedom of choice is related to the arbitrary amplitude and phase of the two plane wave sources of the radiation mode mentioned above. We must choose F_i in such a way that two radiation modes with the same value of the propagation constant, but different values of F_i become mutually orthogonal. But even this requirement does not specify the possible values of F_i uniquely. We are thus free to choose F values according to our own convenience. Of the infinitely many possibilities, we choose the F_i coefficients such that in the limit $n_2 = n_3$, even and odd radiation modes result. In the asymmetric slab waveguide no even or odd modes exist. But the guided modes become either even or odd as n_2 approaches n_3. We obtain the same symmetries for the radiation modes by a suitable choice of the F_i coefficients.

$$F_{1,2} = \frac{1}{(\sigma^2 - \Delta^2)\sin 2\sigma d}\left\{(\sigma^2 - \Delta^2)\cos 2\sigma d + \frac{\Delta}{\rho}(\sigma^2 - \rho^2) \pm \left[(\sigma^2 - \Delta^2)^2 + 2\frac{\Delta}{\rho}(\sigma^2 - \Delta^2)(\sigma^2 - \rho^2)\cos 2\sigma d + \frac{\Delta^2}{\rho^2}(\sigma^2 - \rho^2)^2\right]^{\frac{1}{2}}\right\}. \quad (50)$$

The + sign (− sign) belongs to the odd (even) mode in the limit $n_2 = n_3$, $\Delta = \rho$. We identify the mode label ν again with the transverse propagation constant ρ, and obtain from the normalization condition (9) and (42) the relation between the amplitude coefficient C_r and the power P

$$C_r^2 = \frac{4\omega\mu_0 P}{\pi|\beta|}\left[\frac{\Delta}{\rho}(\cos\sigma d - F_i \sin\sigma d)^2 + \frac{\sigma^2}{\Delta\rho}(\sin\sigma d + F_i\cos\sigma d)^2 + 1 + \frac{\sigma^2}{\rho^2}F_i^2\right]^{-1}. \quad (51)$$

We have thus obtained two distinct sets of radiation modes, the propagation constants of which lie in the range (45). The two sets are distinguished by the labels $i = 1$ and $i = 2$. F_1 and F_2 follow from (50) if we use both signs of the square root expression. It is noteworthy that the following relation holds.

$$F_1 F_2 = -1. \quad (52)$$

This listing contains the complete set of TE guided and radiation modes consistent with the restriction (18). All modes are mutually orthogonal to each other. We have concentrated on the forward traveling modes. The backward traveling modes are obtained simply by changing the sign of β, ($\beta^{(-)} = -\beta^{(+)}$).

VI. GUIDED TM MODES

The TM modes are very similar to the TE modes except that the roles of the field components are interchanged. We now list only the H_y component and obtain the E_x and E_z components of the field by differentiation

$$E_x = \frac{i}{n^2\omega\epsilon_0}\frac{\partial H_y}{\partial z} \tag{53}$$

$$E_z = \frac{-i}{n^2\omega\epsilon_0}\frac{\partial H_y}{\partial x} \tag{54}$$

$$\mathcal{H}_y = D_g e^{-\gamma x} \qquad 0 \leqq x < \infty \tag{55}$$

$$\mathcal{H}_y = D_g\left(\cos \kappa x - \frac{n_1^2}{n_2^2}\frac{\gamma}{\kappa}\sin \kappa x\right) \qquad -d \leqq x \leqq 0 \tag{56}$$

$$\mathcal{H}_y = D_g\left(\cos \kappa d + \frac{n_1^2}{n_2^2}\frac{\gamma}{\kappa}\sin \kappa d\right)e^{\theta(x+d)} \qquad -\infty < x \leqq -d. \tag{57}$$

The parameters κ, γ and θ are defined by (27) through (29). The eigenvalue equation is

$$\tan \kappa d = \frac{n_1^2\kappa(n_3^2\gamma + n_2^2\theta)}{n_2^2 n_3^2\kappa^2 - n_1^4\gamma\theta} \tag{58}$$

and the amplitude coefficient is given by

$$D_g^2 = \frac{4\omega\epsilon_0 P}{|\beta|} \times \frac{n_1^2 n_2^4\kappa^2}{(n_2^4\kappa^2 + n_1^4\gamma^2)\left[d + \dfrac{n_1^2 n_2^2}{\gamma}\dfrac{\kappa^2+\gamma^2}{n_2^4\kappa^2 + n_1^4\gamma^2} + \dfrac{n_1^2 n_3^2}{\theta}\dfrac{\kappa^2+\theta^2}{n_3^4\kappa^2 + n_1^4\theta^2}\right]}. \tag{59}$$

VII. TM RADIATION MODES

The TM radiation modes must again be split up into two ranges. In the range $0 \leqq \rho \leqq (n_2^2 - n_3^2)^{\frac{1}{2}}k$, the fields have the form

$$\mathcal{H}_y = D_r \cos \rho x + \frac{n_2^2}{n_1^2}\frac{\sigma}{\rho} G_r \sin \rho x \qquad 0 \leqq x < \infty \tag{60}$$

$$\mathcal{H}_y = D_r \cos \sigma x + G_r \sin \sigma x \qquad -d \leqq x \leqq 0 \tag{61}$$

$$\mathcal{H}_y = (D_r \cos \sigma d - G_r \sin \sigma d) e^{-i\Delta(x+d)} \qquad -\infty \leqq x \leqq -d. \tag{62}$$

The boundary conditions require that G_r is related to D_r in the following way:

$$G_r = \frac{n_1^2 \Delta \cos \sigma d - i n_3^2 \sigma \sin \sigma d}{n_1^2 \Delta \sin \sigma d + i n_3^2 \sigma \cos \sigma d} D_r. \tag{63}$$

The parameters σ, ρ and Δ are defined by (39) through (41), $-i\Delta$ is a real positive quantity. The amplitude coefficient D_r is related to the power coefficient P

$$D_r^2 = \frac{4 n_2^2 \omega \epsilon_0 P}{\pi |\beta|} \frac{1}{1 + \dfrac{n_2^4}{n_1^4}\dfrac{\sigma^2}{\rho^2}\dfrac{G_r^2}{D_r^2}}. \tag{64}$$

In the range $(n_2^2 - n_3^2)^{\frac{1}{2}}k \leqq \rho \leqq \infty$ the radiation modes have the form $(i = 1, 2)$

$$\mathcal{H}_y = S_r\left(\cos \rho x + \frac{n_2^2}{n_1^2}\frac{\sigma}{\rho} R_i \sin \rho x\right) \qquad 0 \leqq x < \infty \tag{65}$$

$$\mathcal{H}_y = S_r(\cos \sigma x + R_i \sin \sigma x) \qquad -d \leqq x \leqq 0 \tag{66}$$

$$\mathcal{H}_y = S_r\left\{(\cos \sigma d - R_i \sin \sigma d) \cos \Delta(x + d) + \frac{n_3^2}{n_1^2}\frac{\sigma}{\Delta}(\sin \sigma d + R_i \cos \sigma d) \sin \Delta(x + d)\right\} \qquad -\infty < x \leqq -d. \tag{67}$$

The coefficients R_i are again arbitrary. Our choice

$$R_{1,2} = \frac{1}{(n_3^4\sigma^2 - n_1^4\Delta^2)\sin 2\sigma d}\Bigg\{(n_3^4\sigma^2 - n_1^4\Delta^2)\cos 2\sigma d + \frac{n_3^2}{n_2^2}\frac{\Delta}{\rho}(n_2^4\sigma^2 - n_1^4\rho^2) \pm \Bigg[(n_3^4\sigma^2 - n_1^4\Delta^2)^2 + \frac{n_3^4}{n_2^4}\frac{\Delta^2}{\rho^2}(n_2^4\sigma^2 - n_1^4\rho^2)^2 + 2\frac{n_3^2}{n_2^2}\frac{\Delta}{\rho}(n_3^4\sigma^2 - n_1^4\Delta^2)(n_2^4\sigma^2 - n_1^4\rho^2)\cos 2\sigma d\Bigg]^{\frac{1}{2}}\Bigg\} \quad (68)$$

causes the modes with index $i = 1$ to be orthogonal to the modes with index $i = 2$ and, in addition, assures that these modes become even and odd in the limit $n_2 = n_3$. [The + sign (− sign) belongs to the odd (even) mode.] The normalization is given by

$$S_\nu^2 = \frac{4\omega\epsilon_0 P}{\pi|\beta|}\left\{\frac{1}{n_2^2} + \frac{n_2^2}{n_1^4}\frac{\sigma^2}{\rho^2}R_i^2 + \frac{n_3^2}{n_1^4}\frac{\sigma^2}{\rho\Delta}(\sin\sigma d + R_i\cos\sigma d)^2 + \frac{1}{n_3^2}\frac{\Delta}{\rho}(\cos\sigma d - R_i\sin\sigma d)^2\right\}^{-1}. \quad (69)$$

All amplitude coefficients for the TE and TM modes were taken to be real quantities. This assumption does not lead to a loss of generality since the necessary phase factors are incorporated in the expansion coefficients c_ν.

VIII. COUPLING COEFFICIENTS

With the help of the expressions for the normal modes and the coupling coefficients (10), any problem of the asymmetric slab waveguide with arbitrary irregularities of its refractive index distribution can be solved, provided that the restriction (18) is imposed. Problems caused by core boundary irregularities or by gentle tapers can be solved with the help of the coupling coefficients (13). For convenience, a few coupling coefficients will be worked out explicitly.

As long as the restriction (18) applies, TE modes do not couple to TM modes. All coupling coefficients between TE and TM modes vanish. We restrict the discussion to listing the coupling coefficients between guided TE modes, guided TM modes, and to coupling from a guided TE or TM mode to its respective radiation modes for the case of core boundary irregularities.

The coupling coefficients between two guided TE modes can be obtained from (13) and (25). ($p = +$ or $-$)

$$K_{\mu\nu}^{(\pm,p)} = -\frac{\kappa_\mu\kappa_\nu\left[\dfrac{df}{dz} - \dfrac{\sin\kappa_\mu d\sin\kappa_\nu d}{|\sin\kappa_\mu d\sin\kappa_\nu d|}\dfrac{dh}{dz}\right]}{(\beta_\mu^{(\pm)} - \beta_\nu^{(p)})\left[|\beta_\mu\beta_\nu|\left(d+\dfrac{1}{\gamma_\mu}+\dfrac{1}{\theta_\mu}\right)\left(d+\dfrac{1}{\gamma_\nu}+\dfrac{1}{\theta_\nu}\right)\right]^{\frac{1}{2}}}. \tag{70}$$

The eigenvalue equation (31) was used to express (70) in this simple form. This coupling coefficient (and all others to be listed below) holds for the special case $f(z) \ll \pi/\kappa$, $h(z) \ll \pi/\kappa$ with κ of (29). Instead of using the values of the field at $x = f(z)$ and $x = -d + h(z)$, the field values at $x = 0$ and $x = -d$ were used. In order to see the agreement of this coupling coefficient with the coupling coefficient for the symmetric case [eq. (7) of Ref. 7], it is necessary to note that the core thickness d of this paper corresponds to $2d$ of Ref. 7. In addition, we need to keep in mind that only the Fourier components of $f(z)$ and $h(z)$ with spatial frequency $\beta_\mu^{(\pm)} - \beta_\nu^{(p)}$ contribute to coupling between modes μ and ν. The derivatives appearing in (70) are thus equivalent to the products $i(\beta_\mu^{(\pm)} - \beta_\nu^{(p)})\ f(z)$ and $i(\beta_\mu^{(\pm)} - \beta_\nu^{(p)})h(z)$. Keeping these remarks in mind, complete agreement of (70) with (7) of Ref. 7 is obtained for the special case $n_2 = n_3$, $\gamma = \theta$.

The coupling coefficient for coupling between a guided TE mode ν and a TE radiation mode ρ follows similarly from (13), (25), and (36) for radiation in the range $0 \leqq \rho \leqq (n_2^2 - n_3^2)^{\frac{1}{2}}k$

$$K_{\rho\nu}^{(\pm,p)} = -\frac{(n_1^2 - n_2^2)^{\frac{1}{2}}k\kappa_\nu\rho\left(\sigma\cos\sigma d + \dfrac{\Delta}{i}\sin\sigma d\right)}{(\beta_\rho^{(\pm)} - \beta_\nu^{(p)})}$$

$$\cdot\left\{\frac{df}{dz} - \frac{\sin\kappa_\nu d}{|\sin\kappa_\nu d|}\left(\frac{n_1^2 - n_3^2}{n_1^2 - n_2^2}\right)^{\frac{1}{2}}\frac{\sigma}{\sigma\cos\sigma d + \dfrac{\Delta}{i}\sin\sigma d}\frac{dh}{dz}\right\}$$

$$\cdot\left\{\pi|\beta_\rho\beta_\nu|\left(d + \frac{1}{\gamma_\nu} + \frac{1}{\theta_\nu}\right)\left[\rho^2\left(\sigma\cos\sigma d + \frac{\Delta}{i}\sin\sigma d\right)^2 + \sigma^2\left(\sigma\sin\sigma d - \frac{\Delta}{i}\cos\sigma d\right)^2\right]\right\}^{-\frac{1}{2}}. \tag{71}$$

The coupling coefficient of the guided TE mode ν to the TE radiation mode ρ in the range $(n_2^2 - n_3^2)^{\frac{1}{2}}k \leqq \rho < \infty$ is given by

$$K_{j\rho\nu}^{(\pm,p)} = -\frac{(n_1^2 - n_2^2)^{\frac{1}{2}} k\kappa_\nu}{(\beta_\rho^{(\pm)} - \beta_\nu^{(p)})\left[\pi|\beta_\rho\beta_\nu|\left(d + \frac{1}{\gamma_\nu} + \frac{1}{\theta_\nu}\right)\right]^{\frac{1}{2}}} \cdot \frac{\frac{df}{dz} - \frac{\sin\kappa_\nu d}{|\sin\kappa_\nu d|}\left(\frac{n_1^2 - n_3^2}{n_1^2 - n_2^2}\right)^{\frac{1}{2}}(\cos\sigma d - F_j\sin\sigma d)\frac{dh}{dz}}{\left[\frac{\Delta}{\rho}(\cos\sigma d - F_j\sin\sigma d)^2 + \frac{\sigma^2}{\rho\Delta}(\sin\sigma d + F_j\cos\sigma d)^2 + 1 + \frac{\sigma^2}{\rho^2}F_j^2\right]^{\frac{1}{2}}}. \tag{72}$$

The factors F_1 and F_2 are obtained from (50). The radiation modes do not all propagate along the z-axis. Propagating modes are confined to the range $0 \leqq \rho \leqq n_2 k$.

The reader should not be startled by the fact that the coupling coefficients (70) have the dimension m^{-1} while the coupling coefficients (71) and (72) have the dimension $m^{-\frac{1}{2}}$. The different dimensions are attributable to the fact that the coupling coefficients between guided modes occur under a summation sign while the coupling coefficients that describe coupling to radiation modes occur under an integral sign. The integration is performed with respect to ρ, the dimension of which is m^{-1}. The product $c_\rho K_{\rho\nu} d\rho$ has the dimension m^{-1} in agreement with the dimension of the coupling coefficients for guided modes.†

Finally, we list the coupling coefficients for the TM modes. Coupling between guided TM modes is described by the coupling coefficient ($p = +$ or $-$)

$$K_{\mu\nu}^{(\pm,p)} = -\frac{(n_1^2 - n_2^2)D_{g\nu}D_{g\mu}}{4P(\beta_\mu^{(\pm)} - \beta_\nu^{(p)})\omega\epsilon_0 n_1^2 n_2^4}\left\{(n_2^2\beta_\mu^{(\pm)}\beta_\nu^{(p)} + n_1^2\gamma_\nu\gamma_\mu)\frac{df}{dz}\right.$$
$$- \frac{\sin\kappa_\nu d\sin\kappa_\mu d}{|\sin\kappa_\nu d\sin\kappa_\mu d|}\,\frac{n_1^2 - n_3^2}{n_1^2 - n_2^2}\left[\frac{(n_2^4\kappa_\nu^2 + n_1^4\gamma_\nu^2)(n_2^4\kappa_\mu^2 + n_1^4\gamma_\mu^2)}{(n_3^4\kappa_\nu^2 + n_1^4\theta_\nu^2)(n_3^4\kappa_\mu^2 + n_1^4\theta_\mu^2)}\right]^{\frac{1}{2}}$$
$$\left.\cdot(n_3^2\beta_\mu^{(\pm)}\beta_\nu^{(p)} + n_1^2\theta_\nu\theta_\mu)\frac{dh}{dz}\right\}. \tag{73}$$

The coupling coefficients for the TM modes are far more complicated

† Note that $\int |c_\rho|^2 d_\rho$ is dimensionless so that the dimension of c_ρ is $m^{\frac{1}{2}}$.

than the corresponding coefficients for the TE modes. To simplify the notation, we did not substitute expression (59) for the mode amplitude into (73).

The coefficient for coupling from a guided TM mode to a TM radiation mode in the range $0 \leqq \rho \leqq (n_2^2 - n_2^3)^{\frac{1}{2}}k$ is

$$K_{\rho\nu}^{(\pm,p)} = -\frac{(n_1^2 - n_2^2)D_{g\nu}D_{r\rho}}{4P(\beta_\rho^{(\pm)} - \beta_\nu^{(p)})\omega\epsilon_0 n_1^2 n_2^2}\left\{\left(\beta_\rho^{(\pm)}\beta_\nu^{(p)} - \sigma\gamma_\nu\frac{G_{r\rho}}{D_{r\rho}}\right)\frac{df}{dz}\right.$$
$$-\frac{\sin\kappa_\nu d}{|\sin\kappa_\nu d|}\frac{n_1^2 - n_3^2}{n_1^2 - n_2^2}\left[\frac{n_2^4\kappa_\nu^2 + n_1^4\gamma_\nu^2}{n_3^4\kappa_\nu^2 + n_1^4\theta_\nu^2}\right]^{\frac{1}{2}}\left[\beta_\rho^{(\pm)}\beta_\nu^{(p)}\left(\cos\sigma d - \frac{G_{r\rho}}{D_{r\rho}}\sin\sigma d\right)\right.$$
$$\left.\left. + \sigma\theta_\nu\left(\sin\sigma d + \frac{G_{r\rho}}{D_{r\rho}}\cos\sigma d\right)\right]\frac{dh}{dz}\right\}. \quad (74)$$

The amplitude coefficients $D_{g\nu}$, $D_{r\rho}$ and $G_{r\rho}$ are obtained from (59), (63) and (64).

The coefficient for coupling from a guided TM mode to a TM radiation mode in the range $(n_2^2 - n_3^2)^{\frac{1}{2}}k \leqq \rho \leqq \infty$ is given by

$$K_{i\rho\nu}^{(\pm,p)} = -\frac{(n_1^2 - n_2^2)D_{g\nu}S_{r\rho}}{4P(\beta_\rho^{(\pm)} - \beta_\nu^{(p)})\omega\epsilon_0 n_1^2 n_2^2}\left\{(\beta_\rho^{(\pm)}\beta_\nu^{(p)} - \sigma\gamma_\nu R_i)\frac{df}{dz}\right.$$
$$-\frac{\sin\kappa_\nu d}{|\sin\kappa_\nu d|}\frac{n_1^2 - n_3^2}{n_1^2 - n_2^2}\left[\frac{n_2^4\kappa_\nu^2 + n_1^4\gamma_\nu^2}{n_3^4\kappa_\nu^2 + n_1^4\theta_\nu^2}\right]^{\frac{1}{2}}\left[\beta_\rho^{(\pm)}\beta_\nu^{(p)}(\cos\sigma d - R_i\sin\sigma d)\right.$$
$$\left.\left. + \sigma\theta_\nu(\sin\sigma d + R_i\cos\sigma d)\right]\frac{dh}{dz}\right\}. \quad (75)$$

The amplitude coefficients $D_{g\nu}$, $S_{r\rho}$, and R_i are obtained from (59), (68), and (69). The index i assumes the values 1 and 2, and corresponds to the two types of radiation modes that are distinguished by the + and − signs in (68). The superscripts, + and −, attached to the coupling coefficients are supposed to indicate whether the modes travel in + or − z-direction.

It can be shown that the coupling coefficients derived in this paper specialize to the correct expressions[4,5] of the symmetric slab waveguide in the limit $n_2 = n_3$, $\gamma = \Delta$, $\rho = \theta$.[†]

[†] Equations (9.5–26) and (9.5–27) of Ref. 5 must be divided by n_2^2, eq. (9.5–31) must, correspondingly, be divided by n_2^4.

IX. CONCLUSIONS

We have collected the formulas for the modes of the asymmetric slab waveguide and have used this information to derive the coupling coefficients between guided modes as well as between guided and radiation modes for the case of very slight core boundary imperfections. Also presented is the general theory of coupled modes of dielectric waveguides and the general formulas for the coupling coefficients. The theory collected in this paper is useful for the description of mode conversion and radiation phenomena. Phenomena such as the grating coupler and the interaction of acoustic waves and guided light waves can readily be treated with the theory presented here. For an application to statistical irregularities of the core boundary, the reader is advised to consult Refs. 5, 6, and 7.

APPENDIX

Evaluation of the Integral (12)

We consider the index distribution of the slab waveguide as being smoothed out to avoid the discontinuity at the core boundary. It is assumed that the index varies only in a direction perpendicular to the core boundary. We use a coordinate system x', z', the axes of which are perpendicular, and parallel to the tangent at a particular point of the core boundary as shown in Fig. 2. In this coordinate system, we assume that the refractive index is of the form

$$n^2 = F(x'). \tag{76}$$

For values of $x' \leqq 0$ we have $F = n_1^2$; for values $x' > \xi$ we have $F = n_2^2$. At the end of our discussion, we let $\xi \rightarrow 0$. The scalar product

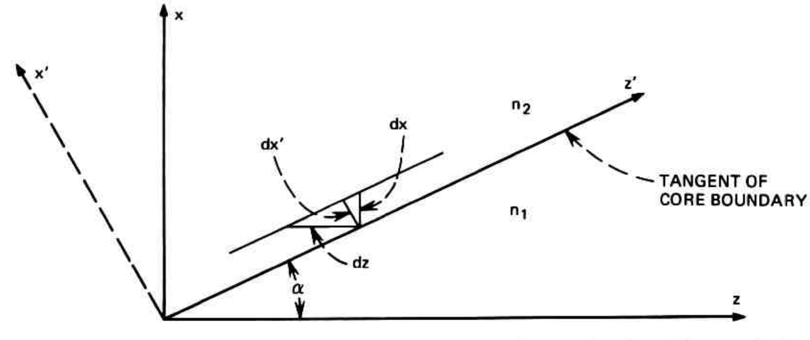


Fig. 2—Sketch of the coordinate systems used for the evaluation of integral (10).

of the electric field vectors can be expressed as

$$\boldsymbol{\mathcal{E}}_\mu^* \cdot \boldsymbol{\mathcal{E}}_\nu = \mathcal{E}_{\mu x'}^* \mathcal{E}_{\nu x'} + \mathcal{E}_{\mu t}^* \mathcal{E}_{\nu t}. \tag{77}$$

The coordinate t indicates a direction parallel to the core boundary. $\mathcal{E}_{\nu t}$ is continuous at the core boundary and can be taken out of the integral. $\mathcal{E}_{\nu x'}$, on the other hand, is discontinuous. We express it in terms of the field just inside of the core and, using (12), write

$$\mathcal{E}_{\nu x'} = \frac{n_1^2}{F(x')} (\mathcal{E}_{\nu x'})_{x'=0}. \tag{78}$$

The scalar product can thus be written in the form

$$\boldsymbol{\mathcal{E}}_\mu^* \cdot \boldsymbol{\mathcal{E}}_\nu = \frac{n_1^4}{F^2(x')} (\mathcal{E}_{\mu x'}^* \mathcal{E}_{\nu x'})_{x'=0} + \mathcal{E}_{\mu t}^* \mathcal{E}_{\nu t}. \tag{79}$$

We thus have to deal with two different integrals. We first consider the integral

$$\int_0^\infty \frac{\partial F(x')}{\partial z} dx = \frac{\partial x'}{dz} \int_0^\infty \frac{\partial F(x')}{\partial x'} dx. \tag{80}$$

We obtain from Fig. 2 the relations

$$\frac{\partial x'}{\partial z} = \sin \alpha \tag{81}$$

and

$$dx = \frac{dx'}{\cos \alpha}. \tag{82}$$

The integral can thus be evaluated

$$\int_0^\infty \frac{\partial F(x')}{\partial z} dx = [F(\xi) - F(0)] \tan \alpha. \tag{83}$$

At the upper core coundary, we have $\tan\alpha = df(z)/dz$, and at the lower core boundary we have $\tan\alpha = dh(z)/dz$. Taking both core boundaries into account, we find with the help of (76),

$$\int_{-\infty}^\infty \frac{\partial n^2}{\partial z} \mathcal{E}_{\mu t}^* \mathcal{E}_{\nu t} dx = -(n_1^2 - n_2^2) \frac{df}{dz} (\mathcal{E}_{\mu t}^* \mathcal{E}_{\nu t})_{x=f} + (n_1^2 - n_3^2) \frac{dh}{dz} (\mathcal{E}_{\mu t}^* \mathcal{E}_{\nu t})_{x=-d+h}. \tag{84}$$

The integral associated with the normal field components is essentially of the form

$$\int_0^\infty \frac{1}{F^2(x')} \frac{\partial F(x')}{\partial z}\, dx = (\tan \alpha) \int_0^\infty \frac{1}{F^2} \frac{\partial F(x')}{\partial x'}\, dx'$$

$$= (\tan \alpha) \frac{F(\xi) - F(0)}{F(0)F(\xi)}. \tag{85}$$

The integral containing the normal field components results in

$$\int_{-\infty}^\infty \frac{\partial n^2}{\partial z} \mathcal{E}^*_{\mu x'} \mathcal{E}_{\nu x'} dx = -(n_1^2 - n_2^2) \frac{n_1^2}{n_2^2} (\mathcal{E}^*_{\mu x'} \mathcal{E}_{\nu x'})_{x=f} \frac{df}{dz}$$

$$+ (n_1^2 - n_3^2) \frac{n_1^2}{n_3^2} (\mathcal{E}^*_{\mu x'} \mathcal{E}_{\nu x'})_{x=-d+h} \frac{dh}{dz}. \tag{86}$$

In (13) we replaced x' with x and t with z. These approximations are valid provided that the inequalities (15) apply. The error is of second order in the derivatives of $f(z)$ or $h(z)$.

REFERENCES

1. Jones, A. L., "Coupling of Optical Fibers and Scattering in Fibers," J. Opt. Soc. *55*, No. 3 (March 1965).
2. Snyder, A. W., "Coupling of Modes on a Tapered Dielectric Cylinder," IEEE Trans. Microwave Theory and Techniques, *MTT-18*, No. 7 (July 1970), pp. 383–392.
3. Synder, A. W., "Mode Propagation in a Nonuniform Cylindrical Medium," IEEE Trans. Microwave Theory and Techniques, *MTT-19*, No. 4 (April 1971), pp. 402–403.
4. Shevchenko, V. V., *Continuous Transitions in Open Waveguides*, Boulder, Colorado: The Golem Press, 1971.
5. Marcuse, D., *Light Transmission Optics*, New York: Van Nostrand Reinhold Company, 1972.
6. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3187–3215.
7. Marcuse, D., "Power Distribution and Radiation Losses in Multimode Dielectric Slab Waveguides," B.S.T.J., *51*, No. 2 (February 1972), pp. 429–454.
8. McKenna, J., "The Excitation of Planar Dielectric Waveguides at *p–n* Junctions, I," B.S.T.J., *46*, No. 7 (September 1967), pp. 1491–1566.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

Batch Input to a Multiserver Queue with Constant Service Times

By A. KUCZURA

(Manuscript received July 26, 1972)

Delay probability formulas for batch input to a finite number of constant-holding-time servers are derived under the assumption of statistical equilibrium. The service-delay distribution (delay until a first request from the batch enters service) is given in terms of the roots of a transcendental equation, while the probability of no service-delay and the average delay are expressed directly in terms of the number of servers, the holding time, and the parameters of the input process. A numerical example with a fixed batch size is discussed.

I. INTRODUCTION

Batch arrivals constitute an important class of input processes in the theory of queues. The investigation of the problem of batch input to a group of constant-holding-time servers was motivated by the existence of installations with multiple Automatic Calling Units (ACU). Customer-based computer equipment controlling the ACU's is capable of originating simultaneous requests. The dial-tone markers, the first common control equipment in a No. 5 Crossbar central office to serve the requests, can be modeled as a group of constant-holding-time servers.

Another example comes from an information transmission system. Messages containing a (small) random number of characters (a batch of characters) arrive according to a Poisson process and must be transmitted to some destination. Delayed messages are stored in a buffer. Since the transmission time per character is usually fixed, this system provides another example of the model studied.

In Section II, the mathematical model used in this study is described and the input process defined; the state equations are written and used to derive the generating function for the equilibrium state probabilities. The probability of no service-delay is found in Section

III, while the average delay is computed in Section IV. The service-delay distribution is given in terms of the roots of a transcendental equation in Section V. A numerical example with a fixed batch size is discussed in Section VI. The effect of this batching scheme on the average delay and the service-delay probability is examined.

II. MATHEMATICAL MODEL

The model studied here is that of a queuing system with a finite number of servers, batch arrivals, and constant holding time. The assumptions are

(*i*) Requests arrive according to a compound Poisson process, that is, requests arrive in groups or batches and the instants at which the batches arrive constitute a Poisson process.

(*ii*) There are c servers and each request has access to any one of them.

(*iii*) All requests have the same constant service time, τ.

(*iv*) The delayed batches wait until service becomes available and are served in order of arrival. The service discipline for requests within a batch is arbitrary, i.e., not specified here.

(*v*) The system is in statistical equilibrium.

Systems with simple Poisson input have been studied as early as 1920, when A. K. Erlang[1] obtained expressions for the probability of delay for arbitrary values of c and the average delay for $c = 1, 2$, and 3. The first complete treatment of such systems was by Pollaczek.[2,3] Crommelin,[4,5] using a method which is simpler than that of Erlang or Pollaczek, also derived general formulas for the probability of delay, the average delay, and the delay distribution. A simplified and concise account of Crommelin's work is given by A. Descloux,[6] who also shows how Pollaczek's formulas can be deduced from Crommelin's results. The development herein is an extension of Crommelin's results to the case of compound Poisson input using the simpler methods employed by Descloux.

We now define the input process. Consider events which happen in groups rather than singly, that is, requests arriving in batches at a group of c servers. For $k = 1, 2, \cdots$, let $N_k(t)$ be a Poisson process with intensity λ_k which governs the arrival of k-sized batches. Assume independence of the processes $N_k(t)$, $k = 1, 2, \cdots$. Let $N(t)$ be the total number of requests that have arrived in the interval $(0, t]$.

Then

$$N(t) = \sum_{k=1}^{\infty} kN_k(t) \tag{1}$$

is called a *compound* Poisson process (Ref. 7, page 271).

The probability that an arriving batch is of size k is equal to λ_k/λ, where

$$\lambda = \sum_{k=1}^{\infty} \lambda_k.$$

From eq. (1), we see that the mean and variance of the number of arrivals per unit time are

$$\mu_1 = \sum_{k=1}^{\infty} k\lambda_k \qquad \text{and} \qquad \mu_2 = \sum_{k=1}^{\infty} k^2\lambda_k, \tag{2}$$

respectively. The generating function of the probability distribution $\pi_n(t) = P\{N(t) = n\}$, $n = 0, 1, 2, \cdots$, is given by

$$\pi(t,z) = \sum_{n=0}^{\infty} \pi_n(t)z^n = e^{t\beta(z)},$$

where

$$\beta(z) = \sum_{n=1}^{\infty} \lambda_n z^n - \lambda,$$

and hence the probabilities $\pi_n(t)$ are given by

$$\pi_n(t) = e^{-\lambda t} \sum_{\mathcal{G}_n} \frac{(\lambda_1 t)^{k_1}(\lambda_2 t)^{k_2}\cdots(\lambda_n t)^{k_n}}{k_1!k_2!\cdots k_n!}, \qquad n = 0, 1, 2, \cdots, \tag{3}$$

where $\mathcal{G}_n$ is the class of all sets of nonnegative integers $\{k_1, k_2, \cdots, k_n\}$ such that $k_1 + 2k_2 + \cdots + nk_n = n$.

The expression given in (3) is not suited for computing $\pi_n(t)$. These probabilities are more conveniently computed from the recurrence relation

$$\begin{aligned} \pi_{k+1}(t) &= \frac{t}{k+1} \sum_{j=0}^{k} (k - j + 1)\lambda_{k-j+1}\pi_j(t), \qquad k = 0, 1, 2, \cdots, \\ \pi_0(t) &= e^{-\lambda t}. \end{aligned} \tag{4}$$

Equation (4) is easily obtained from the relation $k!\pi_k(t) = \pi^{(k)}(t, 0)$, where the superscript denotes differentiation with respect to z.

Special cases of the compound Poisson process are obtained by choosing different convergent sequences of the positive constants

$\lambda_1, \lambda_2, \cdots$. One such sequence is obtained by setting $\lambda_j = \sigma/r^{j-1}$, $r > 1$, $j = 1, 2, \cdots$. This special example has become known as the "stuttering" Poisson process.[8] In this case, a simple expression for $\pi_n(t)$ can be obtained by noting that the generating function has a power series expansion in z, the coefficients of which are given in terms of the Laguerre polynomials L_n, that is,

$$e^{t\beta(z)} = e^{-\lambda t}\left(1 - \frac{z}{r}\right) \sum_{n=0}^{\infty} L_n(-\sigma t r)\left(\frac{z}{r}\right)^n,$$

since

$$\beta(z) = \frac{\sigma z}{1 - z/r} - \lambda \cdot$$

It follows that

$$\begin{aligned} \pi_0(t) &= e^{-\lambda t} \\ \pi_n(t) &= \frac{e^{-\lambda t}}{r^n}\left[L_n(-\sigma t r) - L_{n-1}(-\sigma t r)\right], \qquad n = 1, 2, \cdots. \end{aligned} \tag{5}$$

We will now obtain the equilibrium state equations, and find the probability generating function of the stationary distribution for the general case. Let $X(t)$ be the number of requests (waiting or in service) in the system at time t. Let

$$p_{ij}(t) = P\{X(t) = j \mid X(0) = i\}$$

be the transition probability functions of the process $\{X(t),\ t \geqq 0\}$. It is clear that $X(t)$ is not Markovian. If, however, we examine

$$X_k = X(k\tau), \qquad k = 0, 1, 2, \cdots, \tag{6}$$

we see that this sequence is Markovian and, in fact, $\{X_k, k=0, 1, 2, \cdots\}$ is a homogeneous Markov chain with one-step transition probabilities

$$p_{ij} = P\{X_{k+1} = j \mid X_k = i\}, \qquad k = 0, 1, 2, \cdots,$$

given by

$$p_{ij} = \begin{cases} \pi_j(\tau), & \text{for } 0 \leqq i \leqq c \\ \pi_{j-i+c}(\tau), & \text{for } c < i \leqq j + c \\ 0, & \text{for } j + c < i. \end{cases}$$

We will be interested in the distribution of the number of requests in the system encountered by an arbitrarily arriving batch (a batch arriving at a time point a long way from the origin, i.e., after statistical equilibrium has been reached). But since the instants at which the batches arrive constitute a Poisson process, this distribution is the same as the stationary distribution

$$p_j = \lim_{t\to\infty} p_{ij}(t), \qquad j = 0, 1, 2, \cdots,$$

of the process $\{X(t), t \geqq 0\}$. Moreover, if this limit exists, then so does

$$\lim_{k\to\infty} P\{X_k = j \mid X_0 = i\}$$

and they are equal. Consequently, the distribution of interest to us is given by the stationary distribution of the imbedded Markov chain (6). This distribution is obtained by solving the Chapman-Kolmogorov equations

$$\begin{aligned} p_0 &= \pi_0(\tau)a_c \\ p_n &= \pi_n(\tau)a_c + \sum_{m=c+1}^{n+c} p_m\pi_{n-m+c}(\tau), \qquad n = 1, 2, \cdots, \end{aligned} \tag{7}$$

where

$$a_n = \sum_{m=0}^{n} p_m.$$

We will assume that $\mu_1\tau < c$ so that the stationary distribution $\{p_j\}$ exists.

We need to solve the system (7) for the unknowns p_n. To do this we introduce the probability generating function

$$f(z) = \sum_{n=0}^{\infty} p_n z^n.$$

Multiplying both sides of (7) by z^n and summing over n, we have

$$\begin{aligned} f(z) &= a_c e^{\tau\beta(z)} + \sum_{n=1}^{\infty}\sum_{j=1}^{n} p_{c+j}\pi_{n-j}(\tau)z^n \\ &= a_c e^{\tau\beta(z)} + \sum_{j=1}^{\infty}\sum_{n=j}^{\infty} p_{c+j}\pi_{n-j}(\tau)z^n \\ &= a_c e^{\tau\beta(z)} + \frac{1}{z^c}e^{\tau\beta(z)}[f(z) - g(z)] \end{aligned}$$

where

$$g(z) = \sum_{n=0}^{c} p_n z^n.$$

Thus the probability generating function of the sequence p_n,

$n = 0, 1, \cdots$, is given by

$$f(z) = \frac{g(z) - z^c a_c}{1 - z^c e^{-\tau\beta(z)}}. \tag{8}$$

III. PROBABILITY OF NO SERVICE-DELAY

We say that the service of a batch is delayed if upon its arrival all servers are busy. Hence, the probability of no service-delay will be defined by a_{c-1}, that is, the probability of at most $c - 1$ servers busy. An explicit expression for this probability will now be found.

We start with eq. (8). Since the p_n's are probabilities, $f(z)$ is holomorphic in $|z| \leqq 1$ and, therefore, the zeros of the denominator and numerator in $|z| \leqq 1$ must be the same and have the same multiplicities. We will show that the denominator of (8) has c distinct roots in $|z| \leqq 1$ and that all of them, with the exception of $z = 1$, lie inside the unit circle.

For $|z| = 1 + \delta$, with δ sufficiently small and positive, we have

$$|e^{\tau\beta(z)}| \leqq e^{-\lambda\tau} \exp \tau \left\{ \sum_{k=1}^{\infty} \lambda_k |z|^k \right\} = e^{\tau\mu_1\delta + 0(\delta^2)}$$

where μ_1 is defined by (2).

Since $\tau\mu_1 < c$ by assumption, we have

$$e^{\tau\mu_1\delta+0(\delta^2)} < (1 + \delta)^c = |z|^c,$$

and by Rouché's theorem, the equation

$$e^{\tau\beta(z)} - z^c = 0$$

has exactly c roots within the region $|z| = 1 + \delta$. Let these roots be denoted by $z_1, z_2, \cdots, z_{c-1}, z_c(= 1)$. Then

(*i*) $1, z_1, z_2, \cdots, z_{c-1}$ are distinct.
(*ii*) $|z_n| < 1$ for $n = 1, 2, \cdots, c - 1$.

To prove (*i*), first note that the root $z = 1$ is simple because

$$\lim_{z\to 1} \frac{1 - z^c e^{-\tau\beta(z)}}{z - 1} = \tau\mu_1 - c \neq 0\cdot$$

Similarly, for any root z_i, $i = 1, 2, \cdots, c - 1$,

$$\lim_{z\to z_i} \frac{1 - z^c e^{-\tau\beta(z)}}{z - z_i} = z_i^{c-1} e^{-\tau\beta(z_i)} \left[\tau \sum_{k=1}^{\infty} k\lambda_k z_i^k - c \right]\cdot$$

The first two factors cannot vanish for any admissible choice of the root z_i, so that if z_i is to be a root of second or higher order, we must have

$$\tau(\lambda_1 z_1 + 2\lambda_2 z_2^2 + 3\lambda_3 z_3^3 + \cdots) = c.$$

But this is not possible since

$$\tau|\lambda_1 z_1 + 2\lambda_2 z_2^2 + 3\lambda_3 z_3^3 + \cdots| \leqq \tau\mu_1 < c,$$

and the roots $1, z_1, z_2, \cdots, z_{c-1}$ are therefore all distinct.

To prove (*ii*), suppose that $|z_n| = 1$ for some n, $n = 1, 2, \cdots, c - 1$, then $|\exp[\tau\beta(z_n)]| = 1$ and hence the real part of $\tau\beta(z_n)$ must be zero, that is, $\mathcal{R}[\beta(z_n)] = 0$. Hence we must have

$$\mathcal{R}[\beta(z_n)] = \mathcal{R}\left\{\sum_{k=1}^{\infty} \lambda_k \left(1 - z_n^k\right)\right\} = 0\cdot$$

Since all terms in the sum are nonnegative, we have $\mathcal{R}(1 - z_n^k) = 0$ for all k, and therefore $z_n = 1$, contrary to the assumption. It follows that $|z_n| < 1$ for $n = 1, 2, \cdots, c - 1$.

Since the numerator of (8) is a polynomial of degree c, $f(z)$ has the form

$$f(z) = A\frac{(z-1)(z-z_1)(z-z_2)\cdots(z-z_{c-1})}{1 - z^c e^{-\tau\beta(z)}}. \tag{9}$$

The condition $f(1) = 1$ determines A:

$$A = \frac{\mu_1\tau - c}{(1-z_1)(1-z_2)\cdots(1-z_{c-1})}. \tag{10}$$

In computing a_{c-1} it is convenient to introduce the generating function

$$F(z) = \sum_{n=0}^{\infty} a_n z^n\cdot$$

Then, since $a_n - a_{n-1} = p_n$, $n = 1, 2, \cdots$, we have

$$(1 - z)F(z) = f(z),$$

or

$$F(z) = \frac{f(z)}{1 - z}.$$

Now, making use of (9), we obtain

$$F(z) = -A\frac{(z-z_1)(z-z_2)\cdots(z-z_{c-1})}{1 - z^c e^{-\tau\beta(z)}}.$$

The probability of no service-delay, a_{c-1}, is given by the coefficient of z^{c-1} in the expansion of $F(z)$:

$$a_{c-1} = -A = \frac{c - \mu_1\tau}{(1 - z_1)(1 - z_2)\cdots(1 - z_{c-1})}\cdot \tag{11}$$

An expression for a_{c-1} which does not involve the roots z_i is obtained through an application of the generalized argument-principle (Ref. 9, page 151). That is, suppose $\psi(z)$ is holomorphic and $\phi(z)$ is meromorphic on and inside the contour C. Let α_k, $k = 1, 2, \cdots$, be the zeros with multiplicities r_k, and β_k, $k = 1, 2, \cdots$, the poles with multiplicities s_k of the function $\phi(z)$ inside C. Then the generalized argument-principle states that

$$\frac{1}{2\pi i}\int_C \psi(z)\frac{\phi'(z)}{\phi(z)}dz = \sum_k r_k\psi(\alpha_k) - \sum_k s_k\psi(\beta_k)\cdot$$

Taking the logarithm of eq. (11), we have

$$\log a_{c-1} = \log\,(c - \mu_1\tau) - \sum_{i=1}^{c-1} \log\,(1 - z_i)\cdot$$

We will eliminate the roots z_i from the second term of the right-hand side of the preceding equation. Let

$$\phi(z) = e^{\tau\beta(z)} - z^c,$$

and note that $\phi(z)$ has simple zeros at $z = z_1, z_2, \cdots, z_{c-1}$. Choose $\psi(z)$ as the principal branch of $\log\,(1 - z)$. The generalized argument-principle yields

$$\sum_{n=1}^{c-1} \log\,(1 - z_n) = \frac{1}{2\pi i}\int_C \log\,(1 - z)d[\log \phi(z)] = J$$

where C is the contour $|z| = 1 - \epsilon$ and $\epsilon(>0)$ is chosen so that z_n, $n = 1, 2, \cdots, c - 1$, lie inside C but $z_c(=1)$ is exterior to C. We will now show that

$$J = \log\,(c - \tau\mu_1) + \sum_{n=1}^{\infty}\frac{1}{n}\sum_{j=nc}^{\infty} \pi_j(n\tau)\,,$$

and hence

$$\log a_{c-1} = -\sum_{n=1}^{\infty}\frac{1}{n}\sum_{j=nc}^{\infty} \pi_j(n\tau)\cdot \tag{12}$$

Note first that the principal branch of $\log\,(1 - z)\,\dfrac{d}{dz}\,[z^{c-1}\log\,(1 - z)]$

is holomorphic in $|z| \leqq 1 - \epsilon$. Since its integral on C is zero, we have

$$J = \frac{1}{2\pi i} \int_C \log (1 - z) d \left[\log \frac{1 - z^{-c} e^{\tau \beta(z)}}{1 - z^{-1}} \right].$$

Integration by parts yields

$$J = -\frac{1}{2\pi i} \int_C \log \left[\frac{1 - z^{-c} e^{\tau \beta(z)}}{1 - z^{-1}} \right] \frac{dz}{z - 1}.$$

The integrand above has a simple pole at $z = 1$, and its residue there is equal to $\log (c - \mu_1 \tau)$. Choose $\delta (>0)$ such that the only zeros of $\phi(z)$ in the disk $|z| \leqq 1 + \delta$ are $1, z_1, z_2, \cdots, z_{c-1}$, and let C_1 be the contour $|z| = 1 + \delta$. Noting that the integral of $\log (1 - z^{-1})/(z - 1)$ on C_1 vanishes, we have

$$J = \log (c - \mu_1 \tau) - \frac{1}{2\pi i} \int_{C_1} \log [1 - z^{-c} e^{\tau \beta(z)}] \frac{dz}{z - 1}.$$

Now since $|z^{-c} e^{\tau \beta(z)}| < 1$ on C_1, the power series for $\log [1 - z^{-c} e^{\tau \beta(z)}]$ converges uniformly on C_1, and termwise integration is allowed, so that

$$\begin{aligned} J &= \log (c - \mu_1 \tau) + \frac{1}{2\pi i} \sum_{n=1}^{\infty} \frac{1}{n} \int_{C_1} e^{\tau n \beta(z)} \frac{z^{-nc}}{z - 1} dz \\ &= \log (c - \mu_1 \tau) + \sum_{n=1}^{\infty} \frac{1}{n} \left[1 - \frac{1}{2\pi i} \sum_{j=0}^{\infty} \int_{C_1} e^{\tau n \beta(z)} z^{-nc+j} dz \right]. \end{aligned}$$

Expanding the integrand in powers of z, and integrating term by term, we see that the integral is zero for all terms except one, and that there it is equal to $2\pi i$ times the coefficient of z^{-1}. But the coefficient of z^{-1} is exactly $\pi_{nc-j-1}(n\tau)$, so that

$$\begin{aligned} J &= \log (c - \mu_1 \tau) + \sum_{n=1}^{\infty} \frac{1}{n} \left[1 - \sum_{j=0}^{nc-1} \pi_{nc-j-1}(n\tau) \right] \\ &= \log (c - \mu_1 \tau) + \sum_{n=1}^{\infty} \frac{1}{n} \sum_{j=nc}^{\infty} \pi_j(n\tau), \end{aligned}$$

and this is the result stated earlier.

For the case of a simple Poisson input, we have $\lambda_1 = \lambda$, $\lambda_j = 0$, $j = 2, 3, \cdots$, and (12) reduces to [Crommelin, Ref. 5, eq. (5)]

$$\log a_{c-1} = -\sum_{n=1}^{\infty} \frac{1}{n} \sum_{j=nc}^{\infty} \frac{(\lambda n \tau)^j}{j!} e^{-\lambda n \tau}.$$

Note that this expression and eq. (12) differ only in the terms $\pi_j(n\tau)$ and $((\lambda n\tau)^j/j!)e^{-\lambda n\tau}$, which represent the same probabilities in two different systems: both are equal to the probability of exactly j arrivals in the time interval $n\tau$. As we shall see later, these probabilities appear again in the expressions for the average delay and the service-delay distribution.

For "stuttering" Poisson input, eq. (12) reduces to

$$\log a_{c-1} = -\sum_{n=1}^{\infty} \frac{e^{-[\sigma n\tau r/(r-1)]}}{n} \sum_{j=nc}^{\infty} \frac{1}{r^j} [L_j(-\sigma n\tau r) - L_{j-1}(-\sigma n\tau r)]$$

where the $L_n(\xi)$ are Laguerre polynomials.

IV. AVERAGE DELAY

Under equilibrium conditions, the average delay D is equal to the average amount of waiting per unit of time divided by the average number of arrivals per unit of time. The average amount of waiting per unit of time is equal to

$$\sum_{n=c+1}^{\infty} (n-c)p_n ,$$

so that D is given by

$$D = \frac{1}{\mu_1} \sum_{n=c+1}^{\infty} (n-c)p_n = \frac{1}{\mu_1} \sum_{n=0}^{\infty} np_n - \tau$$

where μ_1 is the average number of arrivals per unit of time defined by (2). An explicit expression for D can be readily obtained by noting that

$$\sum_{n=0}^{\infty} np_n = \lim_{z \to 1} \left[\frac{d}{dz} f(z) \right], \qquad |z| < 1 ,$$

where $f(z)$ is the generating function given by (8). Straightforward differentiation of (8) and application of L'Hospital's rule lead to

$$\frac{D}{\tau} = \frac{1}{\mu_1\tau} \sum_{i=1}^{c-1} \frac{1}{1-z_i} + \frac{\mu_2\tau + \mu_1\tau(\mu_1\tau - 1) - c(c-1)}{2\mu_1\tau(c-\mu_1\tau)} \tag{13}$$

where $z_1, z_2, \cdots, z_{c-1}$ are the roots defined previously. Again we wish to eliminate these roots. The method used in the previous section suggests the application of the generalized argument-principle with

$(1 - z)^{-1}$ as $\psi(z)$, and $\phi(z)$ as before. Thus we have

$$\sum_{n=1}^{c-1} \frac{1}{1 - z_n} = \frac{1}{2\pi i} \int_C \frac{1}{1 - z} \frac{\phi'(z)}{\phi(z)} \, dz = K. \tag{14}$$

Noting that the integral of $(1 - z)^{-1} \dfrac{d}{dz} \{\log [z^{c-1}(1 - z)]\}$ on C is equal to $(c - 1)2\pi i$ (the residue at the simple pole $z = 0$ is $c - 1$), eq. (14) becomes

$$K = (c - 1) + \frac{1}{2\pi i} \int_C \frac{1}{(1 - z)} \, d\left[\log \frac{1 - z^{-c}e^{\tau\beta(z)}}{1 - z^{-1}}\right].$$

Integration by parts yields

$$K = (c - 1) - \frac{1}{2\pi i} \int_C \log \left[\frac{1 - z^{-c}e^{\tau\beta(z)}}{1 - z^{-1}}\right] \frac{dz}{(1 - z)^2}. \tag{15}$$

The integrand in (15) has a pole of second order at $z = 1$, and its residue there is found to be

$$\frac{c(c - 1) - \tau\mu_1(\tau\mu_1 - 1) - \tau\mu_2}{2(c - \tau\mu_1)} - (c - 1).$$

Consequently,

$$K = - \frac{c(c - 1) - \tau\mu_1(\tau\mu_1 - 1) - \tau\mu_2}{2(c - \tau\mu_1)} - \frac{1}{2\pi i} \int_{C_1} \log \left[\frac{1 - z^{-c}e^{\tau\beta(z)}}{1 - z^{-1}}\right] \frac{dz}{(1 - z)^2}. \tag{16}$$

Combining (14) and (16) and noting that the integral of

$$(1 - z)^{-2} \log (1 - z^{-1})$$

vanishes on C_1, we obtain

$$\frac{D}{\tau} = - \frac{1}{2\pi i\mu_1} \int_{C_1} \log[1 - z^{-c}e^{\tau\beta(z)}] \frac{dz}{(1 - z)^2}.$$

Recall now that $|z^{-c}e^{\tau\beta(z)}| < 1$ on C_1, and hence $\log [1 - z^{-c}e^{\tau\beta(z)}]$ has a uniformly convergent power series representation in $z^{-c}e^{\tau\beta(z)}$ on C_1, so that

$$\frac{D}{\tau} = \frac{1}{\mu_1\tau} \frac{1}{2\pi i} \sum_{n=1}^{\infty} \frac{1}{n} \int_{C_1} \frac{e^{n\tau\beta(z)}z^{-nc}}{(1 - z)^2} \, dz. \tag{17}$$

The integrand in (17) has poles at $z = 0$ and $z = 1$ with residues

$$\sum_{k=0}^{nc-1} (nc - k)\pi_k(n\tau)$$

and

$$n(\mu_1\tau - c),$$

respectively, giving us the final result

$$\frac{D}{\tau} = \frac{1}{\mu_1\tau} \sum_{n=1}^{\infty} \frac{1}{n} \sum_{k=1}^{\infty} k\pi_{nc+k}(n\tau). \tag{18}$$

V. SERVICE-DELAY DISTRIBUTION

For the purpose of obtaining the service-delay distribution (delay until a first request from the batch enters service), let us define $g_m(t)$ as the probability that among the requests present at some time t_0, at most m of them are still in the system at time $t_0 + t$. Considering the state corresponding to $g_{mc+c-1}(t)$, we see that, at most, $mc + c - 1$ of the requests preceding the given batch will be in progress at time t later, and consequently, at most, $c - 1$ of them at time $m\tau + t$ later. This is the condition for the service-delay d to be less than $m\tau + t$, or in symbols

$$P\{d < m\tau + t\} = g_{mc+c-1}(t), \qquad 0 \leqq t < \tau. \tag{19}$$

To determine $g_m(t)$, we introduce the generating function

$$G(z,t) = \sum_{m=0}^{\infty} g_m(t)z^m.$$

Upon noting that $\sum_{m=0}^{n} \pi_{n-m}(t)g_m(t) = a_n$, we have

$$\begin{aligned} G(z,t) &= e^{-t\beta(z)} \sum_{m=0}^{\infty} g_m(t)z^m \sum_{n=0}^{\infty} \pi_n(t)z^n \\ &= e^{-t\beta(z)} \sum_{n=0}^{\infty} z^n \sum_{m=0}^{\infty} \pi_{n-m}(t)g_m(t) \\ &= e^{-t\beta(z)} \sum_{n=0}^{\infty} a_n z^n = e^{-t\beta(z)}F(z). \end{aligned}$$

Substituting for $F(z)$ we obtain

$$G(z,t) = -A \frac{(z - z_1)(z - z_2)\cdots(z - z_{c-1})e^{-t\beta(z)}}{1 - z^c e^{-\tau\beta(z)}}. \tag{20}$$

A direct expansion of the right-hand side of (20) in powers of z is not desirable because the coefficients of z involve sums the terms of which have alternate signs and, therefore, are not well suited for computation. To circumvent this difficulty, we first obtain a Laurent series expansion of $G(z,t)$ in the annulus $1 < |z| < |\xi|$, where ξ is that root of $z^c \exp[-\tau\beta(z)] - 1 = 0$ which has the smallest modulus exceeding 1. The existence of such a root can be proved as follows. Since $x^c \exp[-\tau\beta(x)]$ takes on the value 1 at $x = 1$ (x = real part of z), has a positive derivative there, and vanishes at infinity, the equation $z^c \exp[-\tau\beta(z)] - 1 = 0$ has at least one root outside the unit circle.

For $1 < |z| < |\xi|$, the absolute value of $z^{-c} \exp[\tau\beta(z)]$ is less than unity. Expanding the denominator in powers of $z^{-c} \exp[\tau\beta(z)]$ and the exponential function in powers of z, and collecting like-power terms, we obtain, for $1 < |z| < |\xi|$,

$$G(z,t) = \sum_{k=0}^{\infty} \left\{ \sum_{n=0}^{c-1} q_n \sum_{j=0}^{\infty} \pi_{cj+c+k-n}[(j+1)\tau - t] \right\} z^k \qquad (21)$$

where q_n is the coefficient of z^n in the polynomial

$$A(z - z_1)(z - z_2)\cdots(z - z_{c-1}).$$

Since $z = 1$ is the only singularity of $G(z,t)$ in $|z| < |\xi|$ (a simple pole with residue -1), $G(z,t) + (z-1)^{-1}$ is holomorphic in $|z| < |\xi|$ and hence for $|z| < |\xi|$

$$G(z,t) + \frac{1}{z-1} = \text{expansion (21)} + \sum_{n=1}^{\infty} z^{-n}.$$

Therefore, for $|z| < 1$, we must have

$$G(z,t) = \sum_{n=0}^{\infty} z^n + \text{expansion (21)} + \sum_{n=1}^{\infty} z^{-n}.$$

From this equation, we obtain the service-delay distribution

$$P\{d < m\tau + t\} = 1 - \sum_{n=0}^{c-1} q_n \sum_{j=0}^{\infty} \pi_{(m+j+2)c-n-1}[(j+1)\tau - t],$$
$$m = 0, 1, 2, \cdots, \qquad 0 \leqq t < \tau. \qquad (22)$$

VI. A NUMERICAL EXAMPLE

We examine a fixed-size batching scheme which provides some insight into the effect of batch arrivals on the average delay and the probability of service-delay. Suppose customers arrive in batches of size m. Then $\lambda_j = 0$ for $j \neq m$, $\mu_1 = m\lambda_m$ and

$$\pi_j(t) = \begin{cases} e^{-\lambda_m t} \dfrac{(\lambda_m t)^k}{k!}, & \text{for} \quad j = km \\ 0 \quad\quad\quad\quad\quad , & \text{otherwise.} \end{cases}$$

Figure 1 shows the average delay experienced for an arbitrary customer as a function of the occupancy $\rho = \tau\mu_1/c$, for various values of m. We assume that the holding time is unity, and that $c = 4$.

Equation (18) written in the form

$$D = \frac{1}{\mu_1} \sum_{n=1}^{\infty} \frac{1}{n} \left\{ n(\mu_1 - c + ce^{-n\lambda_m}) + \sum_{j=1}^{nc-1} (nc - j)\pi_j(n) \right\}$$

was used to obtain the curves drawn in Fig. 1. We might point out here that the above series converges slowly when the occupancy is near unity. In the interest of speedy computation it may be necessary to solve for the roots of the denominator in eq. (8) and then use (13) to calculate the average delay. The same remarks apply to eq. (12) which is used to compute the probability of no delay.

Because holding times are constant, several interesting phenomena are observed. First, if the batch size is an integer multiple of the number of servers, say $m = kc$, then the mean time until the service of an arriving batch (or the first customer from the batch) begins is the same as the average delay in a one-server system with single Poisson arrivals and holding time k. From the Pollaczek-Khintchine formula, this number is given by

$$\frac{k\rho}{2(1 - \rho)}.$$

Hence the mean delay which an arbitrary customer experiences is the average of the above number and the mean delay experienced by the last customer in the batch to be served. Thus we have

$$D = \frac{k - 1}{2} + \frac{\rho k}{2(1 - \rho)} \qquad (m = kc).$$

Note that as $\rho \rightarrow 0$, $D \rightarrow (k - 1)/2$.

On the other hand, if the number of servers is an integer multiple of the batch size, say $c = jm$, then the system may be viewed as a

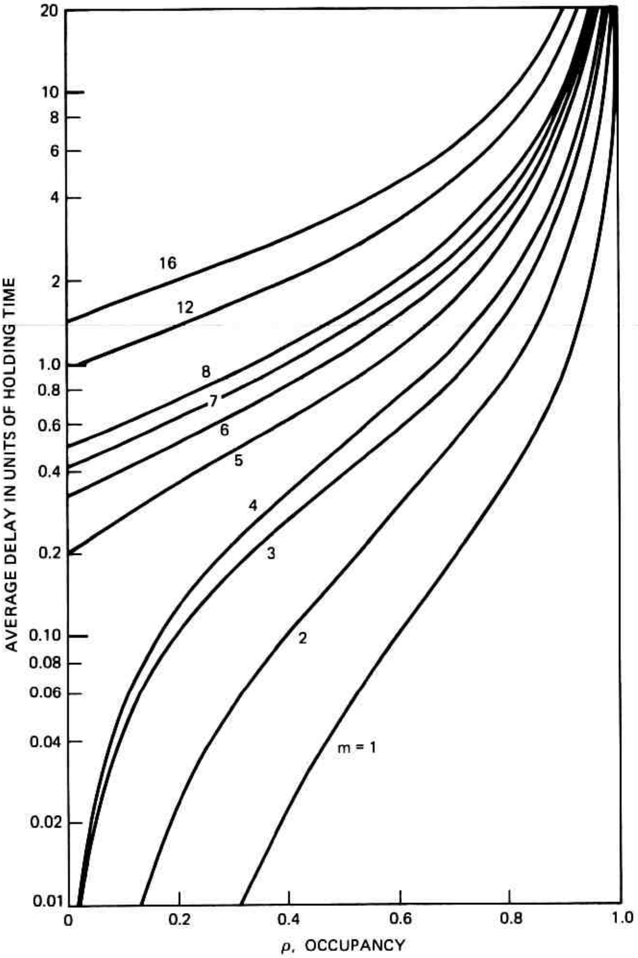


Fig. 1—Average delay in an M/D/4 queue when arrivals occur in batches of size m.

collection of single-server systems with constant holding time and j-phased Erlangian input of mean offered load μ_1/j. This can be seen by imagining that the sets of m servers required to serve the arriving batches are chosen in cyclic order.

Figure 2 shows the probability of service-delay (the probability that the service of an arriving batch is delayed) as a function of the occupancy, for various values of the batch size m and $c = 4$. From eq. (12) we obtained and used the following expression for the service-

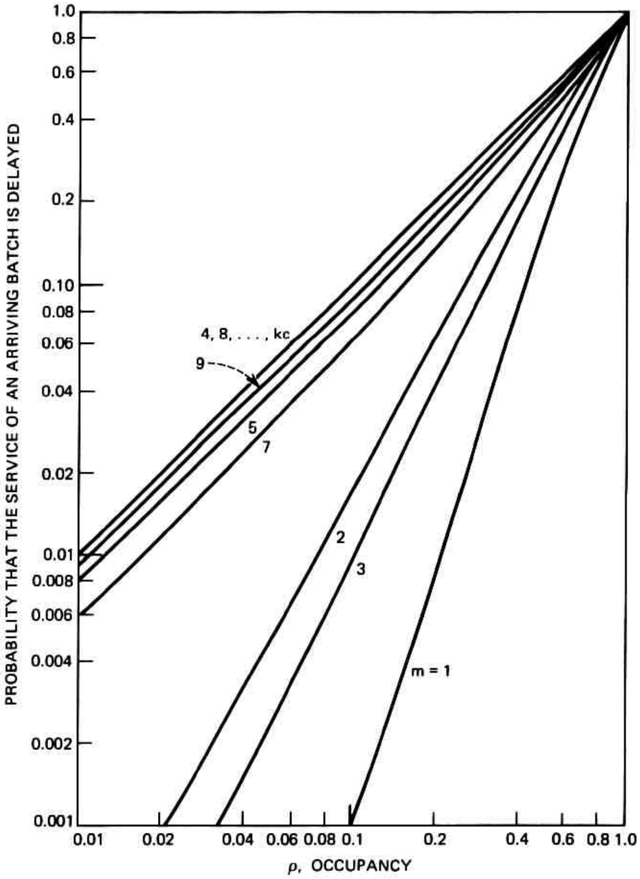


Fig. 2—Probability of service-delay in an M/D/4 queue when arrivals occur in batches of size m.

delay probability:

$$1 - \exp \left\{ - \sum_{n=1}^{\infty} \frac{1}{n} \left[1 - \sum_{j=0}^{nc-1} \pi_j(n) \right] \right\} \cdot$$

Phenomena similar to those observed in Fig. 1 exist here also. For example, if the batch size is an integer multiple of c, say $m = kc$, then the probability of service-delay is the same as the probability of delay in a one-server system with single Poisson arrivals, i.e., it is simply ρ.

On the other hand, if $c = jm$, then the service-delay probability is the same as would be found in a system with single Poissonian arrivals of intensity λ_m and j constant-holding-time servers.

VII. ACKNOWLEDGMENT

I wish to thank A. Descloux of Bell Laboratories for suggesting the problem and for helpful discussions.

REFERENCES

1. Brockmeyer, E., Halstrom, M. L., and Jensen, Arne, *The Life and Works of A. K. Erlang*, Copenhagen: The Copenhagen Telephone Company, 1948.
2. Pollaczek, F., "Über eine Aufgabe der Wahrscheinlichkeitstheorie I," Mathematische Zeitschrift, *32*, 1930, pp. 64–100.
3. Pollaczek, F., "Über eine Aufgabe der Wahrscheinlichkeitstheorie II," Mathematische Zeitschrift, *32*, 1930, pp. 729–750.
4. Crommelin, C. D., "Delay Probability Formulae When the Holding Times Are Constant," P.O. Elec. Engrs. J., *25*, 1932, pp. 41–50.
5. Crommelin, C. D., "Delay Probability Formulae," P.O. Elec. Engrs. J., *26*, 1934, pp. 266–274.
6. Descloux, A., "Delay Systems With Random Input and Constant Holding Time," unpublished work.
7. Feller, W., *An Introduction to Probability Theory and Its Applications*, *1*, 2nd Edition, New York: Wiley, 1957.
8. Galliher, H. P., Morse, P. M., and Simond, M., "Dynamics of Two Classes of Continuous-Review Inventory Systems," Opns. Res., *7*, 1959, p. 362.
9. Ahlfors, L. V., *Complex Analysis*, 2nd Edition, New York: McGraw-Hill, 1966.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

Crosstalk in Uniformly Coupled Lossy Transmission Lines

By J. C. ISAACS, JR., and N. A. STRAKHOV

(Manuscript received July 5, 1972)

The crosstalk between two identical, uniformly coupled, lossy transmission lines is examined. Equations are derived which can be solved to obtain formulas for the near-end crosstalk (NEXT) and far-end crosstalk (FEXT). An example is worked which illustrates the mutual influence of the two lines in terms of the modal voltages and currents. The mutual influence of the two lines is also studied by comparing the results of this example with the "classical" crosstalk formulas which assume weak coupling and neglect the influence of the disturbed line on the disturbing line. It is shown that the influence of the disturbed line on the disturbing line can be neglected for NEXT for most weak coupling situations. For sufficiently high frequencies and/or long line lengths, however, this influence cannot be neglected for FEXT.

I. INTRODUCTION

One of the earliest analyses of crosstalk in coupled transmission lines was made by Campbell;[1] later Shelkunoff and Odarenko[2] used a similar method to analyze the crosstalk in coaxial structures. These "classical" formulas were derived for two parallel transmission lines with weak coupling and matched terminations. One drawback of these analyses is that they do not take into account the effect of the disturbed line on the disturbing line. However, their crosstalk formulas are simple in form and easy to analyze. Also, they are applicable to any parallel, uniformly coupled transmission lines with weak coupling.

Somewhat later an analysis of coupled transmission lines was made by Rice.[3] His results apply under quite general conditions and are expressed in compact matrix notation. However, his results have apparently not influenced current analyses, possibly because the formulas are more complicated to analyze than those in Refs. 1 and 2. Coupling between two pairs under similarly general conditions is given by Kuznetsov and Stratonovich.[4] Although the emphasis is in obtaining results for the time domain, the basic approach is similar to the one we will follow. The specific time domain results do not apply to the transmission lines of interest here because the frequency dependence of the primary constants is not taken into account. More recent analyses[5,6] have relaxed the assumptions of weak coupling and matched lines and do take into account the effect of the disturbed line on the disturbing line. Unfortunately, these analyses focus attention on the lossless case in order to obtain crosstalk formulas which can be readily calculated. While the lossless case may be of interest for some line lengths and frequency ranges, it does not cover many applications which are of great practical interest.

In this paper, a fairly general analytical model is presented for two identical, parallel, uniformly coupled transmission lines with a common ground return. This model does not assume weak coupling, matched terminations, or lossless lines. The resultant crosstalk equations, although somewhat unwieldy, can be evaluated with the aid of a computer.

The motivation for this study was, in part, to assist in the analysis of special cables being utilized in the interconnection of equipment racks. These cables, referred to as flat flexible cables, have conductors that are not twisted and therefore can couple to each other strongly under certain conditions. The results of this study are also of interest to those studying longitudinal mode coupling effects in multipair cable.

II. DERIVATION OF CROSSTALK BETWEEN TRANSMISSION LINES WITH ARBITRARY CONSTANT COUPLING

The starting point for this analysis is the set of coupled differential equations which are assumed to govern the two transmission lines. They are

$$\frac{dE_1}{dx} = -(R + j\omega L)I_1 - j\omega L_c I_2 \tag{1a}$$

$$\frac{dI_1}{dx} = -(G + j\omega C)E_1 - j\omega C_c E_2 \tag{1b}$$

$$\frac{dE_2}{dx} = -(R + j\omega L)I_2 - j\omega L_c I_1 \tag{1c}$$

$$\frac{dI_2}{dx} = -(G + j\omega C)E_2 - j\omega C_c E_1 \tag{1d}$$

where

E_i is the voltage across transmission line i, $i = 1, 2$

I_i is the current flowing in transmission line i, $i = 1, 2$

$\left.\begin{matrix} R \\ L \\ G \\ C \end{matrix}\right\}$ are standard distributed resistance, inductance, conductance, and capacitance, respectively

ω is frequency in radians/s

$\left.\begin{matrix} L_c \\ C_c \end{matrix}\right\}$ are "coupling" inductance and capacitance; the relationship to physical quantities will be derived in a later section.

A number of assumptions are tacitly implied in order for the equations to describe the physical situation. These will now be discussed.

The first and most basic assumption is that only two sets of voltages and currents are involved in the coupling mechanism. This assumption is readily met in the case of unbalanced transmission lines shown in Fig. 1a. However, for balanced transmission lines, depicted in Fig. 1b, other voltages and currents may play a role in the coupling mechanism. They will only be negligible if each transmission line is well balanced with respect to ground.

Another important assumption is that the power propagating down the transmission lines is essentially described by TEM modes. This assumption is required to assure that the telegrapher's equations [i.e., (1) with $L_c = C_c = 0$] are valid.

For the time being, it is not necessary to specify whether or not R, L, G, C, C_c, and L_c are frequency independent. However, if these results are translated from the frequency domain to the time domain, the frequency dependence of these parameters will have to be specified.

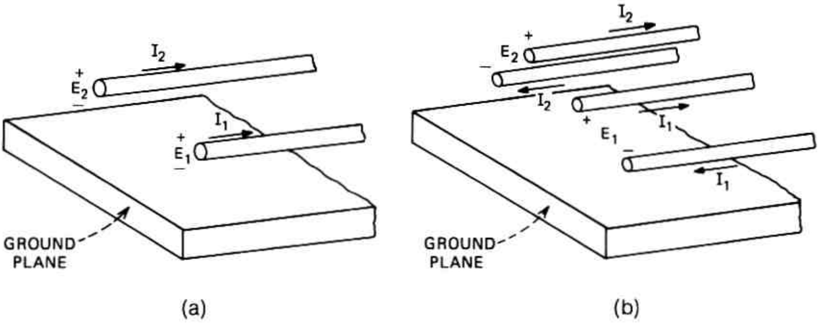


Fig. 1—(a) Unbalanced transmission lines. (b) Balanced transmission lines.

Of course, it is a fundamental assumption of this analysis that the six parameters are independent of x.

Differentiation of (1a) and (1c) with respect to x and substitution of (1b) and (1d) for the appropriate quantities result in

$$\frac{d^2E_1}{dx^2} = A_{11}E_1 + A_{12}E_2 \tag{2a}$$

$$\frac{d^2E_2}{dx^2} = A_{12}E_1 + A_{22}E_2 \tag{2b}$$

where

$$A_{11} = (R + j\omega L)(G + j\omega C) - \omega^2 L_c C_c \tag{3}$$

$$A_{22} = A_{11} \tag{4}$$

$$A_{12} = (R + j\omega L) j\omega C_c + (G + j\omega C) j\omega L_c. \tag{5}$$

Assuming a solution of the form $E_1 = A_1 e^{\gamma x}$ and $E_2 = A_2 e^{\gamma x}$ for (2) yields

$$\gamma = \pm\gamma^+ \quad \text{or} \quad \pm\gamma^-$$

where

$$\begin{aligned}\gamma^+ &= \sqrt{A_{11} + A_{12}} \\ &= \{[R + j\omega(L + L_c)][G + j\omega(C + C_c)]\}^{\frac{1}{2}}\end{aligned} \tag{6}$$

$$\begin{aligned}\gamma^- &= \sqrt{A_{11} - A_{12}} \\ &= \{[R + j\omega(L - L_c)][G + j\omega(C - C_c)]\}^{\frac{1}{2}}\end{aligned} \tag{7}$$

and

$$A_2 = \begin{cases} A_1 & \text{if} \quad \gamma = \pm\gamma^+ \\ -A_1 & \text{if} \quad \gamma = \pm\gamma^-. \end{cases}$$

Therefore, the general solutions for $E_1(x)$ and $E_2(x)$ are expressed as

$$E_1(x) = A^+e^{\gamma^+x} + A^-e^{\gamma^-x} + B^+e^{-\gamma^+x} + B^-e^{-\gamma^-x} \tag{8a}$$

$$E_2(x) = A^+e^{\gamma^+x} - A^-e^{\gamma^-x} + B^+e^{-\gamma^+x} - B^-e^{-\gamma^-x} \tag{8b}$$

where the four constants A^+, A^-, B^+, and B^- will be determined from boundary conditions. The corresponding expressions for the two currents can be obtained by solving (1a) and (1c). After the required algebraic manipulations, one obtains

$$I_1(x) = -\frac{1}{Z^+}A^+e^{\gamma^+x} - \frac{1}{Z^-}A^-e^{\gamma^-x} + \frac{1}{Z^+}B^+e^{-\gamma^+x} + \frac{1}{Z^-}B^-e^{-\gamma^-x} \tag{9a}$$

$$I_2(x) = -\frac{1}{Z^+}A^+e^{\gamma^+x} + \frac{1}{Z^-}A^-e^{\gamma^-x} + \frac{1}{Z^+}B^+e^{-\gamma^+x} - \frac{1}{Z^-}B^-e^{-\gamma^-x} \tag{9b}$$

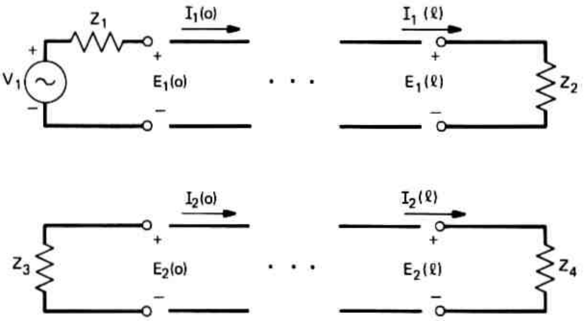


Fig. 2—Boundary conditions imposed on coupled transmission lines.

where

$$Z^+ = \left[\frac{R + j\omega(L + L_c)}{G + j\omega(C + C_c)}\right]^{\frac{1}{2}} \tag{10}$$

$$Z^- = \left[\frac{R + j\omega(L - L_c)}{G + j\omega(C - C_c)}\right]^{\frac{1}{2}}. \tag{11}$$

The boundary conditions that will be imposed are shown in Fig. 2. The corresponding boundary condition equations are:

$$V_1 = Z_1I_1(o) + E_1(o) \tag{12a}$$

$$O = Z_3I_2(o) + E_2(o) \tag{12b}$$

$$O = Z_2I_1(l) - E_1(l) \tag{12c}$$

$$O = Z_4I_2(l) - E_2(l). \tag{12d}$$

Substituting (8) and (9) into (12) results in four equations for the four unknowns A^+, A^-, B^+, and B^-. Solving for these quantities and substituting them into (8) yield a solution* of the form

$$\frac{E_1(x)}{V_1} = \tfrac{1}{2}R^+(x) + \tfrac{1}{2}R^-(x) \tag{13a}$$

$$\frac{E_2(x)}{V_1} = \tfrac{1}{2}R^+(x) - \tfrac{1}{2}R^-(x). \tag{13b}$$

The near-end crosstalk is given by $E_2(o)/E_1(o)$ while† the far-end crosstalk (equal level) is given by $E_2(l)/E_1(l)$.

* In principle, (13) could be derived from eqs. (1.25) and (1.30) of Ref. 3. However, applying the boundary conditions (12) to these equations leads to sufficient algebraic complication that it is easier to derive (13) directly.

† The conventional definition of near-end crosstalk is $E_2(o)/V_1$ which is equivalent to the above definition (except for a factor of 2) under the conditions of loose coupling and matched terminations.

Obtaining expressions for $R^{+}(x)$ and $R^{-}(x)$ involves very extensive algebra for arbitrary impedances and in general does not lend insight into the coupling process. For applications where switching circuits are involved, several special cases of interest partially simplify the algebra in obtaining expressions for the near-end and far-end crosstalk. Some of these cases are:

(*i*) $Z_1 = Z_3$ and $Z_2 = Z_4$,[7] (possible application to analog switching systems).
(*ii*) $Z_1 = Z_3 = Z$, $Z_2 = Z_4 = \infty$ (possible application to switching systems using "totem pole" logic).
(*iii*) $Z_1 = 0$, $Z_2 = Z_3 = Z_4 = \infty$ (possible application to switching systems using simple transistor logic).

These cases all involve somewhat bulky expressions, but they can be obtained with perseverance.

The case that will be studied in detail in the following section is $Z_1 = Z_2 = Z_3 = Z_4$. This case is of special interest for three reasons:

(*i*) The coupling capacitance and inductance can be related easily to physically measurable quantities.
(*ii*) The conditions under which the "classical" crosstalk formulas apply can be studied.
(*iii*) This case is of interest for many applications involving analog circuits.

III. RELATIONSHIP TO PHYSICAL QUANTITIES

The behavior of the coupling process is most easily illustrated by modifying the excitation assumed in (12). Instead of only exciting circuit 1, an excitation will also be applied to circuit 2 as shown in Fig. 3. The set of equations, (12), is modified by letting $Z_1 = Z_2 = Z_3 = Z_4$ and replacing (12b) by

$$\rho V_1 = Z_1 I_2(o) + E_2(o) \tag{14}$$

where ρ is a complex scalar. Obviously, the case $\rho = 0$ corresponds to the situation in Fig. 2 with equal terminating impedances. With this substitution, (13) becomes

$$\frac{E_1(x)}{V_1} = \frac{1+\rho}{2} R_o^{+}(x) + \frac{1-\rho}{2} R_o^{-}(x) \tag{15a}$$

$$\frac{E_2(x)}{V_1} = \frac{1+\rho}{2} R_o^{+}(x) - \frac{1-\rho}{2} R_o^{-}(x) \tag{15b}$$

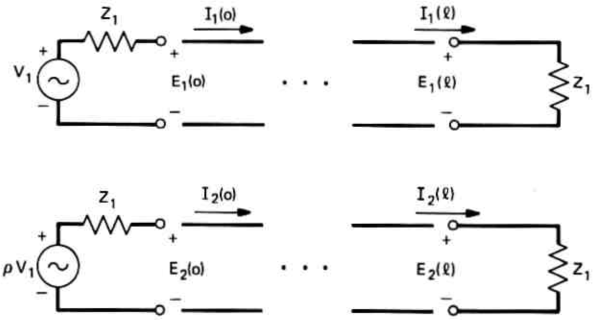


Fig. 3—An alternate means of exciting the coupled transmission lines.

where

$$R_o^+(x) = \frac{P_{00}e^{\gamma^+(l-x)} - P_{10}e^{-\gamma^+(l-x)}}{P_{00}^2 e^{\gamma^+ l} - P_{10}^2 e^{-\gamma^+ l}} \tag{16a}$$

$$R_o^-(x) = \frac{P_{01}e^{\gamma^-(l-x)} - P_{11}e^{-\gamma^-(l-x)}}{P_{01}^2 e^{\gamma^- l} - P_{11}^2 e^{-\gamma^- l}} \tag{16b}$$

and

$$P_{00} = 1 + \frac{Z_1}{Z^+} \tag{17a}$$

$$P_{10} = 1 - \frac{Z_1}{Z^+} \tag{17b}$$

$$P_{01} = 1 + \frac{Z_1}{Z^-} \tag{17c}$$

$$P_{11} = 1 - \frac{Z_1}{Z^-}. \tag{17d}$$

It is now apparent that any excitation of the two coupled transmission lines depicted in Fig. 3 will result in a response which will be a linear combination of the two functions $R_o^+(x)$ and $R_o^-(x)$. Therefore, these functions will be referred to as modes. They will now be examined in somewhat greater detail.

If $\rho = 1$, then (15) reduces to

$$\frac{E_1(x)}{V_1} = \frac{E_2(x)}{V_1} = R_o^+(x). \tag{18}$$

Therefore, if the two lines are energized with equal and in-phase sinusoids, the resulting voltage distributions are given by $R_o^+(x)$. Note that $R_o^+(x)$ contains terms with a + superscript and does not contain any terms with a − superscript. This, in turn, signifies that the propagation constant and the characteristic impedance associated with $R_o^+(x)$ are given by (6) and (10), respectively. This result will now be interpreted in terms of the distributed capacitance and inductance associated with the two transmission lines.

Figure 4 shows a cross section of the two coupled transmission lines, assuming symmetric excitation ($\rho = 1$). According to (18), the voltages on the two lines are equal at every point x; this fact is indicated on Fig. 4. The capacitance per unit length of each conductor to the ground plane is denoted by C_g, while the coupling between conductors is denoted by C_{12}.

Since there is no potential difference across C_{12}, the signals propagating along the two transmission lines are not affected by it. Therefore, each signal propagates along its respective transmission line as if the two lines were uncoupled and with distributed capacitance:

$$C + C_c = C_g. \tag{19}$$

The distributed inductance can be expressed in terms of C_g using the relationship:

$$(C + C_c)(L + L_c) = \mu\epsilon, \tag{20}$$

(See, for example, Chapter I, Sec. 4, eq. (31) of Ref 8.) This formula is applicable to the case where the frequency of excitation is sufficiently high that the magnetic field penetrating the metal conductors contributes a negligible amount to the coupling inductance. Thus

$$L + L_c = \frac{\mu\epsilon}{C_g}. \tag{21}$$

Turning now to the case $\rho = -1$, eq. (15) yields

$$\frac{E_1(x)}{V_1} = -\frac{E_2(x)}{V_1} = R_o^-(x). \tag{22}$$

The propagation constant and characteristic impedance associated with this mode are expressed by (7) and (11), respectively. As with the previous mode, this mode behaves as if the two lines were uncoupled but with primary constants R, G, $C - C_c$, and $L - L_c$. To see how these are related to the physical capacitance, it is convenient to depict the voltages and capacitances as shown in Fig. 5.

As indicated by (22), the voltages on each transmission line are equal but opposite in sign. A vertical line between the two conductors

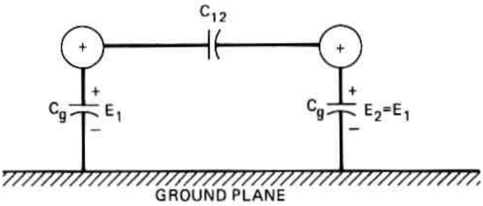


Fig. 4—Cross section of two coupled transmission lines—symmetric excitation.

must therefore constitute a surface of ground potential. Thus, the total capacitance to ground influencing a signal propagating along either line is given by

$$C - C_c = C_g + 2C_{12}. \tag{23}$$

As in the previous case, one may take $(C - C_c)(L - L_c) = \mu\epsilon$ if the frequency is sufficiently high. Thus

$$L - L_c = \frac{\mu\epsilon}{C_g + 2C_{12}}. \tag{24}$$

Combining (19) and (23) to solve for C and C_c yields

$$C = C_g + C_{12} \tag{25}$$

and

$$C_c = -C_{12}, \tag{26}$$

while combining (21) and (24) to solve for L and L_c results in

$$L = \mu\epsilon \frac{C_g + C_{12}}{C_g(C_g + 2C_{12})} \tag{27}$$

and

$$L_c = \mu\epsilon \frac{C_{12}}{C_g(C_g + 2C_{12})}. \tag{28}$$

One final observation is that, in the higher frequency bands of in-

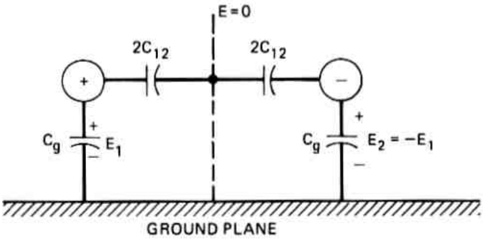


Fig. 5—Cross section of two coupled transmission lines—asymmetrical case.

terest, the following approximations can be made with little error:

$$\gamma^+ \cong \frac{R}{2}\frac{1}{Z^+} + j\omega\sqrt{\mu\epsilon} \tag{29a}$$

$$\gamma^- \cong \frac{R}{2}\frac{1}{Z^-} + j\omega\sqrt{\mu\epsilon} \tag{29b}$$

$$Z^+ \cong \left(\frac{L + L_c}{C + C_c}\right)^{\frac{1}{2}} \tag{29c}$$

$$Z^- \cong \left(\frac{L - L_c}{C - C_c}\right)^{\frac{1}{2}}. \tag{29d}$$

(See Chapter II, Sec. 13, eqs. (18), (8), and (6) of Ref. 8.)

Substituting eqs. (25) to (28) into the above equations yields

$$\gamma^+ \cong \frac{1}{2}\frac{RC_g}{\sqrt{\mu\epsilon}} + j\omega\sqrt{\mu\epsilon} \tag{30a}$$

$$\gamma^- \cong \frac{1}{2}\frac{R(C_g + 2C_{12})}{\sqrt{\mu\epsilon}} + j\omega\sqrt{\mu\epsilon} \tag{30b}$$

$$Z^+ \cong \frac{\sqrt{\mu\epsilon}}{C_g} \tag{30c}$$

$$Z^- \cong \frac{\sqrt{\mu\epsilon}}{C_g + 2C_{12}}. \tag{30d}$$

Thus, the $R_o^-(x)$ mode has a higher loss and smaller characteristic impedance than the $R_o^+(x)$ mode.

It is now possible to outline a measurement procedure that will yield all quantities required to evaluate (13). Since the effective dielectric constant surrounding most physical transmission lines is determined by the detailed geometry of the insulation and shields surrounding each conductor, the quantity $\sqrt{\mu\epsilon}$ will be assumed unknown for the following procedure, even though μ and ϵ may be known for each constituent material in the transmission line.

Step 1. Measure C_g and C_{12}.

Step 2. Terminate the coupled pairs in four equal impedances Z_1 and energize the two lines from the same voltage generator. The generator frequency should be in the range for which the approximations leading to (29) and (30) are valid. In other words, terminate and excite the lines as shown in Fig. 3 with $\rho = 1$.

Step 3. Adjust all four impedances, Z_1, until $R_o^+(l)$ is a maximum. It is easy to show that in this case $Z_1 = Z^+$. Using (30c) and the value of C_g from Step 1 gives $\sqrt{\mu\epsilon}$.

Step 4. With $Z_1 = Z^+$, measure $R_o^+(l)$ which equals $(\frac{1}{2}) \exp(-\gamma^+ l)$. With γ^+ given by (30a), R can be solved for directly, given the length of the coupled lines, l.

Step 5. The remaining quantities, γ^- and Z^-, can now be evaluated using (30b) and (30d). Note that there is an additional check on the value of $\sqrt{\mu\epsilon}$ through the imaginary part of γ^+.

IV. COMPARISON OF RESULTS WITH CLASSICAL CROSSTALK FORMULAS

We now use the results of the previous section to analyze the "classical" crosstalk formulas as derived by Shelkunoff and Odarenko.[2] Their analysis assumed uniform weak coupling between two parallel transmission lines terminated in their characteristic impedances. Assuming the two transmission lines had identical primary and secondary constants, they derived the following formulas for near-end crosstalk (NEXT) and far-end crosstalk (FEXT):

$$N(\omega) = \frac{Z_{12}}{4Z_0\gamma_0(\omega)}\left(1 - e^{-2\gamma_0(\omega)l}\right), \tag{31}$$

$$F(\omega) = \frac{Z_{12}'}{2Z_0}\, l, \tag{32}$$

where Z_{12}, Z_{12}' are the mutual impedances between the two lines for NEXT and FEXT, respectively, Z_0 and γ_0 the secondary quantities of an isolated line [i.e., (10) and (6) with $L_c = C_c = 0$], and l the length of the lines. The above formulas were derived neglecting the effects of the disturbed line on the disturbing line. In the discussion that follows, we shall examine the validity of this assumption.

Referring to (6), (7), (10), and (11), and assuming $L_c \ll L$, $C_c \ll C$, and $Z_1 = Z_0$, it can be shown that

$$Z^{\pm} \cong Z_0 \pm \delta \tag{33a}$$

$$\delta = \frac{j\omega}{2\gamma_0}\left(L_c - C_c Z_0^2\right) \tag{33b}$$

$$\left|\frac{\delta}{Z_0}\right| \ll 1 \tag{33c}$$

and

$$\gamma^{\pm} = \gamma_0 \pm \epsilon \tag{34a}$$

$$\epsilon = j\omega \frac{(L_c + C_c Z_0^2)}{2Z_0} \tag{34b}$$

$$\left|\frac{\epsilon}{\gamma_o}\right| \ll 1. \tag{34c}$$

Now using (33), (34), (13), (16), and (17) one can show that

$$N(\omega) = \frac{E_2(o)}{E_1(o)} \cong \frac{8\delta}{Z_0} \frac{\left(1 - e^{-2\gamma_0 l}\left[\cosh\,(2\epsilon l) - j\frac{\delta}{Z_0}\sinh\,(2\epsilon l)\right]\right)}{16} \tag{35}$$

where the approximation is obtained by only assuming weak coupling. Now for sufficiently small $|\epsilon l|$, the term in brackets is approximately unity so that (35) becomes

$$\begin{aligned} N(\omega) &\cong \frac{\delta}{Z_0}(1 - e^{-2\gamma_0 l}) \\ &= \frac{j\omega}{4\gamma_0 Z_0}(L_c - C_c Z_0^2)(1 - e^{-2\gamma_0 l}). \end{aligned} \tag{36}$$

This agrees with the Shelkunoff and Odarenko result, (31), with $Z_{12} = j\omega(L_c - C_c Z_0^2)$.

For larger values of l, the exponential term in (35) can usually be neglected for lossy lines. In the lossless case, the term in brackets will cause a departure from the classical formula for sufficiently large $|\epsilon l|$; however, in weak coupling situations, the length and/or frequency required to invalidate the approximation $\cosh\,(2\epsilon l) \cong 1$ are usually large enough to invalidate the lossless assumption. Thus, for most practical situations involving weak coupling, eq. (31) is adequate.

We now consider far-end crosstalk. Again referring to (13), (16), and (17), letting $x = l$, and assuming the conditions for weak coupling exist, it can be shown that

$$F(\omega) \cong \frac{e^{\gamma^- l} - e^{\gamma^+ l}}{e^{\gamma^- l} + e^{\gamma^+ l}}. \tag{37}$$

Substituting (34a) into (37) results in

$$\begin{aligned} F(\omega) &\cong \frac{e^{-\epsilon l} - e^{\epsilon l}}{e^{-\epsilon l} + e^{\epsilon l}} \\ &= -\tanh\,(\epsilon l). \end{aligned} \tag{38}$$

Referring to (34b), for sufficiently large ω, Z_0 is a real constant and ϵ is an imaginary number; thus, (38) can be written as

$$F(\omega) \cong -j \tan(-j\epsilon l) \tag{39a}$$

$$\cong -\epsilon l \qquad \text{for} \qquad |\epsilon l| \leqq \frac{\pi}{6}$$

$$= -j\omega \frac{(L_c + Z_0^2 C_c) l}{2Z_0}. \tag{39b}$$

Therefore, for $|\epsilon l| \leqq (\pi/6)$, eq. (39b) agrees with (32) with $Z_{12}' = -j\omega(L_c + Z_0^2 C_c)$. Shelkunoff and Odarenko[2] point out that (32) must not be carried to an absurd conclusion: namely, that most of the far-end power will reside in the disturbed circuit for sufficiently long transmission lines. They conjecture that, in the limiting case, the far-end power will divide equally between the two lines. Equation (39a) indicates that the far-end power oscillates back and forth between the two lines as a function of l (or frequency, since ϵ is a function of ω). Equation (39a) is valid over a larger range of l than (32) (or 39b), although it is not valid for all l, since it is based on an approximation, (34), which is multiplied by l. To be more specific, (6) and (7) can be written as

$$\gamma^{\pm} = [(\gamma_0^2 - \omega^2 L_c C_c) \pm j\omega\gamma_0(Z_0 C + L_c/Z_0))]^{\frac{1}{2}}$$

$$= \gamma_0 \left[1 \pm \frac{j\omega}{\gamma_0 Z_0} (L_c + C_c Z_0^2) - \frac{\omega^2}{\gamma_0} L_c C_c \right]^{\frac{1}{2}}. \tag{40}$$

Assuming the conditions for weak coupling ($L_c \ll L$, $C_c \ll C$), the third term in the brackets is much smaller than the second term, and the second term in the brackets has a magnitude much less than unity, so (34) is a good first-order approximation to (40). Now for Z_0 a real constant, ϵ is an imaginary quantity. However, for any given frequency the higher order terms from (40) contain real parts which will dominate the behavior of the exponential terms in (37) for sufficiently large l. Thus, referring to (37), in the limit as $l \to \infty$,

$$\lim_{l \to \infty} |F(\omega)| = \lim_{l \to \infty} \left| \frac{1 - e^{(\gamma^+ - \gamma^-)l}}{1 + e^{(\gamma^+ - \gamma^-)l}} \right| = 1. \tag{41}$$

The same result is reached by fixing l and letting $\omega \to \infty$.

In summary, when weak coupling conditions exist, the effect of the disturbed line on the disturbing line can be neglected for most NEXT calculations; however, for a sufficiently large l and/or ω, the effect of the disturbed line on the disturbing line cannot be neglected for FEXT calculations. This is because, for certain values of l and/or ω, the far-end power in the disturbed line will be comparable to the far-end power in the disturbing line.

V. CONCLUSION

An analytical model for analyzing crosstalk between two identical, parallel, uniformly coupled transmission lines with ground return has been presented. Using this model, formulas were developed for the two sets of modal voltages and currents on the transmission lines. It was found that each mode has associated with it a propagation factor and characteristic impedance which, in general, are different for each mode.

By applying different sets of excitation voltages to the two lines (changing boundary conditions), the effect of each line on the other can be analyzed in terms of the modal quantities. Using this technique, formulas were derived for the coupling capacitance and inductance in terms of the distributed capacitance and distributed inductance of an isolated line, the distributed capacitance to ground for the nonisolated lines, and the permeability and permittivity of the medium surrounding the transmission lines.

The mutual influence of the two lines was also studied by assuming weak coupling between them and then deriving NEXT and FEXT formulas using this model. These formulas were compared with the classical formulas which do not take into account the influence of the disturbed line on the disturbing line. In the case of NEXT, the effect of the disturbed line on the disturbing line was found to be negligible for most practical cases. In the case of equal level FEXT, however, the effect of the disturbed line on the disturbing line can be quite significant for sufficiently large line length and/or frequency.

REFERENCES

1. Campbell, G. A., "Dr. Campbell's Memoranda of 1907 and 1912," B.S.T.J., *14*, No. 4 (October 1935), pp. 558–572.
2. Shelkunoff, S. A., and Odarenko, T. M., "Crosstalk Between Coaxial Transmission Lines," B.S.T.J., *16*, No. 2 (April 1937), pp. 144–164.
3. Rice, S. O., "Steady State Solutions of Transmission Line Equations," B.S.T.J., *20*, No. 2 (April 1941), pp. 131–178.

4. Kuznetsov, P. I., and Stratonovich, R. L., *The Propagation of Electromagnetic Waves in Multiconductor Transmission Lines*, 1964, London: Pergamon Press, Ltd., distributed by the MacMillian Company, New York. See Chapter 6 which is entitled: "The Propagation of Electromagnetic Waves Along Two Parallel Single-Wire Lines."
5. Amemiya, H., "Time-Domain Analysis of Multiple Parallel Transmission Lines," RCA Review, *28*, 1967, pp. 241–276.
6. Dvorak, V., "Numerical Solution of the Transient Response of a Distributed Parameter Transformer," IEEE Trans. on Circuit Theory (May 1970), pp. 270–273.
7. Remec, M. J., private communication.
8. King, R. W. P., *Transmission Line Theory*, New York: Dover Publications Inc., 1965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

Applications for Quantum Amplifiers in Simple Digital Optical Communication Systems

By S. D. PERSONICK

(Manuscript received August 14, 1972)

Previously published results on the performance of optical direct detection digital receivers using avalanche detectors are extended to the case where incoherent noise due to quantum amplifiers in the transmission medium is present at the detector. These calculations are applied to determine the usefulness of quantum amplifiers in simple digital transmission systems where the optical source instability results in a required amplifier bandwidth which may be orders of magnitude greater than the modulation bandwidth. It is concluded that practical applications exist where quantum amplifiers can be used in analog repeaters between regenerating repeaters in a hybrid digital system; and also as front ends of regenerating repeaters to increase their sensitivities.

I. INTRODUCTION

Quantum amplifiers can be used in optical communication systems even if the optical sources are only partially coherent. They can serve as optical analog repeaters between regenerating repeaters in a hybrid digital system to compensate for transmission loss (see Fig. 1), and also as the front ends of regenerating repeaters which demodulate back to baseband.

This paper investigates the applications for quantum amplifiers in simple digital communication systems employing "on-off" intensity modulation. It will be assumed that due to source instability, the optical system bandwidth may be orders of magnitude greater than the bandwidth of the modulation, and that the quantum amplifiers have limited gain.

We shall calculate Chernov bounds on the required signal energy per pulse at the detector of a digital repeater (to be described in detail below) to achieve a 10^{-9} error rate as a function of the received sponta-

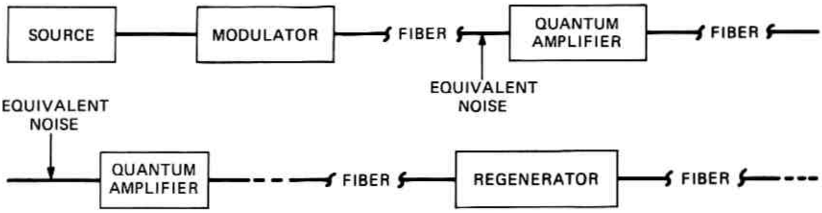


Fig. 1—Fiber communication system.

neous emission noise level from quantum amplifiers, or of the incoherent background noise level for some typical values of dark current of the detector, mean detector gain, and thermal noise introduced by circuitry following the detector. For these Chernov bound calculations, we shall assume a unilateral gain detector. Numerical results for other detectors and parameter values can be obtained by using the moment generating functions to be derived below with the results of two previous papers[1,2] concerning Chernov bounds for direct detection intensity modulation systems using avalanche gain.

We shall also derive some signal-to-noise ratio results which can be used to approximate the energy required per pulse to achieve a desired error rate. These signal-to-noise ratio results are consistent with previous published work of other authors.[3-5]

Throughout this paper it will be assumed that the modulation consists of varying the intensity of the transmitted signal in each baud interval to produce pulses at the regenerating repeater which are of one of two amplitudes, and such that the pulses are approximately constant in a baud interval of length T seconds. Generalization to other pulse shapes should be straightforward using the results below.

II. A MODEL FOR THE QUANTUM AMPLIFIER NOISE

Throughout this paper we shall model the source as follows. Its voltage in a single spatial mode will be given by

$$E_s(t) = \sqrt{2}re\{Ae^{i(\Omega+\omega)t}\} \tag{1}$$

where $|\omega| < 2\pi B/2$.

That is, the source will be nominally at optical frequency $\Omega/2\pi$ but due to source instability there will be an uncertainty of bandwidth B. (The conclusions and numerical results that follow also hold if the source is a randomly phase-modulated sinusoid having a bandwidth B of the form $E_s(t) = \sqrt{2}re\{Ae^{i(\theta(t)+\Omega t)}\}$.)

If the modulated signal is to be transmitted over a channel with quantum amplifiers and possibly with optical filters as well, then these devices must have a bandwidth of at least $B + 1/T$ to accommodate the modulated signal for all possible values of ω. (The term $1/T$ is due to the increase in bandwidth of the source due to the modulation.)

At the regenerating repeater input, the classical field will be modeled as follows (assuming only a single spatial mode)

$$E_r(t) = \sqrt{2} re\{m(t)e^{i(\Omega+\omega)t} + n(t)\} \tag{2}$$

where $m(t)$, the modulation, assumes one of the two possible pulse amplitudes (a pulse which is approximately constant in each baud interval T) and $n(t)$ is a complex Gaussian random process which represents the incoherent spontaneous emission noise introduced in the quantum amplifiers or represents incoherent background noise.[6] In each baud interval T, we can expand the field complex envelope in a Fourier series;[†] taking only enough terms to include the "system" bandwidth B'.[‡]

$$\epsilon_r(t) = m(t)e^{i\omega t} + n(t) = \sum_{-(L-1)/2}^{(L-1)/2} a_k \left[\frac{e^{i(2\pi kt/T)}}{\sqrt{T}}\right] \tag{3}$$

where L (the number of temporal modes) is given by $L = B'T \geqq 1 + BT$. Defining

$$m_k = \frac{1}{\sqrt{T}} \int_{\text{baud interval}} m(t)e^{i\omega t}e^{-i(2\pi kt/T)}dt\,,$$

$$n_k = \frac{1}{\sqrt{T}} \int_{\text{baud interval}} n(t)e^{-i(2\pi kt/T)}dt\,, \tag{4a}$$

we have for each value of k, $a_k = m_k + n_k$. Because we have a digital system, the signal components, m_k, take on one of two values for each k. The noise components n_k are complex Gaussian random variables.

$$\langle n_k n_j^* \rangle = N_o \delta_{kj}, \quad \langle n_k n_j \rangle = 0 \tag{4b}$$

where N_o is the classical incoherent noise spectral height[§]: $\langle n(t)n^*(\tau)\rangle = N_o\delta(t - \tau)$, and $\langle x \rangle$ stands for the expected value of x.

[†] A more rigorous and general approach taken in the Appendix is to expand the received field in a Karhunen-Loéve expansion[5] using the autocorrelation function of the noise $n(t)$ at the detector input as the kernel. The approach taken here is justified on grounds of simplicity and intuitiveness.

[‡] The system bandwidth B' is the minimum of the quantum amplifier bandwidth, the detector optical bandwidth, and the bandwidths of any filters in the optical path preceding the detector. Of course $B' > B + (1/T)$, if we are to accommodate all possible signals with the unstable source described above.

[§] That is, the number of watts of incoherent power falling on the detector in the bandwidth B' is N_oB'.

At the regenerating repeater it will be assumed that the field falls upon a detector with internal gain (e.g., an avalanche detector) and causes the detector to emit "primary" hole-electron pairs at rate (average pairs/second)

$$\lambda(t) = \frac{\eta}{\hbar\Omega}|\epsilon_r(t)|^2 \tag{5}$$

where $\hbar$ = Planck's constant/2π, Ω = optical frequency in radians/s, η = detector quantum efficiency, and $\epsilon_r(t)$ was defined in eq. (3) above.

Due to internal gain, each primary "count" (hole-electron pair) produces a random number of additional secondary counts. Because the modulation pulse is approximately constant throughout a baud interval, we will be interested in the total number of counts produced by the detector due to signal and incoherent noise in each baud interval. The moment generating function[7] of the random total number of counts, N, produced in each baud interval, T, is defined as

$$M_N(s) = \sum_{n=o}^{\infty} e^{sn}p(n) \tag{6}$$

where

$$p(n) = \text{probability that } N = n.$$

From previous work[1] we have

$$M_N(s) = M_C(\psi_G(s)) \tag{7}$$

where

$$\psi_G(s) \quad \text{is} \quad \ln[M_G(s)].$$

$M_G(s)$ is the moment generating function of the random internal gain G and $M_C(s)$ is the moment generating function of the total number of primary counts, C.

We can evaluate $M_C(s)$ as follows. Define the quantity Λ as

$$\Lambda = \frac{\eta}{\hbar\Omega}\int_{\text{baud interval}} |\epsilon_r(t)|^2 dt = \frac{\eta}{\hbar\Omega}\sum_{-(L-1)/2}^{(L-1)/2} |a_k|^2 \tag{8}$$

where Λ is the average number of received primary counts in a baud interval given $\epsilon_r(t)$. Λ is a random variable, since the $\{a_k\}$ are random variables having the following joint complex Gaussian probability density

$$p\{a_1, a_2 \cdots a_L\} = \prod_{k=1}^{L} \frac{1}{\pi N_o} e^{-|a_k - m_k|^2/N_o}.$$

The probability distribution of the total number of primary counts

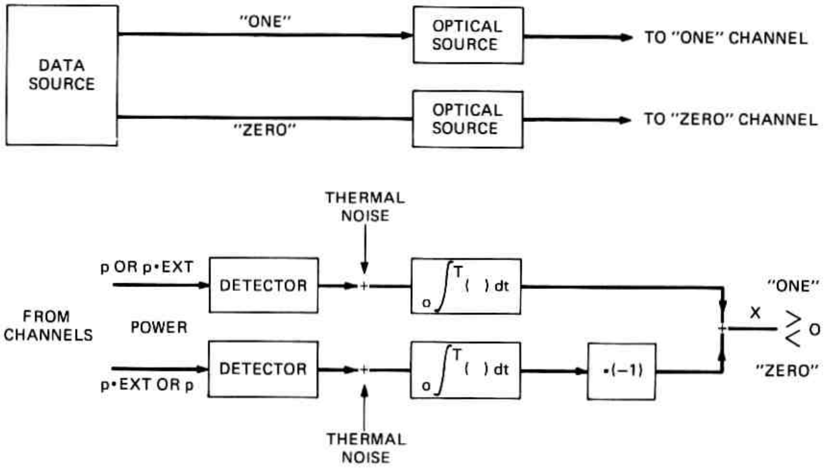


Fig. 2—Twin-channel system.

C in a baud interval *given* Λ is Poisson, i.e.,

$$p(c\,|\,\Lambda) = \frac{\Lambda^c e^{-\Lambda}}{c!} = \text{probability that } C = c \text{ given } \Lambda .$$

It follows that $M_C(s)$ is given by

$$\begin{aligned} M_C(s) &= \int_o^\infty \left[\sum_o^\infty p(c\,|\,\Lambda) e^{sc} \right] p(\Lambda) d\Lambda \\ &= \int_{-\infty}^{\infty} e^{\Lambda(e^s-1)} p(\Lambda) d\Lambda \\ &= \left[1 - \frac{\eta N_o}{\hbar\Omega} (e^s - 1) \right]^{-L} \\ &\qquad \times e^{\{(\eta/\hbar\Omega)\,\Sigma\,|m_k|^2(e^s-1)/[1-(\eta/\hbar\Omega)N_o(e^s-1)]\}}. \end{aligned} \tag{9}$$

III. SIGNAL-TO-NOISE RATIO RESULTS

From (7) and (9) we obtain the mean number of counts, $\langle N \rangle$, emitted by the avalanche detector in a baud interval as follows

$$\begin{aligned} \langle N \rangle &= \frac{\partial}{\partial s} M_N(s) = \frac{\partial}{\partial[\psi_G(s)]} M_C(\psi_G(s)) \frac{\partial}{\partial s} \psi_G(s) \bigg|_{s=0} \\ &= \bar{G}[m^2 + LN_o]\eta/\hbar\Omega \end{aligned} \tag{10}$$

where

$$m^2 = \int_{\text{baud interval}} |m(t)|^2 dt = \sum_{-(L-1)/2}^{(L-1)/2} |m_k|^2,$$

$N_o = \langle |n_k|^2 \rangle$ = classical spectral height of the incoherent noise at the detector input,

$L \geqq [B + 1/T]T = BT + 1,$

and $\bar{G}$ is the mean avalanche gain.

The variance of the total number of counts is

$$\langle N^2 \rangle - \langle N \rangle^2 = \frac{\partial^2}{\partial s^2} M_N(s)\bigg|_{s=0} - \left(\frac{\partial}{\partial s} M_N(s)\bigg|_{s=0}\right)^2$$

$$= \underbrace{(m^2 + LN_o)\frac{\eta}{\hbar\Omega}\overline{G^2}}_{\text{shot noises}} + \underbrace{(LN_o^2 + 2N_o m^2)(\bar{G})^2\left(\frac{\eta}{\hbar\Omega}\right)^2}_{\text{beat noises}\dagger} \qquad (11)$$

where $\overline{G^2}$ is the mean square avalanche gain.

Consider a typical twin-channel digital system, shown in Fig. 2. There is light incident on each detector containing "on-off" modulated signal pulses of duration T and incoherent noise. A channel is in the "on" state when its signal pulse has optical power p. In the "off" state the signal pulse power is $p \cdot EXT$, where EXT is small compared to unity. During each baud interval, one or the other channel is "on." The detectors are assumed to have internal random gain (e.g., avalanche gain or photomultiplier gain) and there are assumed to be thermal noises added to the detector outputs due to the amplifiers following the detectors. It is assumed that the signaling rate is slow enough so that each signal pulse of light of duration T produces an output current from its detector of duration T that does not overlap with the currents from other pulses. The detector output current pulses plus the corresponding noises are integrated in each period T (or equivalently filtered). The output variable x is compared to the threshold after each integration to decide which channel is "on." An error is made if $x > 0$ when the "zero" channel is on, or vice-versa.

The baseband noise-to-signal ratio is defined as the variance of the output voltage x divided by the square of the mean of the output voltage x.

† The term "beat noise" has been used in literature[8] to describe those noise terms at the output of a square law detector which are due to fluctuations in the instantaneous power of a carrier which has a fluctuating amplitude.

$$\frac{\langle x^2\rangle - (\langle x\rangle)^2}{\langle x\rangle^2} = \underbrace{\frac{4\mathrm{k}\theta T}{Re^2\lambda_s^2(1 - EXT)^2\bar{G}^2}}_{\text{thermal noise}} + \underbrace{\frac{2\lambda_d + 2L\overset{\swarrow}{\lambda_n} + \lambda_s(1 + EXT)}{\lambda_s^2(1 - EXT)^2}}_{\text{shot noises}} \frac{\overline{G^2}}{(\bar{G})^2} + \underbrace{\frac{2L\overset{\swarrow}{\lambda_n^2} + 2\lambda_s\overset{\swarrow}{\lambda_n}(1 + EXT)}{\lambda_s^2(1 - EXT)^2}}_{\text{beat noises}} \tag{12}$$

where

$\mathrm{k}\theta$ = Boltzman's constant·absolute noise temperature referred to the integrator input.

R = integrator equivalent thermal noise input resistance.

λ_d = mean dark current counts per detector per interval T before avalanche gain.

$\lambda_s = m^2\eta/\hbar\Omega$ = mean signal counts per interval T in "on" channel before avalanche gain.

$L\lambda_n$ = mean incoherent noise counts in either channel per baud interval T before avalanche gain.

$L \geqq BT + 1$, and equals the number of temporal modes detected.

EXT = Signal power in "off" channel/signal power in "on" channel.

In eq. (12), terms which are due to the incoherent spontaneous emission noises of the quantum amplifiers (or background noise) are marked with arrows.

We see that the optical incoherent noise, when detected to baseband, causes additional shot noise and also contributes two beat noise terms. One of these is proportional to the signal λ_s and one is proportional to L. One can use these signal-to-noise ratio results to approximate the error rate by assuming that the output variable x is roughly Gaussian in distribution.

In the next section we shall generate some curves that may give a clearer picture of the effects of L, λ_n, λ_s, etc., on performance.

IV. CHERNOV BOUNDS

The moment generating function defined in (9) was used with previously published results[1,2] on avalanche photo-diode gain statistics to obtain Chernov upper bounds on the energy per pulse required at the input of a digital twin-channel regenerating repeater of Fig. 2 to achieve a desired error rate as a function of the other parameters.

The general Chernov bound is given as follows.[7] Let χ be a random variable with moment generating function $M_\chi(s)$. Let $\Pr_\chi(x > \gamma)$ be the probability that an outcome x of χ exceeds γ. Then it follows that

$$\Pr_\chi(x > \gamma) \leqq e^{[\psi_\chi(s) - s\gamma]} \quad \text{for} \quad s > 0 \tag{13}$$

where

$$\psi_\chi(s) = \ln[M_\chi(s)].$$

The bound is optimized for s such that $(\partial\psi(s)/\partial s) = \gamma$ provided that value of s is greater than zero.

Similarly,

$$\Pr_\chi(x < \gamma) \leqq e^{[\psi_\chi(s) - s\gamma]} \quad \text{for} \quad s < 0 \tag{14}$$

where the optimal value of s is given by $(\partial\psi(s)/\partial s) = \gamma$ provided that value of s is less than zero.

To obtain Chernov bounds upon the probability of error for the twin-channel system of Fig. 2, one needs the moment generating

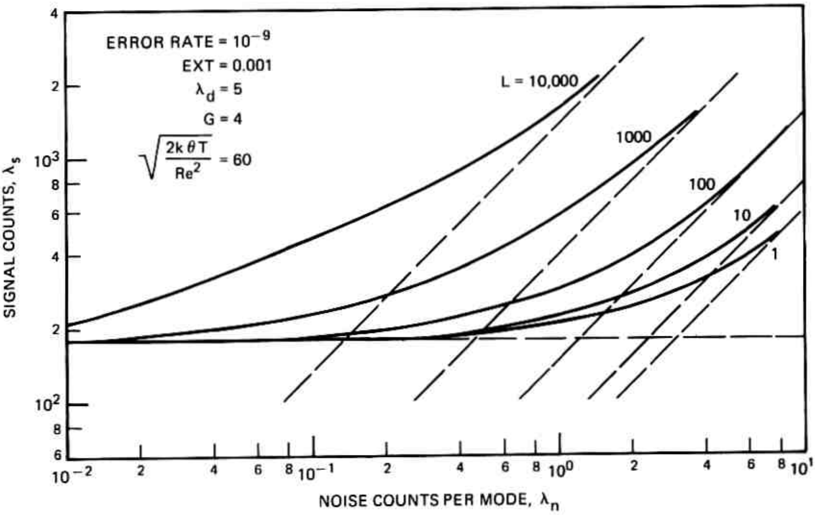


Fig. 3—Required energy per pulse normalized by $\eta/\hbar\Omega$ vs the incoherent noise level N_o at the detector, also normalized by $\eta/\hbar\Omega$.

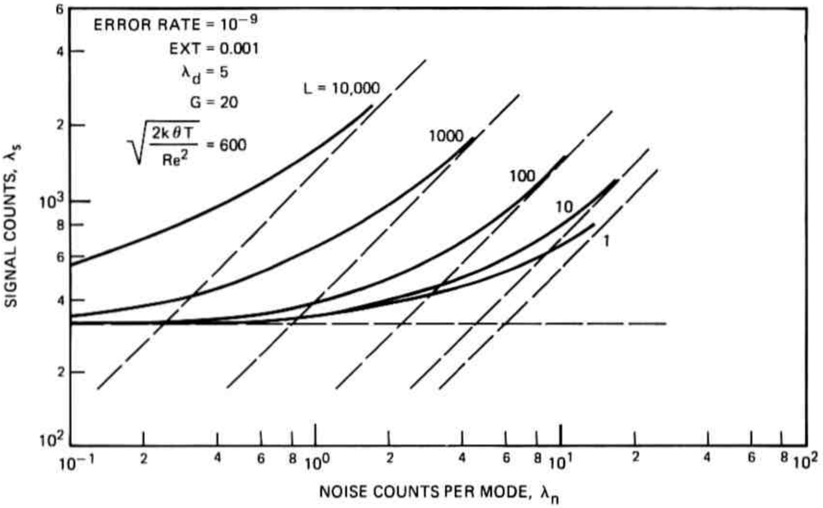


Fig. 4—Same as Fig. 3.

function of the output variable x. This can be obtained using (9) and the results of the Refs. 1 and 2, which are too detailed to duplicate here.

From simple cases where error rates can be calculated exactly, the differences in required power between those results and the bounds are typically a few dB or less. Experimental results also confirm the tightness of the bounds. Therefore, in this paper we shall take the liberty of comparing the effects of various parameters upon the required energy per pulse to achieve a desired error rate by comparing the bounds.

It was decided that the calculations should be presented graphically in two ways.

First, in Figs. 3 to 5, the required energy per pulse normalized by $\eta/\hbar\Omega$ (i.e., the mean number of detected signal photons per pulse) is plotted vs the incoherent noise level N_o at the detector also normalized by $\eta/\hbar\Omega$. This is done for various values shown of L, mean avalanche gain, thermal noise, dark current, error rate, and extinction ratio, for a low-noise unilateral gain avalanche detector (i.e., a detector in which only one type of carrier causes ionizing collisions, and where carrier injection is from one end of the high field region). The avalanche gains used in these calculations do not minimize the required energy per pulse for the given values of the other parameters, but were used for illustration.

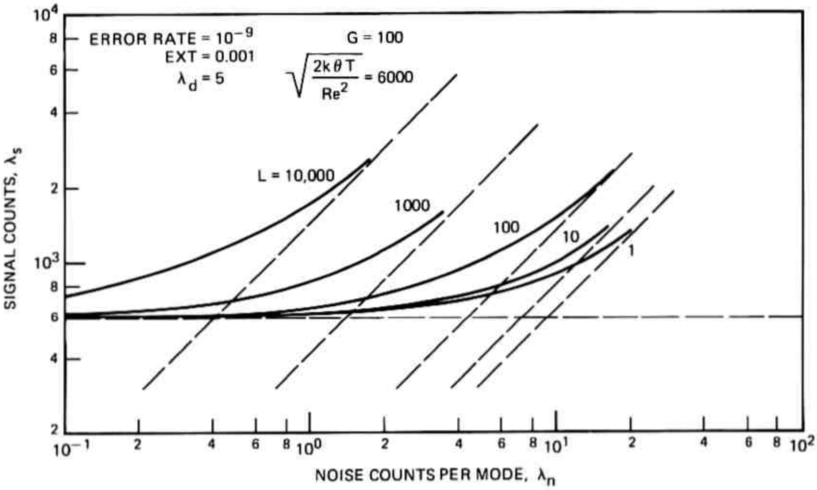


Fig. 5—Same as Fig. 3.

It was recognized that in a hybrid system, if the loss between the regenerating repeater and the analog repeater closest to it is increased, then the signal energy per pulse at the regenerating repeater

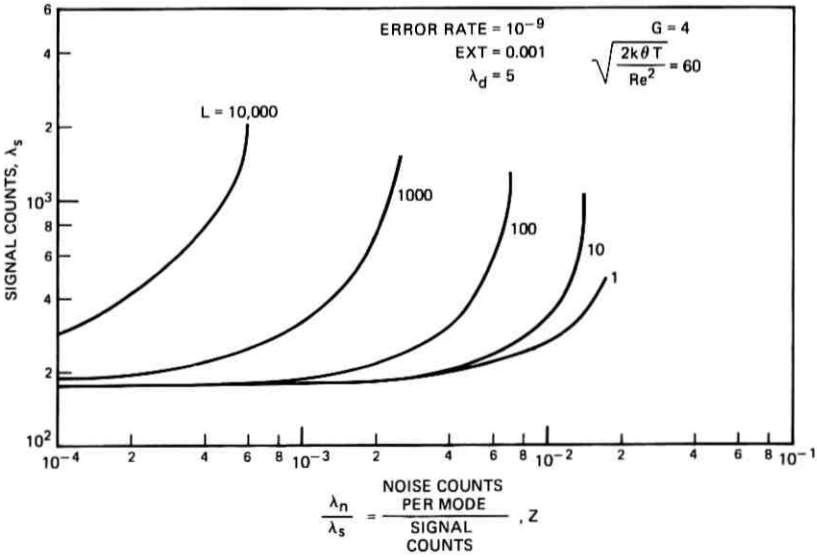


Fig. 6—Required energy per pulse normalized by $\eta/\hbar\Omega$ vs the ratio Z of spontaneous emission noise spectral height to signal energy.

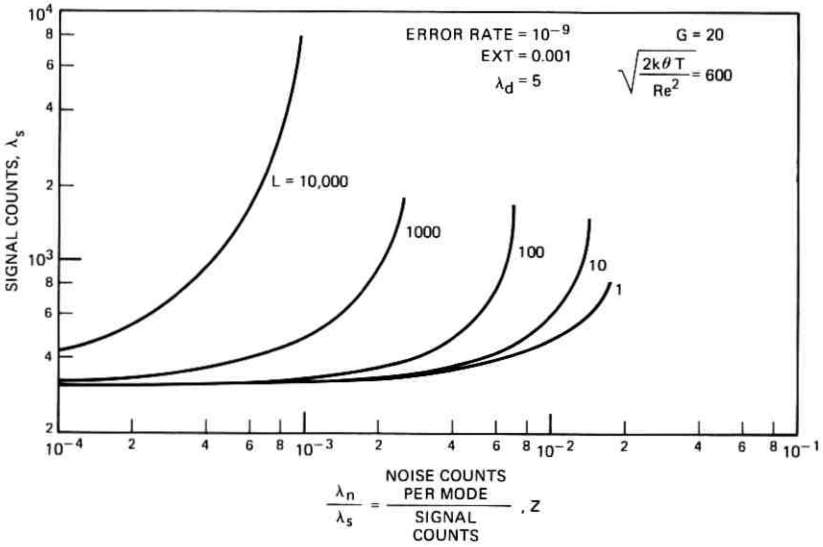


Fig. 7—Same as Fig. 6.

input will decrease while the ratio of signal energy per pulse to spontaneous emission noise spectral height at the regenerating repeater input will remain fixed. Thus in Figs. 6 to 8, the required energy per pulse normalized by $\eta/\hbar\Omega$ is plotted vs the ratio Z of spontaneous

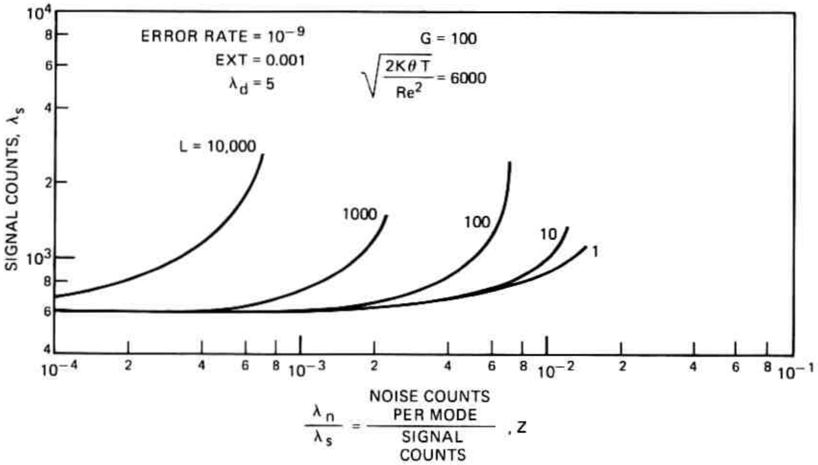


Fig. 8—Same as Fig. 6.

emission noise spectral height to signal energy, for the various values shown of other parameters.

V. APPLICATIONS AND EXAMPLES

5.1 *Analog Repeaters*

Suppose one used quantum amplifiers in analog repeaters placed between regenerating repeaters so as to increase the distance between regenerating repeaters. See Fig. 1. Each quantum amplifier introduces a spontaneous emission noise which has spectral height *referred to its input* given by[6]

$$N_{\text{input}} = F\hbar\Omega\left(\frac{G_q - 1}{G_q}\right) \tag{15}$$

where G_q is the quantum amplifier power gain and F is a noise figure which can be near unity for good quantum amplifiers and is typically less than 10.[†] If the input to analog repeater k is α_k nepers (in power) higher than the signal level at the input of the regenerating repeater, then the total spontaneous emission noise spectral height N_o at the input of the regenerating repeater is

$$N_o = \sum_1^R F_k\hbar\Omega\frac{(G_{qk} - 1)}{G_{qk}}e^{-\alpha_k} \tag{16}$$

where R = number of analog repeaters.

The ratio of N_o in (16) to the signal energy per pulse, $p \cdot T$ at the regenerating repeater input, see Fig. 2, is the parameter Z defined in Section IV above. Since the incoherent noise and the signal both experience equal loss per unit length from the fiber, the ratio Z is constant between the regenerating repeater and the analog repeater closest to it.

Example: Suppose we make the following assumptions. A twin-channel system is used with a unilateral gain detector having mean gain 100 and with all the other parameter values necessary above so that the Chernov bound curves of Fig. 8 are applicable. The source is a Nd:YAlG laser having bandwidth 1 Å at wavelength 1 μm, i.e., $3 \cdot 10^{10}$ Hz. The modulation rate is 300 Mb/s so that $T \sim 3.33 \times 10^{-9}$ s.

[†] F is related to the population inversion in the amplifying medium which is assumed constant in this analysis.

We then have $L = 100$. There are 10 analog repeaters and they are spaced so that the signal level is the same at the input to each one.

From Fig. 8 (assuming that the upper bounds are tight enough so that we can comment upon the effects of various parameters on the required energy per pulse by observing their effects upon the bounds[†]) we see that when Z is less than 10^{-3}, the required signal energy at the regenerator input is 600 counts, i.e., $p \cdot T = \hbar\Omega/\eta \cdot 600$. This value of signal energy is the same as that which would be required if no spontaneous emission noise were present ($Z = 0$).

Thus for spontaneous emission noise to be negligible in this example, we must have the ratio of the signal energy per pulse at the regenerator input to N_o larger than 10^3.

This means [from (16)] that at the analog repeater inputs the signal level must exceed $10^3 \cdot \hbar\Omega RF[(G_q-1)/G_q]$ where

R = number of analog repeaters = 10 in this example
G_q = gain of analog repeater (assumed the same for all repeaters)
F = noise figure of an analog repeater (assumed the same for all repeaters).

Looking again at Fig. 8, we see that for $L = 100$, Z can be as large as 5×10^{-3} before the required signal level at the regenerating repeater becomes large and enters the sensitive region. This means that the signal level at the inputs to the analog repeaters might be as low as $200 \cdot \hbar\Omega RF[(G_q - 1)/G_q]$ in which case the signal required at the regenerating repeater is somewhat larger, but still not extremely sensitive to small changes in Z. Suppose $F = 10$, $\eta = 1$, $G_q = 100$, and the maximum power output of any repeater is 1 mW. Suppose the loss of the medium is 10 dB/km. When spontaneous emission noise is negligible, we need $600\ \hbar\Omega = 1.2 \times 10^{-16}$ joules per pulse at the input to the regenerating repeater and we have 3.33×10^{-12} joules per pulse at the output. Without analog repeaters we can have about 44.5 dB of loss or 4.45 km between regenerating repeaters. Suppose on the other hand we use 10 analog repeaters starting where the signal level is 200 $\hbar\Omega RF[(G_q - 1)/G_q] = 4 \times 10^{-15}$ joules per pulse (i.e., $Z = 5 \times 10^{-3}$); or about 28.8 dB (2.88 km) from the regenerating repeater output. The string of 10 analog repeaters spaced at 20-dB intervals spans 200 dB or 20 km of distance; and we can have an additional 13 dB or 1.3 km of distance to the next regenerating repeater input resulting in the required 2×10^{-16} joules per pulse at that regenerating repeater input.

[†] See comment Section IV.

The total distance between regenerating repeaters is now about 24.2 km.[†]

It seems prudent that for a given value of L, one should avoid values of Z which are so large that small changes in Z result in large changes in the required signal energy at the regenerating repeater. Such small changes in Z might come about if the source power or quantum amplifier gains fluctuated slightly.

5.2 *Regenerator Repeater Front End*

Suppose that in the example above we had just one quantum amplifier (or equivalently, the spontaneous emission noise from any additional quantum amplifiers was negligible).

In the absence of spontaneous emission noise, the energy per pulse required at the regenerative repeater input is approximately 600 $\hbar\Omega/\eta$. Now suppose we place the quantum amplifier immediately before the regenerating repeater. If the gain is sufficiently large, then we can operate with Z as large as $7 \cdot 10^{-3}$. This means the energy per pulse at the input to the quantum amplifier need only be about $\hbar\Omega F/(7 \times 10^{-3}) \approx 140\hbar\Omega F$ (for large G_q). Thus, we see that if $140F < 600/\eta$, then the quantum amplifier increases the sensitivity of the regenerative repeater over that associated with an avalanche detector alone (in this example with $L = 100$).

For other values of L *in this example,* the condition for a quantum amplifier front end to increase the regenerative repeater sensitivity is

$$\frac{F}{Z_{\max}} < \frac{600}{\eta}$$

where $Z_{\max}$ is the maximum value of Z for reasonable required energy per pulse at the input to the regenerating repeater (following the quantum amplifier).

For other systems with different types of avalanche detectors and different parameters (avalanche gain, dark current, etc.) the number 600 in the above equation should be replaced by the required mean number of detected counts in the absence of a quantum amplifier.

[†] A slightly larger total distance between regenerating repeaters can be obtained by starting the chain of analog repeaters 20 dB (rather than 28.8 dB) from the regenerating repeater output. In that case $Z \cong 5 \cdot 10^{-4}$ and the next regenerating repeater can be about 45 dB from the last analog repeater for a total span of 24.5 km between regenerating repeaters. Placing the analog repeaters as described in the above example allows some margin for overload.

5.3 *Background Noise*

As a final comment, it is clear from eq. (12) that if the incoherent noise spectral height, N_o, at the regenerating repeater input is small enough so that $(\eta/\hbar\Omega)N_o \ll \overline{G^2}/(\bar{G})^2$ then only the additional shot noise term is important amongst the three noise terms associated with the incoherent noise.

This inequality always holds for the case where the incoherent noise is background (thermal) radiation in equilibrium at temperatures below 10^4 °K, since for thermal background radiation we have

$$N_o(\text{Thermal}) = \frac{\hbar\Omega}{e^{\hbar\Omega/k\theta} - 1},$$

$k\theta$ = Boltzman's constant · absolute temperature.

At room temperature and at a wavelength of 1 μm, $\hbar\Omega/k\theta \approx 50$.

Therefore, in analyses where incoherent background radiation is included, one usually only includes the additional shot noise term $LN_o(\eta/\hbar\Omega) = L\lambda_n$ in the signal-to-noise ratio formulae.

VI. CONCLUSIONS

We have shown that quantum amplifiers can have applications in both analog repeaters to extend the distance between regenerating repeaters and as front ends of regenerating repeaters. Their usefulness is a function of the ratio of the optical bandwidth of the system to the modulation bandwidth; but is not limited to small values of this ratio. To choose system parameters, for example, the required signal levels at the analog and regenerating repeater inputs, various component parameters such as the mean avalanche gain, avalanche detector type, source bandwidth, baseband thermal noise, etc., must be given. Computations in addition to those presented, upper bounding the error rates, can be carried out with previously published Chernov bound results;[1,2] or approximate error-rate calculations can be made using the signal-to-noise ratio results of Section III above.

APPENDIX

Use of the Karhunen-Loéve Expression

Starting with eq. (2) of the text, we could expand the received complex envelope $\epsilon_r(t) = m(t)e^{i\omega t} + n(t)$ in a baud interval in terms of the Karhunen-Loéve eigenfunctions of the band limited incoherent noise $n(t)$, i.e., define

$$R_n(t,u), \quad \{\psi_k(u)\}, \quad \text{and} \quad \{\gamma_k\}$$

as follows

$$R_n(t,u) = \langle n(t)n^*(u)\rangle$$

$$\int_{\text{baud interval}} \psi_k(u)R_n(t,u)du = \gamma_k\psi_k(t) \qquad \text{for } t\epsilon \text{ baud interval,}$$
$$k = 1, 2, 3 \cdots.$$

Then

$$\epsilon_r(t) = \sum_1^\infty a_k\psi_k(t) \qquad \text{for} \quad t\epsilon(0,T)$$

where

$$a_k = m_k + n_k$$

$$m_k = \int_{\text{baud interval}} m(t)e^{i\omega t}\overset{*}{\psi}_k(t)dt$$

$$n_k = \int_{\text{baud interval}} n(t)\overset{*}{\psi}_k(t)dt$$

$$\langle n_k\overset{*}{n}_j\rangle = \gamma_k\delta_{kj}, \qquad \langle n_kn_j\rangle = 0$$

and

$$\int_{\text{baud interval}} \psi_k(t)\overset{*}{\psi}_j(t)dt = \delta_{kj}.$$

Then we would find that $M_C(s)$ of eq. (9) could be more rigorously given by

$$M_C(s) = \left[\prod_{k=1}^\infty \left[1 - \frac{\eta\gamma_k}{\hbar\Omega}(e^s - 1)\right]\right]$$
$$\exp\left\{(\eta/\hbar\Omega)\sum_{k=1}^\infty \left\{|m_k|^2(e^s - 1)\Big/\left[1 - \frac{\eta}{\hbar\Omega}\gamma_k(e^s - 1)\right]\right\}\right\}.$$

Thus in eq. (9) N_o has been rigorously replaced by γ_k for each k and the finite number of terms L has been replaced by an infinite number of terms.

If we make the reasonable assumption that the incoherent noise is flat with spectral height N_o in a band of width $B' + 1/T$ then

$$\begin{aligned}\gamma_k &\approx N_o \quad \text{for} \quad 1 \geqq k \geqq L \\ &\approx 0 \quad \text{otherwise}\end{aligned} \tag{17}$$

where

$$L = B'T + 1.$$

Thus the form for $M_C(s)$ derived in the main text is identical to the more rigorous result under this approximation.

REFERENCES

1. Personick, S. D., "New Results on Avalanche Multiplication Statistics with Applications to Optical Detection," B.S.T.J., *50*, No. 1 (January 1971), pp. 167–189.
2. Personick, S. D., "Statistics of a General Class of Avalanche Detectors with Applications to Optical Communication," B.S.T.J., *50*, No. 10 (December 1971), pp. 3075–3095.
3. Steinberg, H., "The Use of a Laser Amplifier in a Laser Communication System," IEEE Proc. (June 1963), p. 943.
4. Arams, F., and Wang, M., "Infrared Laser Preamplifier System," Proc. IEEE (March 1965), p. 329.
5. Karp, S., and Clark, J. R., "Photon Counting, a Problem in Classical Noise Theory," IEEE Trans. Inform. Theory, *IT 16* (November 1970), pp. 672–680.
6. Marcuse, D., *Engineering Quantum Electrodynamics*, New York: Harcourt Brace Jovanovich, 1970, pp. 177–197.
7. Van Trees, H. L., *Detection Estimation and Modulation*, Vol. 1, New York: Wiley and Sons, 1967, pp. 118–132.
8. Arnaud, J. A., "Enhancement of Optical Receiver Sensitivities by Amplification of the Carrier," IEEE J. Quantum Elec., *QE 4* (November 1968), pp. 893–899.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 1, January, 1973
Printed in U.S.A.

A Proper Model for Testing the Planarity of Electrical Circuits

By A. J. GOLDSTEIN and D. G. SCHWEIKERT

(Manuscript received August 25, 1972)

The question of whether an electrical circuit can be laid out on a plane, without resorting to crossovers or multilayer wiring, is usually answered by testing the planarity of a graph representing the circuit.

Two commonly used representations are shown to be inadequate. We present the following new representation, and show it to be complete and unrestrictive: The graph has one node for each circuit module, and one node for each net; for every net with k modules, there is a "star" of k edges connecting the net's node to each of the modules of the net.

I. INTRODUCTION

Electrical networks frequently consist of a set of modules (beam-leaded chips, DIPs, etc.), and a set of electrical interconnections or "nets" among two or more modules. Each net specifies a set of modules to be interconnected with a single conducting path. The planar design problem consists of placing the modules and the net wiring in the plane. The question of whether the interconnections can be accomplished in the plane without resorting to crossovers or multilayer wiring is usually answered by testing the planarity of a graph representing the circuit.

This graph is typically constructed by one of two mappings:

(*i*) Module-to-Node Mapping. The modules are represented by the nodes (or points) of the graph; and the nets are represented by its edges (or lines); or

(*ii*) Module-to-Edge Mapping. The modules are represented by the edges and the nets are represented by the nodes.

Since the edge of a graph connects exactly two nodes, these mappings are not uniquely defined and *a priori* design decisions must be made which may be either improper or restrictive, and may produce spurious crossovers (see Sections II and III).

We give a unique representation that maps both nets and modules into the nodes of a graph, G. In this representation, a k-module net will appear as a "star" with an edge from its node to each of the modules in the net. We show that this mapping is a complete and unrestrictive representation of an electrical circuit. The main result of this paper (Section IV) is that the network can be laid out in the plane without crossovers if and only if G is planar. Thus the practical problem of planarity of these networks is solved since there are good computer algorithms for testing planarity.[1–5] Such algorithms will do a good but not optimal job of minimizing crossovers in a nonplanar graph. As with other mappings, we are ignoring certain practical restrictions, such as a specified cyclic terminal order for a module. Usually, these restrictions can be forced on the graph by auxiliary strategies.

The representation presented here is similar to that given by Engl and Mylnski:[6] in order to properly represent a k-node net, the conventional definition of an undirected edge, i.e., a set of two nodes, was generalized to a set of k nodes. We demonstrate here that such generalized concepts are unnecessary. By mapping both nets and modules into nodes, we retain the conventional definition of an edge, which greatly simplifies the presentation and proof, and most importantly, permits the use of conventional planarity testing algorithms.

II. INADEQUACY OF THE MODULE-TO-NODE MAPPING

Since nets map into edges, and an edge connects exactly two nodes, there is an inherent restriction to two-module nets. A common embellishment of the module-to-node mapping, is to decompose a k-module

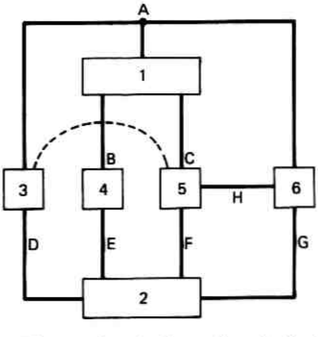


Fig. 1—Planar circuit (ignoring dashed net).

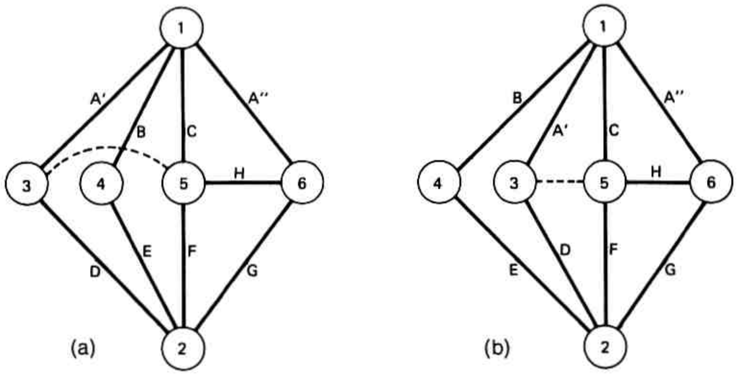


Fig. 2a.—Planar graph constructed using the module-to-node mapping on Fig. 1. Note A′ and A″ are adjacent on 1.

Fig. 2b—Alternative planar graph. Terminals for A′ and A″ are not adjacent on 1.

net ($k > 2$) into a string of $k - 1$ two-module nets.* Thus $k - 2$ of the modules are formally permitted to have two terminals contacting the same net. Since the cyclic order of edges leaving a node is irrelevant in deciding whether a graph is planar, these two terminals may not be adjacent in a planar layout of the graph. If not adjacent, these two terminals may necessitate a crossover inside the module; we will term this a "module crossing."

For example, the electrical circuit in Fig. 1 has a three-module net A. If A is represented as two two-module nets A′ (3, 1) and A″ (1, 6) then the module-to-node mapping yields a graph having a planar layout shown in Fig. 2a. Since the A′ and A″ terminals on 1 are adjacent, they can be merged, and planarity is legitimately indicated.

However, this graph has a second, and equally acceptable, planar layout (see Fig. 2b) in which the A′ and A″ terminals on Module 1 are not adjacent, and a physical realization (see Fig. 3) of this second layout may require an unnecessary crossover inside Module 1, i.e., a module crossing.

If one adds the additional net (3, 5) (shown as a dashed line in Fig. 1) then the graph has only one planar layout (Fig. 2b), and that layout requires a module crossing for its physical realization (Fig. 3).

These two examples demonstrate that the module-to-node mapping, by arbitrarily inserting two terminals per module for certain nets, cannot distinguish layouts which are physically planar from those

* The use of the complete graph for k nodes (all pair-wise connections), commonly but inaccurately used[7] in graphical representations for partitioning and placement algorithms, is clearly unacceptable here since the complete graph for five or more nodes is nonplanar.

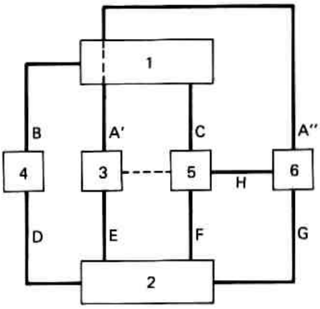


Fig. 3—Physical implementation of graph in Fig. 2b. Note "module crossing" at 1.

which use module crossings. Furthermore, if a planar layout of the graph requires the use of module crossings, there may or may not be an alternative planar layout of the graph which does *not* require the use of module crossings.

When a k-module net is decomposed into two-module connections, it is possible to choose a decomposition which will produce a nonplanar graph even though the circuit is planar. For example, the circuit in Fig. 4 is planar, and the module-to-node mapping will produce a planar graph if net A is decomposed into the string of three two-module nets: (1, 2), (2, 3), (3, 4). However, one may have chosen the alternative decomposition (1, 3), (2, 4), (3, 4) which yields the nonplanar graph shown in Fig. 5.

Certain technologies permit a limited amount of "under-module" wiring, which may permit the required module crossing in the above examples. However, even if this capability exists, there are two

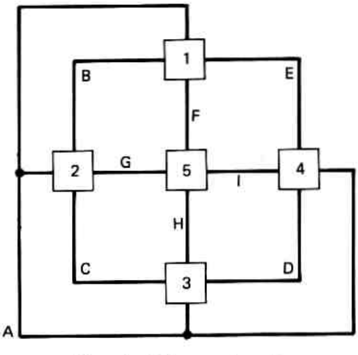


Fig. 4—Planar circuit.

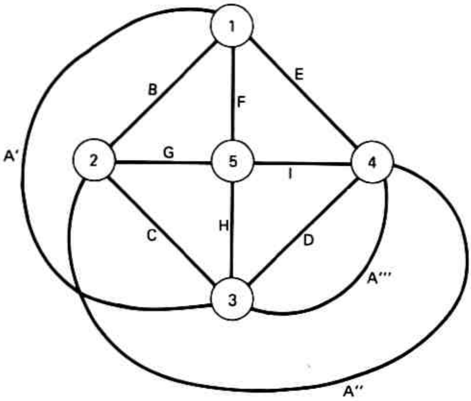


Fig. 5—Nonplanar graph resulting from an inappropriate decomposition of Net A in Fig. 4.

objections to the use of this mapping: (*i*) the set of two-module nets which best represent the k-module net ($k > 2$) is difficult to determine *a priori*, and an arbitrary choice may result in unnecessary crossovers; and (*ii*) unnecessary module crossings may result.

III. INADEQUACY OF THE MODULE-TO-EDGE MAPPING

A module which connects to $k > 2$ nets cannot be simply represented as a single edge. A typical elaboration of this mapping[8] is to represent a k-net module as a ring of k two-net modules. For example, the four-net Module 2 in the circuit above (see Fig. 1–ignoring the dashed connection) could map into the four edges shown in Fig. 6a. With similar representations for Modules 1, 5, and 6, the module-to-edge mapping for this circuit has the planar layout shown in Fig. 7.

However, without the obviously planar schematic in Fig. 1 for guidance, one may have arbitrarily chosen the equally acceptable

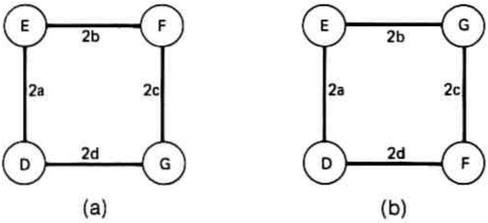


Fig. 6—Alternative decompositions of Module 2 in Fig. 1.

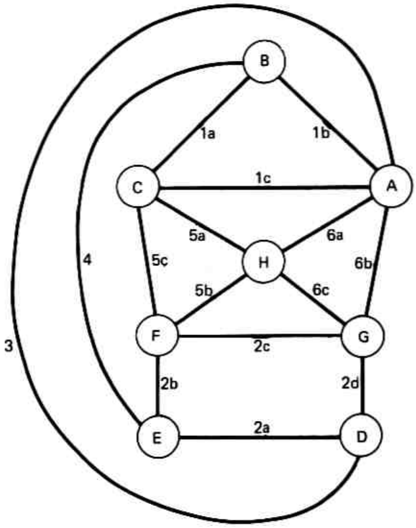


Fig. 7—Planar graph constructed using the module-to-edge mapping on Fig. 1.

representation of Module 2 shown in Fig. 6b. In this case, the graph is not planar.

Basically, the representation of a k-net module ($k > 3$) as a ring of edges, requires the specification of the sequence of terminals leaving a module–a specification which may not be required by the physical problem. As demonstrated in the above example, an arbitrary choice of terminal sequence may be restrictive and may yield a false indication of nonplanarity.

For certain designs, where the modules are predesigned and the terminal sequence *is* specified, the choice of ring sequence is obvious and not a restriction, but a practical requirement. Note, however, that the ring may appear as a mirror image in the planar layout of the graph; where the module cannot be physically mirrored, additional restrictions are necessary.

IV. MODULE-AND-NET-TO-NODE MAPPING

The previous two mappings fail to produce graphs which always reflect the planarity aspects of the circuit. In this section, we construct a graph, G, to represent the circuit and show that the circuit is planar if and only if G is planar. The graph G constructed from the net information has one node for each module plus a "net node" for every net. For every k-module net there is a "star" of k edges connecting the net

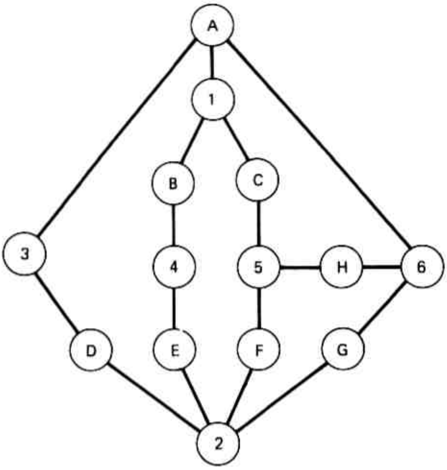


Fig. 8—Planar graph constructed using the new module-and-net-to-node mapping on Fig. 1.

node to each module in the net. Using this mapping, Fig. 8 shows the planar graph representing the circuit of Fig. 1.

The star-like subgraph is selected somewhat arbitrarily and can be replaced by any tree attached to the net's modules. Recall (Section I) that the cyclic order of edges at a module is unrestricted.

Theorem: The circuit is planar if and only if G is planar.

Proof: If G is planar, then clearly the circuit is planar. Conversely, suppose the circuit is planar. Consider the planar subgraph of any net. (Since they are electrically unnecessary, we may assume the subgraph has no loops.) We will modify it to form a star. First, create a node s at any point of the subgraph which is not a terminal. Continue to modify the subgraph by repeating the following process at s until a star subgraph results: (cf. Fig. 9).

> Choose an edge (s, t) of the modified subgraph with t having at least two edges. Let (t, u) be the first edge at t in, say, clockwise order from (t, s). Create a new subgraph by replacing the edge (t, u) by an edge (dashed in Fig. 9) from s to u "running parallel" and on the left side of the path s, t, u. If t now has only two edges, then delete t and coalesce its two edges into one.

Since the subgraph of the net was planar, the resulting star subgraph is also planar and has a node s corresponding to the net. By replacing

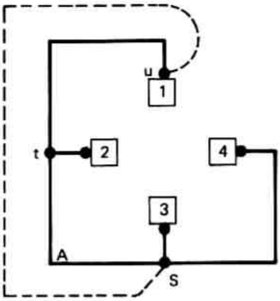


Fig. 9—Net A of planar circuit in Fig. 4.

every net subgraph by a star subgraph, we obtain a planar graph, G, of the desired type. Q. E. D.

Two observations may substantially reduce the size of the graph which is tested for planarity. Since a two-module net results in a star with only two edges, it is clear that planarity is unchanged if this net node is deleted and the two edges are coalesced into one. Similarly, a two-net module results in a node with only two edges connected to it; again, that module node can be deleted and the two edges coalesced into one.

REFERENCES

1. Auslander, L., and Parter, S. V., "On Imbedding Graphs in a Sphere," J. Math. Mech., *10*, 1961, pp. 517–523.
2. Goldstein, A. J., "An Efficient and Constructive Algorithm for Testing Whether a Graph Can Be Embedded in the Plane," Proc. Conf. on Combinatorics and Graphs, Princeton, May 1963.
3. Fisher, G. J., and Wing, O., "Computer Recognition and Extraction of Planar Graphs from the Incidence Matrix," IEEE Trans. Circ. Theory, *13*, 1966, pp. 154–163.
4. Hopcroft, J., and Tarjan, R., "Planarity Testing in V Log V Steps," Proc. of IFIP Congress, Ljubljana, Yugoslavia, 1971, Booklet TA2, pp. 18–22.
5. Tarjan, R., "An Efficient Planarity Algorithm," Stanford University Report STAN-CS-244-71, November 1971.
6. Engl, W. L., and Mylnski, D. A., "Topological Synthesis Procedure for Circuit Integration," Proc. 1969 IEEE Int. Solid-State Circ. Conf., pp. 138–139.
7. Schweikert, D. G., and Kernighan, B. W., "A Proper Model for the Partitioning of Electrical Circuits," Proc. 9th Design Automation Workshop, Dallas, 1971, pp. 57–62.
8. Rose, N. A., and Oldfield, J. V., "Printed-Wiring-Board Layout by Computer," Electronics and Power (October 1971), pp. 376–379.

Contributors to This Issue

GEN M. CHIN, Electronic Technology, 1967, RCA Institutes; Bell Laboratories, 1967—. Since joining the System Elements Research Department, Mr. Chin has been involved in the development of high-speed digital circuits and laser modulation. He is now working on Pierce data ring switching systems.

DENIS J. CONNOR, B.A.Sc., 1963, M.A.Sc., 1965, and Ph.D., 1969, University of British Columbia; Bell Laboratories, 1969—. A member of the Visual Communications Research Department, Mr. Connor is currently working on techniques for the efficient coding of television signals.

A. JAY GOLDSTEIN, B.S. (Physics), 1948, and M.A. (Mathematics), 1951, Pennsylvania State University; Ph.D. (Mathematics), 1955, Massachusetts Institute of Technology; mathematics faculty of Polytechnic Institute of Brooklyn, 1954–1957; Bell Laboratories, 1957—. Mr. Goldstein has worked on network analysis and synthesis, computer-oriented combinatoric algorithms, and interactive computing systems. He is now supervisor of the Mathematical Techniques Group.

B. GOPINATH, M.S. (Mathematical Physics), 1964, University of Bombay, India; M.S.E.E. and Ph.D. (E.E.), 1968, Stanford University; Postdoctoral Research Associate, Stanford University, 1967–1968; Bell Laboratories, 1968—. Mr. Gopinath's primary interest, as a member of the Mathematics of Physics and Networks Department, is in the applications of mathematical methods to physical problems.

BARRY G. HASKELL, B.S., 1964, M.S., 1965, and Ph.D. (Electrical Engineering), 1968, University of California; Research Assistant, University of California, 1965–68; Bell Laboratories, 1968—. Mr. Haskell is engaged in TV picture processing studies. Member, Phi Beta Kappa, Sigma Xi, IEEE.

James C. Isaacs, Jr., B.E.E., 1964, M.S.E.E., 1967, Ph.D. (E.E.), 1970, University of Virginia; Bell Laboratories, 1970—. Mr. Isaacs is a member of the Exploratory Transmission Media Department, and is currently engaged in studies relating transmission media characteristics to system performance. Member, IEEE, Eta Kappa Nu.

Dawon Kahng, B.Sc. (Physics), 1955, Seoul University, Korea; M.Sc. (E.E.), 1956, and Ph.D. (E.E.), 1959, Ohio State University; Bell Laboratories, 1959—. Mr. Kahng has worked on feasibility studies of MOS transistors and hot electron devices, and on silicon epitaxial film doping profile studies. Since 1964, he has been supervising a group concerned with the development of surface barrier high-frequency diodes, and with studies of large gap and ferroelectric semiconductors, and, more recently, luminescence in the visible and charge coupled devices. Member, IEEE, Sigma Xi, Pi Mu Epsilon, AAAS; life member, Korean Physical Society.

Bernard B. Kosicki, B.A., 1961, Wesleyan University; M.A., 1962, and Ph.D., 1967, Harvard University; Bell Laboratories, 1967—. Mr. Kosicki has worked on the growth and structural properties of GaN thin films and in areas related to the evaluation of gallium arsenide for use in silicon diode array camera tubes used in the *Picturephone®* system. He is presently concerned with advanced silicon device processing, as it applies to fabrication of large charge coupled devices. Member, Phi Beta Kappa, Sigma Xi.

Anatol Kuczura, B.S. (Engineering Physics), 1961, University of Illinois; M.S. (Mathematics), 1963, University of Michigan; M.S.E.E., 1966, New York University; Ph.D. (Mathematics), 1971, Polytechnic Institute of Brooklyn; Bell Laboratories, 1963—. From 1963 to 1966, Mr. Kuczura worked in military systems engineering. Since 1966, he has been engaged in research on the application of probability theory and stochastic processes to the analysis of telephone traffic and queuing. Member, ORSA, SIAM, American Mathematical Society, Mathematical Association of America, AAAS, Chi Gamma Iota, Pi Mu Epsilon.

Dietrich Marcuse, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of two books. Member, IEEE, Optical Society of America.

Debasis Mitra, B.Sc. (E.E.), 1964, and Ph.D. (E.E.), 1967, University of London; United Kingdom Atomic Energy Authority Research Fellow, 1965–1967; University of Manchester, U.K., 1967–1968; Bell Laboratories, 1968—. Mr. Mitra, a member of the Mathematics of Physics and Networks Department, is interested in the application of mathematical methods to physical problems.

F. W. Mounts, E.E., 1953, and M.S., 1956, University of Cincinnati; Bell Laboratories, 1956—. Mr. Mounts has been concerned with research in efficient methods of encoding pictorial information for digital television systems. Member, IEEE, Eta Kappa Nu.

S. D. Personick, B.E.E., 1967, City College of New York; S.M., 1968, E.E., 1969, and Sc.D., 1969, Massachusetts Institute of Technology; Bell Laboratories, 1967—. Mr. Personick is engaged in studies of optical communication systems.

D. G. Schweikert, B.E., 1959, Yale University; Ph.D., 1966, Brown University; Bell Laboratories, 1966—. Mr. Schweikert's current interests are in the exploratory development of computer aids to the design of large-scale integrated circuits. His previous interests involved general scientific computing, including work in underwater acoustics at General Dynamics/Electric Boat (1961–1964). Member, Sigma Xi, Tau Beta Pi, Association for Computing Machinery, Society for Industrial and Applied Mathematics.

M. M. Sondhi, B.S. (Honours), 1950, Delhi University (Delhi, India); D.I.I.Sc., 1953, Indian Institute of Science (Bangalore, India);

M.S., 1955, and Ph.D., 1957, University of Wisconsin; Bell Laboratories, 1962—. Mr. Sondhi is working on problems concerning the processing and transmission of speech signals and modeling the detection of auditory and visual signals by human beings.

Nicholas A. Strakhov, B.S.M.E., 1959, Massachusetts Institute of Technology; M.E.E., 1961, New York University; Ph.D., 1967, New York University; Bell Laboratories, 1959—. Mr. Strakhov has been designing and developing electronic test sets for transmission media maintenance. Since 1967 he has been engaged in analysis of transmission media properties with particular emphasis on crosstalk in multipair cable. He is currently supervising a group responsible for developing cable design rules. Member, Sigma Xi, Pi Tau Sigma, IEEE.

Michael F. Tompsett, B.A. (Physics), 1962, and Ph.D. (E.E.), 1966, Cambridge University, England; English Electric Valve Company, Chelmsford, England, 1966–1969; Bell Laboratories, 1969—. Mr. Tompsett is presently engaged in the development of charge coupled devices. Member, Institution of Electrical Engineers (London), Institute of Electrical and Electronic Engineers, Institute of Physics (London).

Gerard White, B.Sc., 1963, and Ph.D., 1966, University of Wales, Bangor; Bell Laboratories, 1967—. Mr. White is a member of the Electronic and Computer Systems Research Laboratory where he has been engaged in studies of Gunn effect devices, high-speed communication circuits, and optical communication systems. His current research interests are in the field of Pierce data ring switching systems. Senior Member, IEEE; Member, Sigma Xi.