THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 53 March 1974 Number 3

Codes Which Detect Deception

By E. N. GILBERT, Mrs. F. J. MacWILLIAMS, and N. J. A. SLOANE

(Manuscript received May 15, 1973)

We consider a new kind of coding problem, which has applications in a variety of situations. A message x is to be encoded using a key m to form an encrypted message $y = \Phi(x, m)$, which is then supplied to a user G. G knows m and so can calculate x. It is desired to choose $\Phi(\cdot,\cdot)$ so as to protect G against B, who knows x, y, and $\Phi(\cdot,\cdot)$ (but not m); B may substitute a false message y' for y. It is shown that if the key can take K values, then an optimal strategy for B secures him a probability of an undetected substitution $\geqq K^{-\frac{1}{2}}$. Several encoding functions $\Phi(\cdot,\cdot)$ are given, some of which achieve this bound.

I. INTRODUCTION

The gambling casino has often supplied a vivid and concrete setting for problems in probability theory,[1] stochastic processes,[2] hypothesis testing,[3] information theory,[4] and coding theory,[5] and we shall use it to describe our problem.

There are two main participants, the owner of the casino G (standing for good guy) and the manager B (the bad guy). B has been reporting the daily takings from the slot machines to be less than they actually are and keeping the difference for himself. To prevent this, G proposes to install in each slot machine a key generator of which he possesses an exact duplicate and an encoder which will encrypt the day's takings x using a key m to produce an encrypted message

$$y = \Phi(x, m). \tag{1}$$

(See Figs. 1 and 2.) The device will punch y onto a paper tape. At suitable intervals B will mail the tape to G, who will calculate x from y and m. From time to time G will visit the casino to change the key generator. We assume that B knows x and $\Phi(\cdot,\cdot)$ (but cannot change them), and y (which he can change), but does not know m. G knows y, m, and $\Phi(\cdot,\cdot)$.

If B attempts to give G a false message y_o', there may be no x' satisfying $y_o' = \Phi(x', m)$, and then G will discover B's deception. But if B can solve (1) for m, then he can successfully substitute a false message x' by giving G the correctly encrypted message $y' = \Phi(x', m)$. The problem is to design $\Phi(\cdot,\cdot)$ so as to make it as difficult as possible for B to deceive G without being caught.

Clearly, the problem is applicable to other situations (vending machines, cash registers, etc.) and in fact was first presented to us by G. J. Simmons of Sandia Corporation in connection with monitoring the production of certain materials in the interests of arms limitation.

The problem resembles the one normally encountered in cryptography in that a key m is used to encrypt a clear text x into an encoded form $y = \Phi(x, m)$. But there is an important difference. Since B knows x already, many of the standard cryptographic codes would allow B to recover the key m.

To prevent B from using (1) to learn the key, G must construct $\Phi(\cdot,\cdot)$ so that (1) has several solutions m. Then B will probably pick a wrong key m_o and G will discover that B's encrypted mesage y_o' is incompatible with the correct key. As one might expect, to provide many solutions to (1) G must use a large number K of possible keys.

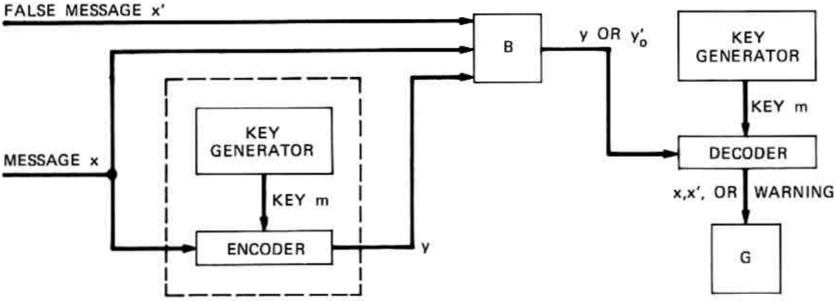


Fig. 1—Encoding to detect substitution.

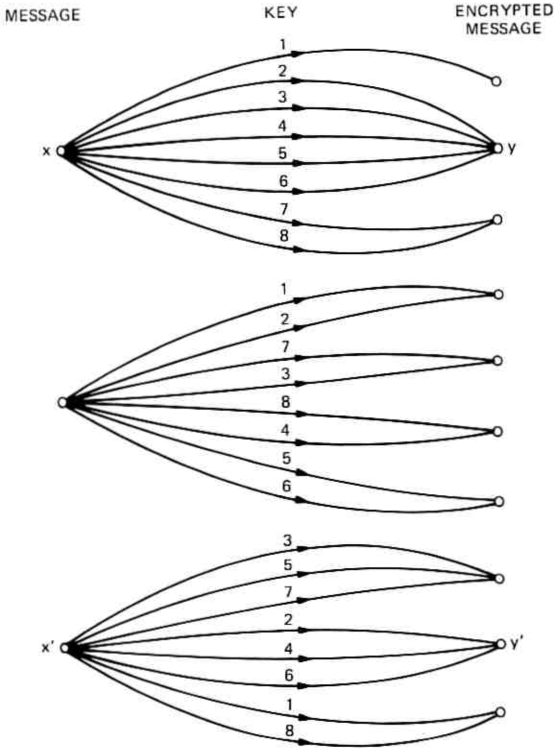


Fig. 2—Diagram of a code.

When B tries to substitute a false message, his probability of escaping detection will be called p_o. The probability p_o for an optimal B strategy will be called $p_o^{\bullet}$. We will show that $p_o^{\bullet} \geqq K^{-\frac{1}{2}}$. Although Section IV will construct a code which is best-possible in the sense of achieving $p_o^{\bullet} = K^{-\frac{1}{2}}$, this equality can be achieved only by severely restricting the number N of possible messages x. More useful codes must compromise among three conflicting goals for G: small $p_o^{\bullet}$, small K, and large N. We give two such codes, one random (Section VII) and one systematic (Section VIII).

Throughout most of this paper we imagine that B has a particular, but unknown, false message x' to substitute for x. We assume that x is equally likely to be any one of the N possibilities and that B picks x' at random from the remaining $N - 1$ messages. Then p_o is an average of the probabilities $p_o(x, y, x')$ of success when B substitutes a given x' for given x, knowing y.

In Section IX, B uses a different strategy. There B is content to succeed in *any* deception. Given x and y, B now picks x' to maximize the chance of escaping detection. Merely keeping p_o small does not protect G against this if individual terms $p_o(x, y, x')$ are large. With proper design, the systematic code of Section VIII still defeats B.

II. THE AUTHENTICATOR

A convenient special form for the encryption (1) is

$$y = (x; z), \tag{2}$$

i.e., y is the clear text x followed by a string z of extra digits or letters. Here z is some function of x and m. G will use z to test the received message y for authenticity. For this reason z will be called an *authenticator*.

Although (2) is a special case of (1), nothing is lost by restricting the encryption to this special form. Indeed, if some other $\Phi_o(x, m)$ in (1) provides a good code, one can always create a code of the form (2) by taking $z = \Phi_o(x, y)$, i.e.,

$$y = \Phi(x, m) = [x; \Phi_o(x, m)].$$

Including x as part of y cannot help B; he knows x already. Giving x to G explicitly cannot hinder him in detecting a deception by B. Thus the new code is at least as good for G as the old one.

Whether or not to use a code of the form (2) is purely a matter of convenience. However, the form (2) has a special property which we can now require without loss for all codes. It is that different clear text messages x_1, x_2 cannot be encoded into the same y, i.e.,

$$\Phi(m_1, x_1) \neq \Phi(m_2, x_2) \tag{3}$$

holds for all m_1, m_2 if $x_1 \neq x_2$. Then a typical code has a *diagram* like Fig. 2 which portrays clear messages x as points in the left column and encrypted messages y as points in the right column. The lines directed from left to right are labeled by the key names $1, \cdots, K$ to show how these keys encode each x into a y. Because of (3) the encrypted messages y fall into disjoint clusters, each cluster containing all possible images of a particular x.

III. PROBABILITY OF DECEPTION

B successfully deceives G with probability $p_o \geqq K^{-1}$ just by guessing a key m_o at random with all K keys equally likely. Better strategies use B's knowledge of x and y to restrict his guess to keys satisfying (1).

Usually B need not guess $m_o = m$, the correct key. B still succeeds if

$$\Phi(x', m_o) = \Phi(x', m). \tag{4}$$

In Fig. 2, B would pick m_o to be one of 2, 3, 4, 5, or 6; if $m = 2$ then the guesses $m_o = 2$, 4, or 6 all succeed.

An important qualitative feature of a code is the size of the bundle of lines leading from the message x to the encrypted message y in the code diagram (Fig. 2). G must make these bundles large enough to prevent B from guessing m with high probability. But if the bundles are too large, B will succeed often because many keys m_o satisfy (4). In compromising between the two extreme bundle sizes, G cannot limit B to a probability $p_o = 1/K$. In fact, we now show that B can always use a strategy which succeeds with probability

$$p_o \geqq K^{-\frac{1}{2}}. \tag{5}$$

In order to prove (5) we will have to place some natural restrictions on the behavior of G and B.

(*a*) B does not attempt to deceive G by replacing x by $x' = x$. If we allowed B that kind of "deception," B could succeed with probability $p_o = 1$ and (5) would be a weak result.

(*b*) All N messages x are equally likely. Although this requirement could be relaxed, some condition like it must be imposed to forbid G from using one particular message x_1 almost exclusively. In that case G could let all keys encrypt x_1 to the same y_1 but give all other messages x' K distinct encrypted forms. B would then have $p_o < K^{-\frac{1}{2}}$ but G would receive little information from each message.

(*c*) Another restriction on G might be that he use the K keys at random, equally likely and independent of x. We won't need this restriction on G to prove (5). If G uses the keys in any other way he only helps B increase p_o.

(*d*) We will prove that (5) holds even if B picks x' at random from the $N - 1$ messages different from x, all equally likely. This only strengthens (5) because there may be better strategies for B.

Knowing how the message x, x', and key m are distributed, we can compute the joint probability $P(x, y, x')$. This probability is the weight used in averaging $p_o(x, y, x')$ to get

$$p_o = \sum_{x,y,x'} P(x, y, x')p_o(x, y, x'), \tag{6}$$

as mentioned in Section I. The probability $p_o(x, y, x')$, that B succeeds in substituting x', knowing x and y, depends on how B uses x, y, x' to determine a false encrypted message y'_o. B knows the function $\Phi(\cdot, \cdot)$ and the key distribution. From these, he can compute the conditional probability distribution $P(y'|x, y, x')$ of the correctly encrypted false message $y' = \Phi(x', m)$. B maximizes his chance of success by using a false message y'_o which maximizes $P(y'|x, y, x')$. Then B achieves

$$p_o(x, y, x') = \operatorname*{Max}_{y'} P(y'|x, y, x') \tag{7}$$

and maximizes p_o in (6). Since (7) is optimal for B we give the corresponding p_o value a special name p_o^*.

As a preliminary to (5) we now relate p_o^* to the average uncertainty U which B has about the correctly encrypted false message y'. U is a conditional entropy

$$\begin{aligned} U &= H(y'|x, y, x') \\ &= -\sum_{x,y,x',y'} P(x, y, x', y') \log P(y'|x, y, x'). \end{aligned} \tag{8}$$

Lemma: If B chooses y'_o to make (7) hold, then

$$p_o = p_o^* \geqq 2^{-U}. \tag{9}$$

Equality holds in (9) if and only if all the possible encrypted messages y' for each (x, y, x') having $P(x, y, x) \neq 0$ are equally likely and there are exactly 2^U such y'.

The proof does not require restrictions (*a*), (*b*), (*c*), or (*d*). Use (7) to write $P(y'|x, y, x') \leqq p_o(x, y, x')$ in (8). Sum on y' and use the convexity of the function $-\log p$ to get

$$\begin{aligned} U &\geqq -\sum_{x,y,x'} P(x, y, x') \log p_o(x, y, x') \\ &\geqq -\log \sum_{x,y,x'} P(x, y, x') p_o(x, y, x'). \end{aligned}$$

Now (9) follows from (6).

The derivation used two inequalities. Both must become equalities if equality holds in (9). $P(y'|x, y, x') = p_o(x, y, x')$ requires all possible y' to be equally likely for given x, y, x'. In the convexity argument, equality requires all $-\log p_o(x, y, x')$ terms to be equal to U.

We now bound p_o^* in terms of the uncertainty $H(m)$ associated with the choice of key.

Theorem 1: Suppose (7) and restrictions (a), (b), (d) all hold. Then

$$p_o = p_o^* \geqq 2^{-\frac{1}{2}H(m)}. \tag{10}$$

First note that y' is determined by $y' = \Phi(m, x')$ if m, x' are known. Then y' contains less information than (m, x'):

$$U = H(y' | x, y, x') \leqq H(m, x' | x, y, x') = H(m | x, y, x'). \tag{11}$$

But the conditional probability for m given x, y, x' depends only on x, y, so (11) becomes

$$U \leqq H(m | x, y). \tag{12}$$

Also,

$$H(m) \geqq H(m | x) = H(m, y | x) = H(y | x) + H(m | x, y)$$

so (12) provides

$$U \leqq H(m) - H(y | x). \tag{13}$$

But

$$U = H(y' | x, y, x') \leqq H(y' | x').$$

Because of constraint (d), x' is equally likely to be any one of the N messages. Then, by (b), x and x' have the same distribution, $H(y' | x') = H(y | x)$, and finally

$$U \leqq H(y | x). \tag{14}$$

Now compare (13) and (14). If $H(y | x) \leqq \frac{1}{2}H(m)$, then $U \leqq \frac{1}{2}H(m)$ follows from (14). If $H(y | x) \geqq \frac{1}{2}H(m)$, then $U \leqq \frac{1}{2}H(m)$ follows from (13). In either case, (10) follows from the lemma.

Remark: The bound (10) implies (5), and in fact reduces to (5) when restriction (c) holds.

IV. PROJECTIVE PLANE CODES

Since p_o^* is the largest probability of success obtainable by B, a code for which equality holds in (10) guarantees G the minimum p_o against optimal behavior by B. This section designs such a code. We now assume that G behaves according to (c) of Section III, for that will make

$$p_o^* = K^{-\frac{1}{2}}.$$

If equality is to hold in (10), all the inequalities used in proving Theorem 1 must become equalities. We now review these inequalities to obtain requirements on the code.

The requirements are most easily stated in terms of the bundles of keys in the code diagram, Fig. 2.

(*i*) Every pair of bundles, from x_1 to y_1 and x_2 to y_2, with $x_2 \neq x_1$, have exactly one key in common.
(*ii*) Every bundle contains $K^{\frac{1}{2}}$ keys.
(*iii*) There are $K^{\frac{1}{2}}$ bundles at each x.

To prove (*i*), (*ii*), (*iii*), begin with (11) and write $H(y'|x, y, x') = H(m, x'|x, y, x')$. If, for some x, y, x', more than one key m satisfied $y' = \Phi(m, x')$ then there would be more conditional uncertainty about the pair (m, x') than about y'. Thus equality in (11) requires

(*i'*) Every pair of bundles, from x_1 to y_1 and x_2 to y_2, $x_2 \neq x_1$, have at most one key in common.

Equality in (9) requires that the keys in any bundle from x to y be distributed equally over $2^U = 2^{\frac{1}{2}H(m)} = K^{\frac{1}{2}}$ images y' of any x'. Each of these keys leads from x' to a different y' [by (*i'*)]. Then the bundle x to y has $K^{\frac{1}{2}}$ keys, which proves (*ii*). Now (*iii*) follows from (*ii*) because there are only K keys. Requirements (*ii*) and (*iii*) also guarantee $H(y|x) = \frac{1}{2}\log K = \frac{1}{2}H(m)$, which is needed for equality in (13) and (14).

To strengthen (*i'*) to (*i*) consider the $K^{\frac{1}{2}}$ bundles leaving x and the $K^{\frac{1}{2}}$ bundles leaving x'. There are $K^{\frac{1}{2}} \cdot K^{\frac{1}{2}} = K$ pairs of bundles. (*i'*) permits each pair to have at most one key in common. But each key is common to some pair. Since there are K keys, (*i*) must hold.

One can find trivial codes which satisfy (*i*), (*ii*), (*iii*) but which have only a few messages x. For instance, the K keys might be arranged in a $K^{\frac{1}{2}} \times K^{\frac{1}{2}}$ square matrix and each row (or column) be designated as the bundle for a distinct encrypted form of x_1 (or x_2). Since this code has $N = 2$ it is not very useful. In order to force N to be large we need another requirement.

Since (*i*) requires a pair (m_1, m_2) of different keys to belong to at most one bundle, the number of pairs of keys having a common bundle is $N\binom{K^{\frac{1}{2}}}{2}$. This number must be no greater than the unrestricted number of pairs of keys $\binom{K}{2}$, so that

$$\begin{aligned} \tfrac{1}{2}NK^{\frac{1}{2}}(K^{\frac{1}{2}} - 1) &\leqq \tfrac{1}{2}K(K-1) \\ N &\leqq K^{\frac{1}{2}} + 1. \end{aligned} \tag{15}$$

The condition for equality in (15) is

(*iv*) Every pair (m_1, m_2) of different keys belongs to exactly one common bundle.

We now add requirement (*iv*) in order to have a code with the largest possible N. Note that even for this code (15) indicates only about half as many message bits as key bits.

A code satisfying (*i*), (*ii*), (*iii*), (*iv*) can be constructed from any finite projective plane. Recall that a projective plane is a set of points and lines in which:

(*v*) Each pair of different lines has a unique point in common, and
(*vi*) Each pair of different points belongs to a unique line.

The most easily visualized projective plane is an infinite one based on the surface of a sphere. The lines and points of this projective plane are the great circles and pairs of diametrically opposite points on the sphere. A well-known technique (see Refs. 6, 7) uses a Galois field $GF(q)$, where q is a prime power, to construct a projective plane having $q^2 + q + 1$ points and $q^2 + q + 1$ lines.

The code will be obtained by using certain points and lines of a projective plane as the names of messages, keys, and bundles. First pick any line S to serve a special role. Using the sphere as a model, we call S the *equator*. Points on the equator will represent messages x. Points not on the equator will represent keys m. Lines other than the equator represent encrypted messages y (bundles). Each x and m determines a unique line (not S because it contains m) which we use as the name of y in (1).

Figure 3 shows the projective plane constructed from $GF(2)$. It has $2^2 + 2 + 1 = 7$ points. Six of the seven lines are shown as straight lines and the seventh, which we may take as the equator S, is a circle.

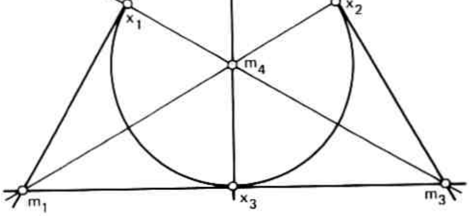


Fig. 3—A projective plane.

The three points on S are the messages and the remaining four are keys. The six straight lines are bundles, containing two keys each.

One can easily verify (*i*) and (*iv*) using (*v*) and (*vi*). Moreover, in the projective plane based on $GF(q)$, $q + 1$ lines pass through each point and $q + 1$ points lie on each line. Each line different from S contains one message x and q keys.

Each x lies on S and on q other lines. Then (*ii*) and (*iii*) hold if

$$q = K^{\frac{1}{2}}. \tag{16}$$

The equator contains $N = 1 + q = 1 + K^{\frac{1}{2}}$ points, as we expect from (15), and (*iv*) holds. When G uses this code, B will know that m is one of q keys on the line y. For any $x' \neq x$, these keys lie on q different lines through x' and B has $p_o^{\bullet} = 1/q = K^{-\frac{1}{2}}$.

A Galois field $GF(q)$ exists if and only if q is a power of a prime, $q = p^n$. Then (16) requires K to be an even power of a prime: $K = p^{2n}$ in this design.

V. IMPLEMENTATION

This section simplifies the code of Section IV into a form that is easily realized by a logic circuit.

The usual construction for a projective plane begins by defining the points as vectors, having three components taken from $GF(q)$. Two vectors $\mathbf{v}_1$, $\mathbf{v}_2$ are regarded as two names for the same point if they differ only by a scalar multiple, i.e., if $\mathbf{v}_2 = \alpha\mathbf{v}_1$ for some $\alpha \in GF(q)$. The zero vector (0, 0, 0) is not used as a point. Lines are sets of points satisfying a linear homogeneous constraint. A line L can then be described by a nonzero vector $\mathbf{L} = (a, b, c)$ with the understanding that the points on L are the vectors $\mathbf{v} = (r, s, t)$ satisfying

$$\mathbf{L}\cdot\mathbf{v} = ar + bs + ct = 0.$$

Take the equator to be the line specified by the vector $\mathbf{S} = (0, 0, 1)$. Then messages x are points having third coordinate zero. By applying appropriate scalar multipliers, each x can be written either as (0, 1, 0) or as $(1, s, 0)$ with $s \in GF(q)$. The remaining points, which can be written in the standard form $(i, j, 1)$, are the q^2 keys.

To make the logic circuit as simple as possible we agree not to use (0, 1, 0) as a message. There remain $N = q = K^{\frac{1}{2}}$ messages, all of the form $(1, s, 0)$. The q lines through $(1, s, 0)$ all have vectors $(-s, 1, c)$ where

$$si - j = c \tag{17}$$

holds for all keys $(i, j, 1)$ on the line.

Only a single element s of $GF(q)$ need be transmitted to specify the vector $(1, s, 0)$ and hence x. Likewise, the key input in Fig. 1 requires only the pair (i, j). The encrypted message y [a line with vector $(-s, 1, c)$] can be transmitted just as a pair (s, c). That amounts to using c as an authenticator z. The encoder is a computer which uses (17) to produce the authenticator value c from the inputs s, i, j. G uses a similar computer to test that his received s, c and known i, j satisfy (17).

For example, the code obtained from the projective plane of Fig. 3 is:

message	*key*	*encrypted message*
0	00 or 01	00
	10 or 11	01
1	00 or 11	10
	01 or 10	11

Again the code obtained from the projective plane with 13 points based on $GF(3) = \{0, 1, 2\}$ is:

message	*key*	*encrypted message*
0	00, 01, 02	00
	10, 11, 12	01
	20, 21, 22	02
1	00, 12, 21	10
	01, 10, 22	11
	02, 11, 20	12
2	00, 11, 22	20
	02, 10, 21	21
	01, 12, 20	22

Tables for constructing larger Galois fields will be found in Refs. 8, 9, 10, and circuits for doing arithmetic in these fields in Refs. 10, 11, 12. A field $GF(2^b)$ is convenient if the message originates in binary form. Then x and z each consist of b binary digits while $2b$ digits (b for i and b for j) are required for the key.

VI. BLOCK DESIGNS

Projective planes are special cases of more complicated structures called balanced incomplete block designs (BIBD). The technique used in Section IV generalizes directly to produce new codes based on

BIBD's. The new codes do not achieve $p_o^* = K^{-\frac{1}{2}}$, but they provide good solutions for some new values of K not of the form p^{2n}.

A (b, v, r, k, λ) *BIBD* is another system of points and sets of points. The sets are now called *blocks* instead of lines. There are v points in total and each block contains exactly k points. Each point belongs to r blocks and each pair of points is a subset of λ blocks. These conditions determine the number b of blocks. For $bk = vr$ and $r(k-1) = \lambda(v-1)$ must hold in a BIBD (Ref. 6, p. 96; Ref. 7, p. 100).

Examples:

(1) The projective plane formed from $GF(q)$ (see Section IV): $b = v = q^2 + q + 1$, $r = k = q + 1$, $\lambda = 1$.

(2) The affine plane formed from $GF(q)$ (Ref. 7, p. 176): $b = q^2 + q$, $v = q^2$, $r = q + 1$, $k = q$, $\lambda = 1$.

(3) Many other examples are known: see, for example, Refs. 6, 7, 13, 14, and recent volumes of the journals *Sankhya*, *Annals of Mathematical Statistics*, and the *Journal of Combinatorial Theory*.

Given any BIBD with $\lambda = 1$, we may form a code as follows. Proceeding as in Section IV, we select a particular block S to serve as the "equator." Points on S will represent messages x. Points not on S will represent keys m. Blocks other than the equator represent encrypted messages y (bundles). Each x and m determines a unique block different from S which we use as the name of the y in (1).

There are $N = k$ messages, $K = v - k$ keys, $b - 1$ encrypted messages, and $k - 1$ keys per bundle. Since $\lambda = 1$, the $k - 1$ keys in the bundle from x to y belong to distinct bundles leaving x'. Then $p_o^* = 1/(k-1) = 1/(N-1)$.

When the BIBD is a projective plane these formulas become again $K = q^2$, $N = 1 + K^{\frac{1}{2}}$, and $p_o^* = K^{-\frac{1}{2}}$. For affine planes $K = q^2 - q$, $N = q < 1 + K^{\frac{1}{2}}$, and $p_o^* = 1/(q-1) > K^{-\frac{1}{2}}$. Thus, for given K, the affine plane has both smaller N and larger p_o^* than one would expect from the projective plane. The larger p_o^* should be expected since (*ii*), (*iii*) fail.

To have (*ii*), (*iii*) hold, r and k should be as close as possible. In most known BIBD's other than the projective and affine planes, r and k are considerably different. For example, consider the BIBD with parameters $b = 195$, $v = 91$, $r = 15$, $k = 7$, $\lambda = 1$ (number 111 in Hall's list[7]). The code obtained from this design has $K = 84$ keys, $N = 7$ messages, and $p_o^* = \frac{1}{6}$. For comparison, the projective plane code based on $GF(9)$ is superior on all counts, having $K = 81$, $N = 10$, and $p_o^* = \frac{1}{9}$.

VII. RANDOM CODES

The projective plane code in Section IV obtains $p_o^* = K^{-\frac{1}{2}}$, the smallest possible value, but it has only $N = 1 + K^{\frac{1}{2}}$ messages. Codes with $N \gg K$ have more interest. To see how large the corresponding p_o^* might be, this section examines a code constructed at random. Now N can be made as large as desired. The main result will be that p_o^* still need not exceed $K^{-\frac{1}{2}}$ by a large factor.

The random code will have one free parameter A. Each x is allowed A possible encoded forms y. For each of the K keys the y in (1) is chosen at random from the A possibilities, all equally likely. The K choices are made independently. It may well happen that one of the A possibilities is never chosen in the K trials. In that case the code diagram, Fig. 2, will show fewer than A bundles from x. The code has a p_o^* which depends on the random choices. We will look for the expected value $E(p_o^*)$. Specific codes, with the given N and K and having p_o^* less than this expectation, surely exist.

All the data about $\Phi(\cdot,\cdot)$ that B needs when substituting x' for x are contained in a table showing how the encrypted messages y, y' depend on the key m. Figure 4 shows a convenient table as an $A \times A$ array of cells, each cell containing a list of all keys which determine a (y, y') pair. Figure 4 corresponds to the pair of messages labeled x, x' in Fig. 2. Let $\nu(y, y')$ be the number of keys in the (y, y') cell.

Knowing y, B examines the corresponding column in Fig. 4. Since the K keys are equally likely,

$$P(y' \mid x, y, x') = \nu(y, y') / \sum_{y_1} \nu(y, y_1). \tag{18}$$

The optimal strategy, by which B achieves (7), is to pick y_o' to maximize $\nu(y, y')$. In Fig. 4 the row y_o' intersects the y column in a cell with the largest number of keys. There may be $k > 1$ such cells in the y column, in which case B may as well pick one of the k rows equally likely, at random.

$E(p_o^*)$ can now be described as the solution to a distribution problem. Imagine that the correct key is key #1 and that it occupies the cell in column 1 and row 1. Distribute the $K - 1$ remaining keys at random

		y	
		3,5	7
y'		2,4,6	
	1		8

Fig. 4—Table of keys.

over the A^2 cells. Let $p_{n,k}$ be the probability that the (1, 1) cell contains $\nu(1, 1) = n$ keys, that $k - 1$ other cells in column 1 contains n keys, and that moreover all of the $A - k$ remaining cells in column 1 contain fewer than n keys. Then

$$E(p_o^*) = \sum_{n,k} k^{-1} p_{n,k} \tag{19}$$

is the probability that B picks the first row for y_o'.

The exact formula for $p_{n,k}$ is cumbersome. It is not hard to simulate the distribution experiment on a computer in order to estimate $E(p_o^*)$ when K is less than a few hundred. This has been done, but only as a check on the simpler approximate calculation which follows.

When A is large, each key has a small probability A^{-2} of belonging to the cell (y, y'). After a large number $K - 1$ of independent trials, the number $\nu(y, y')$ of keys in the cell will have approximately a Poisson distribution with mean

$$\lambda = (K - 1)/A^2. \tag{20}$$

Accordingly, we treat numbers $\nu(y, y')$ as independent Poisson random variables with mean λ. The number $\nu(1, 1)$ is special because we started the distribution by placing key #1 in cell (1, 1); $\nu(1, 1) - 1$ is the Poisson variable for this cell. Poisson approximation has the disadvantage that the total number of keys $\sum_{y,y'} \nu(y, y')$ is itself a random variable. However, the mean number of keys is K and there is high probability that there will be close to K keys if K is large. The effect of this approximation should be worse for small K than for large K. The Poisson approximation and the simulation do give the same $E(p_o^*)$ to within a few percent even for $K = 25$.

Table I — E(p_0^*) for random designs

$\lambda =$	$\frac{1}{16}$	$\frac{1}{4}$	1	4	16	$K^{-\frac{1}{2}}$
$K = 25$		0.47	0.44	0.54		0.2
64	0.46	0.34	0.32	0.38	0.57	0.125
100	0.40	0.29	0.27	0.32	0.46	0.1
256	0.27	0.21	0.19	0.22	0.32	0.06
400	0.23	0.17	0.16	0.18	0.26	0.05
1,024	0.15	0.12	0.11	0.12		0.03
4,096	0.087	0.069	0.062	0.068	0.092	0.015
10,000	0.062	0.047	0.042	0.046	0.061	0.01
40,000	0.036	0.026	0.023	0.024	0.032	0.005
100,000	0.025	0.018	0.015	0.016	0.021	0.003
1,045,576	0.0084	0.0063	0.0054	0.0055	0.0069	0.001

To simplify writing an expression for $p_{n,k}$, let b_n and B_n denote the probabilities that a Poisson random variable has value exactly n or at most n.

$$\begin{aligned} b_n &= \lambda^n e^{-n}/n\,! \\ B_n &= b_0 + b_1 + \cdots + b_n. \end{aligned}$$

Then

$$p_{n,k} = b_{n-1} b_n^{k-1} B_{n-1}^{A-k} \binom{A-1}{k-1}. \tag{21}$$

In (21), b_{n-1} is the probability that cell (1, 1) contains n keys, $b_n^{k-1}\ B_{n-1}^{A-k}$ is the probability that a particular set of $k - 1$ other cells have n keys but all $A - k$ others have $n - 1$ keys or less, and the binomial coefficient counts the different sets of $k - 1$ cells. Now insert (21) into (19) and sum on k to get

$$E(p_o^*) = \sum_{n=1}^{\infty} (n/\lambda A)\{B_n^A - B_{n-1}^A\}. \tag{22}$$

Table I gives values of $E(p_o^*)$, computed from (22). For fixed K, a broad minimum of $E(p_o^*)$ occurs near $\lambda = 1$. Then (20) shows that the minimum occurs when $A = K^{\frac{1}{2}}$, approximately. Thus, even when G designs his code by random means, he should pick A to make (*ii*) and (*iii*) of Section IV hold as nearly as possible.

Although (22) is only an approximate solution to the problem, it is also a generating function for the exact solution. Let $e(K)$ denote the exact expected value of p_o^* when the number of keys is K. Instead of $e(K)$, eq. (22) provides

$$\sum_K \frac{(\lambda A^2)^{K-1}}{(K-1)\,!} \exp\,(-\lambda A^2) e(K),$$

i.e., a sum of terms $e(K)$ weighted by the probability that the Poisson experiment produces $K - 1$ keys in addition to key #1. In principle, one could multiply the sum in (22) by exp (λA^2), expand the result into a series in powers of λ, and identify the coefficient of λ^{K-1} as $A^{2(K-1)} e(K)/(K - 1)\,!$. The result for $e(K)$ is unpleasant and (22) is accurate enough. In an experiment to estimate $e(64)$, 2000 trials were made for each of $\lambda = \frac{1}{4}$, 1, 4. The fractions of trials in which B succeeded were 0.31, 0.30, 0.37.

VIII. SYSTEMATIC CODES

This section constructs a systematic code with large N by means of another generalization of the projective plane code of Section IV.

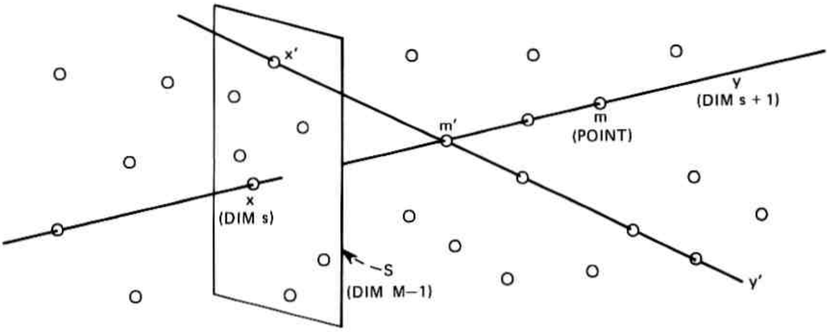


Fig. 5—Code designed from projective space of dimension M.

Unlike the random code, which had N as a free parameter, this code will specify a particular N. That disadvantage is offset by a smaller value of $E(p_o^*)$ and by a more important advantage discussed in Section IX.

Figure 5 will illustrate the code design. Given a field $GF(q)$, one can construct a projective space $PG(M, q)$ of dimension M in which points are again equivalence classes of nonzero vectors, now having $M + 1$ components. $M = 3$ in Fig. 5. The number of points is

$$f(M) = (q^{M+1} - 1)/(q - 1) = 1 + q + \cdots + q^M. \qquad (23)$$

Each set of points satisfying a system of $M - D$ independent linear homogeneous equations is a D-dimensional subspace $PG(D, q)$ containing $f(D)$ of the points of $PG(M, q)$. The number of D-dimensional subspaces of $PG(M, q)$ is[15]

$$\begin{aligned} g(D, M) &= \frac{f(M)f(M-1)\cdots f(M-D)}{f(D)f(D-1)\cdots f(0)} \\ &= \frac{(q^{M+1}-1)(q^M-1)\cdots(q^{M+1-D}-1)}{(q^{D+1}-1)(q^D-1)\cdots(q-1)}. \end{aligned} \qquad (24)$$

Proceeding as in Sections IV and VI, we again select a particular subspace S of dimension $M - 1$ to serve as the "equator." In Fig. 5, S is a projective plane. We again identify messages x with subspaces of S. But now S has subspaces of dimension $0, 1, \cdots, M - 2$ and so we can specify the dimension s of the messages as another parameter of the design. In Fig. 5, $s = 0$; another code might use $s = 1$. Given M, s, the number of distinct messages is

$$N = g(s, M - 1). \qquad (25)$$

Again, the points not in S will be keys. There are

$$K = f(M) - f(M - 1) = q^M \tag{26}$$

keys.

The key m (a point) and message x (of dimension s) determine a unique $(s + 1)$-dimensional space which will represent y. Since y has $f(s + 1)$ points and $f(s)$ of them belong to S, y contains $f(s + 1) - f(s) = q^{s+1}$ keys. Now (*ii*), (*iii*) of Section IV need not hold. Instead, for each x, the q^M keys fall into

$$A = q^{M-s-1} \tag{27}$$

bundles of

$$K/A = q^{s+1}$$

keys each. In Fig. 5, $A = q^2$, $K/A = q$.

To find p_o^* consider the matrix, Fig. 4, corresponding to a particular pair x, x'. The q^{s+1} keys in a given column y need not be distributed one to a row [as in (*i*) of Section IV]. Each cell in the matrix contains all the keys belonging to an intersection between $(s + 1)$-dimensional spaces through x and x'. If x and x' themselves intersect in an r-dimensional space $x \cap x'$ then the cell contains the q^{r+1} keys of an $(r + 1)$-dimensional space through $x \cap x'$. B must choose one of $q^{s+1}/q^{r+1} = q^{s-r}$ equally likely rows; his probability of correctly guessing y' is

$$p_o(x, y, x') = q^{r-s}. \tag{28}$$

Now (6) and (28) provide

$$p_o^* = \sum_r h(r)q^{r-s}, \tag{29}$$

where $h(r)$ is the probability that a randomly chosen x' intersects a specific x in a space of dimension r. In (29), the range of summation is $2s + 1 - M \leqq r \leqq s - 1$ provided $2s + 1 \geqq M$. But if $2s + 1 < M$, as in Fig. 5, then $x \cap x'$ can be empty. In that case the summation (29) extends over $-1 \leqq r \leqq s - 1$.

We now show

$$h(r) = q^{(s-r)^2}g(s - r - 1, M - s - 2)g(r, s)/\{g(s, M - 1) - 1\}, \tag{30}$$

which together with (24) and (29) gives p_o^*. The factor $g(r, s)$ in (30) is the number of different r-dimensional subspaces of x; it suffices to show that the remaining terms of (30) give the probability that a randomly chosen x' intersects x in a particular subspace H of dimension r. Given x, and a subspace H, we can find M basis vectors e_0, e_1,

$\cdots, e_M$ for S such that $e_0, e_1, \cdots, e_r$ span H, and $e_0, e_1, \cdots, e_s$ span x. Each x' contains H and so has a basis containing $e_0, \cdots, e_r$. The remaining $s - r$ basis vectors of x' can have the form

$$v_j = \sum_{j=r+1}^{M} \xi_{i,j} e_j, \qquad i = r + 1, \cdots, s,$$

in which $e_0, e_1, \cdots, e_r$ do not appear. In determining $\xi_{i,j}$ one must not allow x' to intersect x in a space of dimension larger than r. This requirement is equivalent to a condition that the partial sums

$$v_i^o = \sum_{j=s+1}^{M} \xi_{i,j} e_j, \qquad i = r + 1, \cdots, s,$$

of v_i be linearly independent. Then the v_i^o span an $(s - r - 1)$-dimensional subspace x^o of the $(M - s - 2)$-dimensional subspace S^o spanned by $e_{s+1}, \cdots, e_M$. The factor $g(r - s - 1,\ M - s - 2)$ in (30) is the number of ways of choosing x^o. Having chosen H and x^o (and hence ξ_{ij} for $j = s + 1, \cdots, M$), the $(s - r)^2$ numbers

$$\xi_{ij}; \qquad i = r + 1, \cdots, s; \qquad j = r + 1, \cdots, s$$

can be chosen in $q^{(s-r)^2}$ ways to specify x' completely. Now the numerator in (30) is the number of ways of picking an x' to have an r-dimensional intersection with x and the denominator is the number $N - 1$ of messages (different from x) from which B chooses x'.

Now q, M, and s determine N, K, A, p_o^*. Table II gives some of the better designs obtained by taking $q = 2$. These all have $M = 2s + 2$, so that $K/A^2 = 1$ follows from (26) and (27). For given K, the least

Table II — Designs with q = 2

Dimensions		Keys	Inputs	Prob (B wins)
M	s	K	N	p_o^*
2	0	4	3	0.6666
4	1	16	35	0.400
6	2	64	1,395	0.2222
8	3	256	200,787	0.1176
10	4	1,024	1.09×10^8	0.0606
12	5	4,096	2.3×10^{11}	0.0308
14	6	16,384	2×10^{15}	0.0155
16	7	65,536	6×10^{19}	0.0078
18	8	262,144	8×10^{24}	0.0039
20	9	1,048,576	4×10^{30}	0.00195

Table III — Design with q = 2, M = 12, s = 5

$r = \dim(x \cap x')$	$h(r)$	$p_o(x, y, x')$
-1	0.3979	0.015625
0	0.5773	0.03125
1	0.1204	0.0625
2	0.00432	0.125
3	2.9×10^{-5}	0.25
4	3.4×10^{-8}	0.5

p_o^* was always obtained when $K/A^2 = 1$; a similar phenomenon was encountered with random designs having $\lambda = 1$ [cf. eqs. (20)]. The table contains codes having N much larger than K. At the same time, p_o^* is approximately $2/K^{\frac{1}{2}}$, which compares well with the projective plane code.

IX. CHOICE OF x'

Until now B had no control over the choice of x'. We treated x' as a random variable which B accepts as given. But suppose that B has no particular x' in mind; he merely wants to mislead G by substituting any convenient wrong message x'. An optimal strategy for B must again achieve (7) but B will select x' to maximize $p_o(x, y, x')$ for each given x, y.

A code with small p_o^*, for randomly chosen x', may now be a poor one. Table III shows more detail about the code with $q = 2$, $M = 12$, $s = 5$ in Table II. This code had $p_o^* = 0.0308$, as computed from (29). But some false messages x' intersect x in spaces of dimension $r = 4$; if B substitutes one of these, his chance of success is 0.5 [eq. (28)].

Table IV — Effect of changing field, keeping key size approximately fixed

Field	Dimensions		Key bits	K/A^2	Msg bits	Prob (B wins)	
q	M	s	$\log_2 K$		$\log_2 N$	if $r = s - 1$	averaged
256	2	0	16	1	8.01	0.0039	0.0039
41	3	1	16.08	41	10.7	0.0244	0.0250
16	4	1	16	1	16.1	0.0625	0.0078
9	5	2	15.9	9	19.2	0.1111	0.0137
7	6	2	16.86	1	25.5	0.1429	0.0058
5	7	3	16.24	5	28.3	0.2000	0.0096
4	8	3	16	1	32.5	0.2500	0.0078
3	10	4	15.9	1	40.5	0.3333	0.0082
2	16	7	16	1	65.9	0.5000	0.0078

The code is good for randomly chosen x' only because B usually has a message x' with $r = -1$ or 0.

A good code for G must now have $p_o(x, y, x')$ small uniformly, not just on the average. The code of Section VIII achieves this if q is large. For (28) shows $p_o(x, y, x') \leqq 1/q$. Unfortunately for G, increasing q has the effect of decreasing N. Then G must compromise, picking q small enough to obtain large N but large enough so that B's chance of success, $1/q$, is tolerably small. Table IV shows a typical tradeoff between N and $1/q$. The designs in Table IV all have approximately the same key size $K = 2^{16}$. Table IV shows both probabilities of success for B, $1/q$ if B makes $r = s - 1$ and the averaged value (29) if B picks x' at random. If one ignores the designs with $K/A^2 \not\doteq 1$, the averaged probability doesn't change much. To reduce $1/q$ from 0.5 to 0.1 reduces the message size, $\log N$, by a factor of 3.

REFERENCES

1. W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, 2nd edition, New York: Wiley, 1957.
2. L. E. Dubins and L. J. Savage, *How to Gamble if You Must—Inequalities for Stochastic Processes*, New York: McGraw-Hill, 1965.
3. T. M. Cover and M. E. Hellman, "The Two-Armed-Bandit Problem with Time-Invariant Finite Memory," IEEE Trans. Info. Theory, *IT-16*, No. 2 (March 1970), pp. 185–195.
4. J. L. Kelly, Jr., "A New Interpretation of Information Rate," B.S.T.J., *35*, No. 4(July 1956), pp. 917–926.
5. S. W. Golomb, "Run-Length Encodings," IEEE Trans. Info. Theory, *IT-12*, No. 3 (July, 1966), pp. 399–401.
6. H. J. Ryser, *Combinatorial Mathematics*, Carus Math. Monograph 14, Math. Assoc. America, distributed by Wiley, N. Y., 1963.
7. M. Hall, Jr., *Combinatorial Theory*, Waltham, Mass.: Blaisdell, 1967.
8. E. J. Watson, "Primitive Polynomials (Mod 2)," Math. Comp., *41*, No. 79 (July 1962), pp. 368–370.
9. J. D. Alanen and D. E. Knuth, "Tables of Finite Fields," Sankhya, *26*, 1964, pp. 305–328.
10. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2nd edition, Cambridge, Mass.: M.I.T. Press, 1972.
11. T. C. Bartee and D. I. Schneider, "Computation with Finite Fields," Information and Control, *6*, No. 1 (January 1963), pp. 79–98.
12. E. R. Berlekamp, *Algebraic Coding Theory*, New York: McGraw-Hill, 1968.
13. S. Vajda, *Patterns and Configurations in Finite Spaces*, Griffin's Statistical Monograph 22, New York: Hafner, 1967.
14. S. Vajda, *The Mathematics of Experimental Design*, Griffin's Statistical Monograph 23, New York: Hafner, 1967.
15. R. D. Carmichael, *Introduction to the Theory of Groups of Finite Order*, reprinted by Dover, N. Y., 1956.

Copyright © 1974 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 3, March 1974
Printed in U.S.A.

Formulas on Queues in Burst Processes—II

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

(Manuscript received July 26, 1973)

Queues arising in buffers due to either random interruptions of the channel or variable source rates are analyzed in the framework of a single digital system. Two motivating applications are: (i) multiplexing of data with speech on telephone channels and (ii) buffering of data generated by the coding of moving images in Picturephone® *service.*

In the model a source feeds data to a buffer at a uniform rate. The buffer's access to a channel with fixed maximum rate of transmission is controlled by a switch; only when the switch is closed ("on") is the buffer able to discharge. The on-off sequence of the switch is indicated by a burst process which is a key element in this paper. In such a process, long periods during which the switch stays closed alternate with periods, called bursts, during which the on-off sequence is a first-order Markov process. The length of a burst is randomly distributed. This is a generalization of the memoryless burst process considered in an earlier paper.[1] *In that paper we gave formulas for the efficient computation of various functionals of the queues arising in the system. Now we extend these formulas to hold for the generalized class of burst processes.*

I. INTRODUCTION

In a recent paper[1] we considered the problem of buffering the output of a uniform source whose access to a given transmission channel is controlled by a burst process. We gave formulas for efficiently computing various functionals of queues that form in such a communication system when the controlling burst process is memoryless.

In the present paper we generalize the controlling process to one which is first-order Markov within a burst. This generalization considerably increases the usefulness of the formulas. Consider, for example, the two motivating applications discussed in Ref. 1: (*i*) multi-

* The sequence of names was decided by coin tossing.

plexing of data with speech on telephone channels[2–6] and (*ii*) buffering of data generated by the coding of moving images in *Picturephone*® service.[7] For the first application, analysis of data shows[2] that it is necessary to go to a first-order Markov process to adequately model the burst phenomena in speech signals. In the *Picturephone* application, although the correlation of data rates within a frame is negligible, it is quite significant from frame to frame.[8] For frame-to-frame coding, therefore, the present model with memory becomes necessary.

The system under consideration is shown in Fig. 1. The source emits data uniformly at the rate of 1 symbol per unit time. The transmission rate of the channel is $(k + 1)$ symbols per unit time, where k is some positive integer. The on-off pattern of the switch is indicated by a binary burst process: $E(j)$ is either 0 or 1 for $j = 0, 1, 2, \cdots$. If $E(j) = 0$ the switch is closed for the time duration $[j, j + 1)$; otherwise, the switch is open. We assume that there are long periods during which $E(j) = 0$ and that at the end of every such period the buffer is empty. The activity separated by such periods we call a burst. We assume bursts to be independent of each other, and the burst length to have a probability distribution which is either geometric or is a weighted sum of geometric distributions. Within a burst, $\{E(j)\}$ is assumed to be a homogeneous two-state Markov chain with transition probabilities θ_1 and θ_2 given by

$$\theta_1 \triangleq \text{Prob.}\ \{E(j+1) = 1 | E(j) = 0\} \tag{1a}$$

$$\theta_2 \triangleq \text{Prob.}\ \{E(j+1) = 0 | E(j) = 1\}, \qquad j = 0, 1, 2, \cdots. \tag{1b}$$

These two parameters completely specify the Markov chain; the probabilities of the other two possible transitions are, of course, given by

$$1 - \theta_1 = \text{Prob.}\ \{E(j+1) = 0 | E(j) = 0\}$$

and

$$1 - \theta_2 = \text{Prob.}\ \{E(j+1) = 1 | E(j) = 1\}.$$

We shall assume that $0 < \theta_1 < 1$ and $0 < \theta_2 < 1$. If $\theta_1 + \theta_2 = 1$, E_j becomes a Bernoulli sequence of independent random variables, which is the case treated in Ref. 1.

In subsequent sections of this paper we will obtain the results summarized below.

In (*i*), (*ii*), and (*iii*), we assume the switch to be controlled by an infinitely long sequence generated by the Markov chain described by (1); these three results are therefore of interest in situations where the distribution of burst lengths is not known accurately.

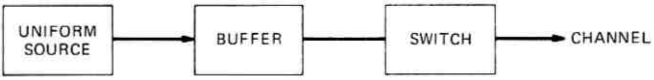


Fig. 1—Switched communication system.

(*i*) We derive a recursive formula for the steady-state distribution of buffer content for finite buffers, the recursion being with respect to the buffer size, N.

(*ii*) Let $T^{(N)}$ be the steady-state probability of a buffer of size N being full when the channel is inaccessible. ($T^{(N)}$, therefore, is the steady-state probability of a transmission fault.) We show that

$$\frac{1}{T^{(N+k+1)}} = \frac{1}{1-\theta_2}\frac{1}{T^{(N+k)}} + \frac{1-\theta_1-\theta_2}{1-\theta_2}\frac{1}{T^{(N+1)}} - \frac{1-\theta_1}{1-\theta_2}\frac{1}{T^{(N)}},$$

where $(k+1)$, as previously defined, is the transmission rate of the channel. We show that the steady-state probability of the buffer being full is $T^{(N)}/(1-\theta_2)$, and therefore satisfies the same recursive relation.

(*iii*) For a buffer of size greater than N, let $F^{(N)}$ denote the mean time to first passage through the level N. We show that $F^{(N)}$ satisfies the recursion

$$F^{(N+k+1)} = \frac{1}{1-\theta_2}F^{(N+k)} + \frac{1-\theta_1-\theta_2}{1-\theta_2}F^{(N+1)} - \frac{1-\theta_1}{1-\theta_2}F^{(N)} + \frac{\theta_1+\theta_2}{1-\theta_2}.$$

The next two results are of interest when the distribution of burst lengths is well-approximated by a weighted sum of geometric distributions.

(*iv*) Let $G^{(N)}$ be the probability of overflow for a buffer of size N *during a burst*. Then if the burst lengths have a geometrical probability distribution with parameter $\rho\{$i.e., Prob. (burst length $= i$) $= \rho^{i-1}(1-\rho)\}$, we show that

$$\frac{1}{G^{(N+k+1)}} = \frac{1}{\rho(1-\theta_2)}\frac{1}{G^{(N+k)}} + \frac{\rho(1-\theta_1-\theta_2)}{1-\theta_2}\frac{1}{G^{(N+1)}} - \frac{1-\theta_1}{1-\theta_2}\cdot\frac{1}{G^{(N)}}.$$

This result generalizes to the case when the burst length distribution is a sum of geometric distributions.

(*v*) We derive a closed expression as well as a recursive formula for the mean time for first passage through a level N *during a burst* conditioned on the occurrence of an overflow. The recursion is with respect to N, and the bursts are assumed to be distributed as in (*iv*).

(*vi*) We determine the asymptotic behavior of all the formulas in (*i*) to (*v*) as $N \to \infty$. For instance, we prove that, as $N \to \infty$, $(1/G^{(N)}) \sim s^N$, where s is the unique positive real root of a particular polynomial, such that $s > 1/\rho > 1$.

The closed expressions are all valid for $k \geqq 1$ and $N \geqq 0$, and the recursions as stated above are valid for $N \geqq 0$. The recursive formulas provide very efficient means for computation of the various functionals, particularly in design studies where a whole range of buffer sizes is to be investigated.

1.1 Notation

Whenever necessary we will use a superscript in parentheses, e.g., $x^{(M)}$, to indicate that the quantity corresponds to a buffer of size M (or to the level M in a buffer of size greater than M). If $\mathbf{x}$ is a vector, then the superscript (M) will also indicate that the vector $\mathbf{x}^{(M)}$ is $(M+1)$-dimensional with components $x_i^{(M)}$, $i = 0, 1, 2, \cdots, M$. These two uses of the superscript are consistent because the dimensions of all vectors defined in this paper are related to buffer size (level) in this manner. Whenever the superscript is missing, the standard value (N) will be implied.

We will use lower-case boldface letters to denote column vectors, upper-case boldface letters to denote matrixes, and a superscript T to denote the transpose. We will denote by $\mathbf{I}$ the identity matrix, by $\mathbf{1}$ the vector whose components are all equal to 1, and by $\mathbf{e}_j$ the vector whose jth component is 1 and the rest 0, e.g., $\mathbf{e}_0^T = (1, 0, \cdots, 0)$.

II. EQUATIONS OF THE PROCESSES

Let $B(t)$ be the number of symbols in the buffer at time t. Then for a buffer of size N

$$B(t+1) = \text{Max}\,[B(t) - k, 0] \quad \text{if} \quad E(t) = 0 \tag{2a}$$

$$= \text{Min}\,[B(t) + 1, N] \quad \text{if} \quad E(t) = 1. \tag{2b}$$

In the last equation the assumption is that if the channel is inaccessible and the buffer is full, then the current source symbol is discarded and the buffer remains full.

In order to study the evolution of the buffer content process, it is convenient to introduce two $(N+1)$-dimensional vectors $\mathbf{p}(t)$

$= \{p_o(t), \cdots, p_N(t)\}$ and $\mathbf{q}(t) = \{q_o(t), \cdots, q_N(t)\}$ defined by the equations

$$p_i(t) \triangleq \Pr\{B(t) = i,\, E(t) = 0\}, \qquad i = 0, \cdots, N \tag{3a}$$

$$q_i(t) \triangleq \Pr\{B(t) = i,\, E(t) = 1\}, \qquad i = 0, \cdots, N. \tag{3b}$$

Under the assumption that $\{E(t)\}$ is the two-state Markov chain defined by (1), it is straightforward to show that $\mathbf{p}(t)$ and $\mathbf{q}(t)$ represent a $2(N + 1)$-state homogeneous Markov chain. For

$$\begin{aligned}
p_0(t+1) &\triangleq \Pr\{B(t+1) = 0,\, E(t+1) = 0\} \\
&= \sum_{i=0}^{k} \Pr\{B(t) = i,\, E(t+1) = 0,\, E(t) = 0\} \\
&= \sum_{i=0}^{k} \Pr\{E(t+1) = 0 \mid E(t) = 0,\, B(t) = i\} \\
&\qquad\qquad\qquad \times \Pr\{B(t) = i,\, E(t) = 0\} \\
&= (1 - \theta_1) \sum_{i=0}^{k} p_i(t),
\end{aligned} \tag{4}$$

where the last step follows from the Markov property of $\{E(t)\}$. Similarly,

$$\begin{aligned}
p_i(t+1) &= (1 - \theta_1)p_{i+k}(t) + \theta_2 q_{i-1}(t), \qquad i = 1, 2, \cdots, N - k, \\
&= \theta_2 q_{i-1}(t), \qquad i = N - k + 1, \cdots, N - 1, \\
&= \theta_2\{q_{i-1}(t) + q_N(t)\}, \qquad i = N.
\end{aligned} \tag{5}$$

Also

$$\begin{aligned}
q_i(t+1) &= \theta_1 \sum_{j=0}^{k} p_j(t), \qquad i = 0, \\
&= \theta_1 p_{i+k}(t) + (1 - \theta_2)q_{i-1}(t), \qquad i = 1, 2, \cdots, N - k, \\
&= (1 - \theta_2)q_{i-1}(t), \qquad i = N - k + 1, \cdots, N - 1, \\
&= (1 - \theta_2)\{q_{i-1}(t) + q_i(t)\}, \qquad i = N.
\end{aligned} \tag{6}$$

Equations (4), (5), and (6) can be written conveniently in matrix notation as

$$\mathbf{p}(t+1) = (1 - \theta_1)\mathbf{B}\mathbf{p}(t) + \theta_2\tilde{\mathbf{A}}\mathbf{q}(t) \tag{7a}$$

$$\mathbf{q}(t+1) = \theta_1\mathbf{B}\mathbf{p}(t) + (1 - \theta_2)\tilde{\mathbf{A}}\mathbf{q}(t). \tag{7b}$$

Here the $(N+1) \times (N+1)$ matrixes $\mathbf{B}$ and $\tilde{\mathbf{A}}$ are defined as

$$\mathbf{B} \triangleq \left[\begin{array}{cccc} \overbrace{1 \quad 1\cdots1}^{(k+1)} & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \\ & & & \\ & & & \end{array}\right] \begin{array}{l} 0 \\ 1 \\ \vdots \\ N-k, \\ \vdots \\ N \end{array} \qquad \tilde{\mathbf{A}} \triangleq \left[\begin{array}{ccccc} 0 & & & & 0 \\ 1 & 0 & & & \\ & 1 & \ddots & & \\ & & \ddots & 0 & \\ & 0 & & 1 & 1 \end{array}\right]. \tag{8}$$

Notice that the composite matrix

$$\left[\begin{array}{cc} (1-\theta_1)\mathbf{B} & \theta_2\tilde{\mathbf{A}} \\ \theta_1\mathbf{B} & (1-\theta_2)\tilde{\mathbf{A}} \end{array}\right] \tag{9}$$

is stochastic (nonnegative elements and every column sums to 1) and independent of t. Equations (7a), (7b) are, therefore, the transition equations of a $2(N+1)$-state homogeneous Markov chain.

2.1 Equations for some new probabilities

For many of the derivations in the succeeding sections (e.g., mean first passage time, probability of no overflow, etc.) it is convenient to define certain new probabilities $r_i(t)$ and $s_i(t)$, $i = 0, 1, \cdots, N$. Consider a buffer of size greater than N and let $X(t)$ be the event $\bigcap_{s=0}^{t} \{B(s) \leqq N\}$, i.e., the event that $B(s)$ does not exceed N at any of the time instants $s = 0, 1, 2, \cdots, t$. Then

$$r_i(t) \triangleq \Pr\{B(t) = i,\, E(t) = 0,\, X(t)\}, \qquad i = 0, \cdots, N, \tag{10a}$$

$$s_i(t) \triangleq \Pr\{B(t) = i,\, E(t) = 1,\, X(t)\}, \qquad i = 0, \cdots, N. \tag{10b}$$

We define the $(N+1)$-dimensional vectors $\mathbf{r}(t)$ and $\mathbf{s}(t)$ with components $\{r_o(t), \cdots, r_N(t)\}$ and $\{s_o(t), \cdots, s_N(t)\}$, respectively.

In a manner analogous to the derivation of eqs. (7a) and (7b), we can derive recurrence relations giving $\mathbf{r}(t+1)$, $\mathbf{s}(t+1)$ in terms of $\mathbf{r}(t)$, $\mathbf{s}(t)$. Thus, for $i = 0, 1, \cdots, N$,

$$\begin{aligned} r_i(t+1) &\triangleq \Pr\{B(t+1) = i,\, E(t+1) = 0,\, X(t+1)\} \\ &= \Pr\{B(t+1) = i,\, E(t+1) = 0,\, X(t)\} \\ &= (1-\theta_1)\Pr\{B(t+1) = i,\, E(t) = 0,\, X(t)\} \\ &\qquad + \theta_2 \Pr\{B(t+1) = i,\, E(t) = 1,\, X(t)\}, \end{aligned} \tag{11}$$

where the last equation follows from the Markov property of $\{E(t)\}$. As before, $B(t+1)$ and $E(t)$ determine the possible values of $B(t)$

and we get

$$\begin{aligned} r_i(t+1) &= (1-\theta_1)\sum_{j=0}^{k} r_j(t), \qquad i=0, \\ &= (1-\theta_1)r_{i+k}(t) + \theta_2 s_{i-1}(t), \qquad i = 1, 2, \cdots, N-k, \\ &= \theta_2 s_{i-1}(t), \qquad i = N-k+1, \cdots, N. \end{aligned} \tag{12}$$

Comparison of eq. (12) with eqs. (4) and (5) shows that for $t = 0, 1, \cdots,$

$$\mathbf{r}(t+1) = (1-\theta_1)\mathbf{Br}(t) + \theta_2\mathbf{As}(t), \tag{13}$$

where $\mathbf{A}$ is obtained from $\tilde{\mathbf{A}}$ by setting to 0 the single nonzero entry on its main diagonal, i.e.,

$$\mathbf{A} = \tilde{\mathbf{A}} - \mathbf{e}_N\mathbf{e}_N^T. \tag{14}$$

Analogously to (13) we can also show that

$$\mathbf{s}(t+1) = \theta_1\mathbf{Br}(t) + (1-\theta_2)\mathbf{As}(t). \tag{15}$$

The transition equations (13) and (15), although very similar to eqs. (7a) and (7b), differ fundamentally from them in that $\mathbf{A}$, and consequently the matrix

$$\begin{bmatrix} (1-\theta_1)\mathbf{B} & \theta_2\mathbf{A} \\ \theta_1\mathbf{B} & (1-\theta_2)\mathbf{A} \end{bmatrix}, \tag{16}$$

are not stochastic.

We close this section by deriving from (13) and (15) a useful second-order recursion involving $\mathbf{s}(t+2)$, $\mathbf{s}(t+1)$, and $\mathbf{s}(t)$. Multiplying (13) by θ_1, (15) by $(\theta_1 - 1)$, and adding we get

$$\theta_1\mathbf{r}(t+1) = (1-\theta_1)\mathbf{s}(t+1) - (1-\theta_1-\theta_2)\mathbf{As}(t). \tag{17}$$

From (15),

$$\mathbf{s}(t+2) = \theta_1\mathbf{Br}(t+1) + (1-\theta_2)\mathbf{As}(t+1). \tag{18}$$

Premultiplying (17) by $\mathbf{B}$ and adding to (18) gives

$$\mathbf{s}(t+2) = [(1-\theta_1)\mathbf{B} + (1-\theta_2)\mathbf{A}]\mathbf{s}(t+1) - (1-\theta_1-\theta_2)\mathbf{BAs}(t), \qquad t = 0, 1, 2, \cdots. \tag{19}$$

As we will have to refer frequently to the recursion (19) it is convenient to define

$$\mathbf{C} \triangleq [(1-\theta_1)\mathbf{B} + (1-\theta_2)\mathbf{A}]$$

and

$$\mathbf{D} \triangleq -(1-\theta_1-\theta_2)\mathbf{BA} \tag{20}$$

so that eq. (19) becomes

$$\mathbf{s}(t+2) = \mathbf{C}\mathbf{s}(t+1) + \mathbf{D}\mathbf{s}(t), \qquad t = 0, 1, 2, \cdots. \tag{21}$$

III. INFINITELY LONG SEQUENCES

When the burst length distribution is not known, useful information can still be obtained by considering the behavior of the buffer content when the switch in Fig. 1 is controlled by infinitely long sequences generated by the Markov chain (1). In this section we derive various functionals for such a situation.

3.1 Stationary distributions for finite buffers

In eqs. (7a), (7b), if we set $\mathbf{p}(t+1) = \mathbf{p}(t) = \mathbf{p}$ and $\mathbf{q}(t+1) = \mathbf{q}(t) = \mathbf{q}$, then the vectors $\mathbf{p} = \{p_0, \cdots, p_N\}$ and $\mathbf{q} = \{p_0, \cdots, q_N\}$ give the limiting distributions[9] as $t \to \infty$ of the buffer content process defined in Section II. The limiting distributions $\mathbf{p}$, $\mathbf{q}$ are thus the solutions of

$$\mathbf{p} = (1-\theta_1)\mathbf{B}\mathbf{p} + \theta_2\tilde{\mathbf{A}}\mathbf{q} \tag{22a}$$

$$\mathbf{q} = \theta_1\mathbf{B}\mathbf{p} + (1-\theta_2)\tilde{\mathbf{A}}\mathbf{q} \tag{22b}$$

with, of course, the normalization

$$\mathbf{1}^T(\mathbf{p}+\mathbf{q}) = 1. \tag{23}$$

In this section we derive a simple formula for computing the vectors $\mathbf{p}$ and $\mathbf{q}$ for a given buffer size $(N+1)$ in terms of p and q for a buffer of size N. As a first step we simplify the problem by eliminating p from eqs. (22a), (22b). Multiplylng (22a) by θ_1 and (22b) by $(\theta_1 - 1)$ and adding gives

$$\mathbf{p} = \left(\frac{1-\theta_1}{\theta_1}\right)\mathbf{q} - \frac{1-\theta_1-\theta_2}{\theta_1}\tilde{\mathbf{A}}\mathbf{q}. \tag{24}$$

Substituting (24) into (22b) gives

$$[\mathbf{I} - (1-\theta_1)\mathbf{B} - (1-\theta_2)\tilde{\mathbf{A}} + (1-\theta_1-\theta_2)\mathbf{B}\tilde{\mathbf{A}}]\mathbf{q} = 0. \tag{25}$$

Premultiplying (24) by $\mathbf{1}^T$ and subtracting from (23) gives

$$\mathbf{1}^T\mathbf{q} = \frac{\theta_1}{\theta_1+\theta_2} \tag{26}$$

since $\mathbf{1}^T\tilde{\mathbf{A}} = \mathbf{1}^T$. It is important to note that the $N+1$ component equations in (25) are not independent. Indeed, since $\mathbf{1}^T\tilde{\mathbf{A}} = \mathbf{1}^T\mathbf{B} = \mathbf{1}^T$, it is clear that the first equation is just the sum of the rest and may therefore be ignored. The remaining N equations are linearly independent and we can solve them for $q_0, \cdots, q_{N-1}$ in terms of q_N, and then obtain q_N from (26). Finally, we can obtain p from (24).

In carrying out the solution of (25) and (26) in this manner the recursion we are looking for becomes obvious if we define the $(N + 1)$-dimensional vector $\mathbf{y}^{(N)}$ with components given by*

$$y_i^{(N)} = q_{N-i}^{(N)}/q_N^{(N)}, \qquad i = 0, \cdots, N. \tag{27}$$

[The meaning of the superscript (N) is given in Section 1.1.] Equations (25) and (27) give

$$y_0^{(N)} = 1 \tag{28a}$$

$$y_1^{(N)} = \frac{\theta_2}{1 - \theta_2} \tag{28b}$$

$$y_i^{(N)} = \frac{y_{i-1}^{(N)}}{1 - \theta_2}, \qquad i = 2, \cdots, k \tag{28c}$$

$$y_{k+1}^{(N)} = \frac{y_k^{(N)}}{1 - \theta_2} + \frac{1 - \theta_1 - \theta_2}{1 - \theta_2} y_1^{(N)} - \frac{\theta_2}{1 - \theta_2} \tag{28d}$$

$$y_i^{(N)} = \frac{y_{i-1}^{(N)}}{1 - \theta_2} + \frac{1 - \theta_1 - \theta_2}{1 - \theta_2} y_{i-k}^{(N)} - \frac{1 - \theta_1}{1 - \theta_2} y_{i-k-1}^{(N)}, \quad i > k + 1. \tag{28e}$$

The important fact about (28) is that the superscript (N) is superfluous. If N is changed to $N + 1$, for instance, in (28) we see that

$$y_i^{(N+1)} = y_i^{(N)}, \qquad i = 0, \cdots, N, \tag{29}$$

and the last component of $y^{(N+1)}$ is

$$y_{N+1}^{(N+1)} = \frac{1}{1 - \theta_2} y_N^{(N)} + \frac{1 - \theta_1 - \theta_2}{1 - \theta_2} y_{N-k+1}^{(N)} - \frac{1 - \theta_1}{1 - \theta_2} y_{N-k}^{(N)}. \tag{30}$$

Thus the vector $\mathbf{y}^{(N+1)}$ is obtained from $\mathbf{y}^{(N)}$ by merely appending to the components of $\mathbf{y}^{(N)}$ one component given by (30). To complete the recursion for $\mathbf{q}^{(N+1)}$, we note from (26) and (27) that

$$\frac{1}{q_{N+1}^{(N+1)}} = \left(\frac{\theta_1 + \theta_2}{\theta_1}\right) \sum_{i=0}^{N+1} y_i^{(N+1)} \tag{31}$$

* Note that $q_N^{(N)} \neq 0$, for otherwise the solution q of (25) is the null vector which cannot satisfy (26).

and therefore, from (29) and (30),

$$\begin{aligned}\frac{1}{q_{N+1}^{(N+1)}} &= \frac{1}{q_N^{(N)}} + \frac{\theta_1+\theta_2}{\theta_1}\,y_{N+1}^{(N+1)}\\ &= \frac{1}{q_N^{(N)}} + \frac{\theta_1+\theta_2}{\theta_1(1-\theta_2)}\\ &\qquad \times \left[y_N^{(N)} + (1-\theta_1-\theta_2)y_{N+1-k}^{(N)} - (1-\theta_1)y_{N-k}^{(N)}\right]. \qquad (32)\end{aligned}$$

Equation (32) gives $q_{N+1}^{(N+1)}$ in terms of the components of $\mathbf{q}^{(N)}$.

3.2 Probability of transmission fault and of buffer being full

Frequently it is adequate to determine the variation with buffer size of the components $p_N^{(N)}$ and $q_N^{(N)}$ rather than of the complete distributions $\mathbf{p}^{(N)}$ and $\mathbf{q}^{(N)}$. Notice that the probability of transmission fault $T^{(N)}$ is, by the definition given in Section I, identical to $q_N^{(N)}$; and the probability that a buffer of size N is full is clearly $p_N^{(N)} + q_N^{(N)}$. It is therefore of interest to obtain recursions for these quantities without having to compute the entire p and q vectors from the recursions derived in Section 3.1.

By premultiplying eqs. (22a) and (22b) by $\mathbf{e}_N^T (\triangleq \{0, 0, \cdots, 0, 1\})$ we get

$$\mathbf{e}_N^T\mathbf{p} = \theta_2\mathbf{e}_N^T\tilde{\mathbf{A}}\mathbf{q} = \frac{\theta_2}{1-\theta_2}\,\mathbf{e}_N^T\mathbf{q} \qquad (33)$$

or

$$\mathbf{e}_N^T(\mathbf{p}+\mathbf{q}) = \frac{\mathbf{e}_N^T\mathbf{q}}{1-\theta_2} = \frac{T^{(N)}}{1-\theta_2}, \qquad (34)$$

i.e.,

$$p_N^{(N)} + q_N^{(N)} = \frac{1}{1-\theta_2}\cdot T^{(N)}.$$

It therefore suffices to obtain a recursion for $T^{(N)}$. Suppressing the superscript (N) from (28e), and summing over the index i from $k+2$ to $N+k+1$, we get

$$\sum_{i=k+2}^{N+k+1} y_i = \frac{1}{1-\theta_2}\sum_{i=k+1}^{N+k} y_i + \frac{1-\theta_1-\theta_2}{1-\theta_2}\sum_{i=2}^{N+1} y_i - \frac{1-\theta_1}{1-\theta_2}\sum_{i=1}^{N} y_j. \qquad (35)$$

Since $T^{(N)} = q_N^{(N)}$, (31) is used to relate $T^{(N)}$ to $\{y_i\}$. Now substituting the values of $\{y_i\}$ given in (28) we obtain

$$\frac{1}{T^{(N+k+1)}} - \frac{1}{1-\theta_2}\frac{1}{T^{(N+k)}} - \frac{1-\theta_1-\theta_2}{1-\theta_2}\frac{1}{T^{(N+1)}} + \frac{1-\theta_1}{1-\theta_2}\frac{1}{T^{(N)}} = 0, \qquad N \geqq 1. \tag{36}$$

Equation (36) is the recursion quoted in Section I.

3.3 Mean first passage time

Let N be a positive integer and let the buffer be of size greater than N. Let an infinitely long burst start at $t = 0$, with the buffer initially empty, and let $F^{(N)}$ denote the mean time required for the buffer content to first exceed N. The manner in which $F^{(N)}$ depends on N is a useful guide in designing an adequate buffer, especially when the distribution of burst lengths is not accurately known. In this section we derive a recursive formula for $F^{(N)}$, the recursion being with respect to the level N.

By definition, the Nth component of the vector $\mathbf{s}(t)$ defined in eq. (10b) is the probability that the level N is exceeded for the first time at the instant $t + 1$. Therefore,

$$\begin{aligned} F^{(N)} &= \sum_{t=0}^{\infty} (t+1) s_N(t) \\ &= \mathbf{e}_N^T \sum_{t=0}^{\infty} (t+1)\mathbf{s}(t). \end{aligned} \tag{37}$$

In the appendix we show that if λ is an eigenvalue of the matrix defined in (16), then $|\lambda| < 1$. This proves the convergence of the series in (37).

We proceed by obtaining an expression for $\sum_{t=0}^{\infty} (t+1)\mathbf{s}(t)$ by the method of generating functions. Let

$$\mathbf{S}(z) \triangleq \sum_{t=0}^{\infty} z^{t+1}\mathbf{s}(t) \tag{38}$$

so that

$$\mathbf{S}'(z) = \sum_{t=0}^{\infty} (t+1) z^t \mathbf{s}(t) \tag{39}$$

and, in particular,

$$\mathbf{S}'(1) = \sum_{t=0}^{\infty} (t+1)\mathbf{s}(t). \tag{40}$$

From the equation, (21), governing the evolution of $\{s(t)\}$ we find that

$$\mathbf{S}(z) = [I - z\mathbf{C} - z^2\mathbf{D}]^{-1}\{z\mathbf{s}(0) + z^2\mathbf{s}(1) - z^2\mathbf{Cs}(0)\}. \qquad (41)$$

It is shown in the appendix that the above matrix inverse exists for all $|z| \leq 1$. Following the procedure already outlined [eqs. (39) and (40)] we find that

$$\sum_{t=0}^{\infty}(t+1)\mathbf{s}(t) = [\mathbf{I}-\mathbf{C}-\mathbf{D}]^{-1}[\mathbf{C}+2\mathbf{D}][\mathbf{I}-\mathbf{C}-\mathbf{D}]^{-1}\{\mathbf{s}(0)+\mathbf{s}(1)-\mathbf{Cs}(0)\}$$
$$+[\mathbf{I}-\mathbf{C}-\mathbf{D}]^{-1}\{\mathbf{s}(0)+2\mathbf{s}(1)-2\mathbf{Cs}(0)\}. \qquad (42)$$

The resulting expression for $F^{(N)}$, from (37) and (42), is further simplified by using the following identities:

$$\mathbf{e}_N^T = \frac{1}{\theta_1}\mathbf{1}^T[\mathbf{I} - \mathbf{C} - \mathbf{D}],$$

and

$$\mathbf{1}^T[\mathbf{C} + 2\mathbf{D}] = (\theta_1 + \theta_2)\mathbf{1}^T + (1 - \theta_2 - 2\theta_1)\mathbf{e}_N^T.$$

Then

$$F^{(N)} = \frac{\theta_1 + \theta_2}{\theta_1}\cdot\mathbf{1}^T[\mathbf{I} - \mathbf{C} - \mathbf{D}]^{-1}\{\mathbf{s}(0) - (1 - \theta_1)\mathbf{Bs}(0) + \theta_1\mathbf{Br}(0)\}$$
$$+ (\mathbf{1}^T\mathbf{r}(0) - \theta_2)/\theta_1. \qquad (43)$$

The above expression for $F^{(N)}$ holds for arbitrary initial states of the buffer. However, as mentioned in the beginning of this section, in deriving a recursive formula for $F^{(N)}$ we will assume the buffer empty at $t = 0$. In that case, $\mathbf{r}(0) = \tau\mathbf{e}_0$ and $\mathbf{s}(0) = (1 - \tau)\mathbf{e}_0$ with $\tau\epsilon[0, 1]$. Substituting in (43) we get, for this special case,

$$F^{(N)} = (\theta_1 + \theta_2)\mathbf{1}^T(\mathbf{I} - \mathbf{C} - \mathbf{D})^{-1}\mathbf{e}_0 + \frac{\tau - \theta_2}{\theta_1}. \qquad (44)$$

We can derive a recursion for the quantity

$$f^{(N)} \triangleq \mathbf{1}^T(\mathbf{I} - \mathbf{C} - \mathbf{D})^{-1}\mathbf{e}_0 \qquad (45)$$

from which the recursion for $F^{(N)}$ will follow immediately. The procedure is very similar to the one used to derive (36). Thus let $\mathbf{x}^T = (x_0, x_1, \cdots, x_N)$ be the solution of

$$(\mathbf{I} - \mathbf{C} - \mathbf{D})\mathbf{x} = \mathbf{e}_0. \qquad (46)$$

Then, since $\mathbf{1}^T(\mathbf{I} - \mathbf{C} - \mathbf{D}) = \theta_1\mathbf{e}_N^T$, we get $x_N = 1/\theta_1$. We may re-place the first of the component equations in (46) by this relation. Exactly as in (27) and (28), we find that the components $x_i^{(N)}$ $(i = 0, \cdots, N)$ of the vector $\mathbf{x}^{(N)}$ are, in reverse order, the first $N + 1$ numbers $\hat{x}_i$ in the sequence generated as follows:

$$\hat{x}_0 = \frac{1}{\theta_1} \tag{47a}$$

$$\hat{x}_i = \frac{1}{1 - \theta_2}\hat{x}_{i-1}, \qquad i = 1, \cdots, k, \tag{47b}$$

$$\hat{x}_i = \frac{1}{1 - \theta_2}\hat{x}_{i-1} + \frac{1 - \theta_1 - \theta_2}{1 - \theta_2}\hat{x}_{i-k} - \frac{1 - \theta_1}{1 - \theta_2}\hat{x}_{i-k-1}, \qquad i > k. \tag{47c}$$

Summing (47c) over i from $k + 1$ to $N + k + 1$ and noting that $f^{(N)} = \sum_{i=0}^{N} x_i$, we get

$$\begin{aligned} f^{(N+k+1)} - \frac{1}{1 - \theta_2} f^{(N+k)} - \frac{1 - \theta_1 - \theta_2}{1 - \theta_2} f^{(N+1)} - \frac{1 - \theta_1}{1 - \theta_2} f^{(N)} \\ = \sum_{i=0}^{k} \hat{x}_k - \frac{1}{1 - \theta_2}\sum_{i=0}^{k-1} \hat{x}_i - \frac{1 - \theta_1 - \theta_2}{1 - \theta_2}\hat{x}_0 \\ = \frac{1}{1 - \theta_2}, \end{aligned} \tag{48}$$

where the last step follows from (47a), (47b). However,

$$f^{(N)} = \{F^{(N)} - (\tau - \theta_2)/\theta_1\}/(\theta_1 + \theta_2).$$

Substituting in (48) we get

$$\begin{aligned} F^{(N+k+1)} = \frac{1}{1 - \theta_2} F^{(N+k)} + \frac{1 - \theta_1 - \theta_2}{1 - \theta_2} F^{(N+1)} - \frac{1 - \theta_1}{1 - \theta_2} \\ + \frac{\theta_1 + \theta_2}{1 - \theta_2}, \qquad N = 0, 1, 2, \cdots. \end{aligned} \tag{49}$$

Interestingly, τ does not appear explicitly in the recursion (49); it does, of course, affect the initial conditions [i. e., the values of $F^{(0)}, \cdots, F^{(k)}$] via eq. (44).

It is interesting to note that the forcing term $(\theta_1 + \theta_2)/(1 - \theta_2)$ in (49) can be eliminated. By direct substitution it is seen that if $\theta_1 \neq k\theta_2$ then $F^{(N)} - (\theta_1 + \theta_2)N/(\theta_1 - k\theta_2)$ satisfies the homogeneous recursion (49). When $\theta_1 = k\theta_2$, the same is true of $F^{(N)} - (\theta_1 + \theta_2)N^2/k(2 - \theta_1 - \theta_2)$. These transformations which reduce (49) to the homogeneous form will be of use when we investigate the asymptotics of solutions in Section V.

IV. BURSTS WITH GEOMETRICALLY DISTRIBUTED LENGTHS

When information is available concerning the distribution of burst lengths we can compute design parameters which are more realistic than the quantities $T^{(N)}$ and $F^{(N)}$ discussed in the preceding sections. Clearly an event is of consequence only if it occurs within a burst. Its probability of occurrence at the tth instant must therefore be weighted by the probability that the burst length exceeds t. If the distribution of burst lengths is the weighted sum of geometric distributions, i.e.,

$$\text{Prob. }\{\text{Burst length} = i\} = \sum_{k=0}^{J} \beta_k(1-\rho_k)\rho_k^{i-1},$$
$$i = 1, 2, \cdots; \quad 0 < \rho_k < 1, \tag{50}$$

then simple recursions can be obtained for such weighted averages. To keep the derivations simple we have only treated the case $J = 1$ since, as shown in Ref. 1, generalization to higher values of J is straightforward. In Sections 4.1 and 4.2 we derive such recursions for the probability of overflow within a burst and for the mean time to first cross a level within a burst.

4.1 Overflow within a burst

For a buffer of size greater than N let $G^{(N)}$ denote the probability that the buffer content exceeds N (at least once) during a burst. It is clear that $G^{(N)}$ also equals the probability that a transmission fault occurs (at least once) during a burst, when the buffer size is N. We call $G^{(N)}$ the probability of overflow.

By its definition in (10), $s_N(t)$ is the probability that the buffer content exceeds N for the first time at $t + 1$. Therefore,

$$\begin{aligned} G^{(N)} &\triangleq \sum_{t=0}^{\infty} s_N(t) \text{ Prob. }\{\text{burst length} \geqq (t+1)\} \\ &= \sum_{t=0}^{\infty} s_N(t)\rho^t \\ &= \mathbf{e}_N^T \sum_{t=0}^{\infty} \rho^t \mathbf{s}(t). \end{aligned} \tag{51}$$

As proved in the appendix, the matrix in (16) has all its eigenvalues strictly within the unit circle. Therefore the series in (51) converges for $\rho \leqq 1$.

Multiplying (21) by ρ^{t+2} and summing over t from 0 to ∞ we get, on re-arranging terms,

$$(\mathbf{I} - \rho\mathbf{C} - \rho^2\mathbf{D}) \sum_{t=0}^{\infty} \rho^t \mathbf{s}(t) = \mathbf{s}(0) + \rho\{\mathbf{s}(1) - \mathbf{C}\mathbf{s}(0)\}$$
$$= [\mathbf{I} - \rho(1 - \theta_1)\mathbf{B}]\mathbf{s}(0) + \rho\theta_1\mathbf{B}\mathbf{r}(0). \quad (52)$$

In the appendix we show that $(\mathbf{I} - \rho\mathbf{C} - \rho^2\mathbf{D})$ is nonsingular for all $\rho \leqq 1$. Therefore

$$G^{(N)} = \mathbf{e}_N^T(\mathbf{I} - \rho\mathbf{C} - \rho^2\mathbf{D})^{-1}[\{\mathbf{I} - \rho(1 - \theta_1)\mathbf{B}\}\mathbf{s}(0) + \rho\theta_1\mathbf{B}\mathbf{r}(0)]. \quad (53)$$

As before, specializing to the interesting case of an initially empty buffer, i.e., $\mathbf{r}(0) = \tau\mathbf{e}_0$, $\mathbf{s}(0) = (1 - \tau)\mathbf{e}_0$, with τ in $[0, 1]$, we get

$$G^{(N)} = [(1 - \tau)(1 - \rho) + \rho\theta_1]\mathbf{e}_N^T(\mathbf{I} - \rho\mathbf{C} - \rho^2\mathbf{D})^{-1}\mathbf{e}_0. \quad (54)$$

We can obtain a recursion for $G^{(N)}$ by a procedure almost identical to that used in obtaining the recursion for $T^{(N)}$. Note that if $\mathbf{z}^{(N)}$ is a vector such that

$$(\mathbf{I} - \rho\mathbf{C} - \rho^2\mathbf{D})\mathbf{z}^{(N)} = \mathbf{e}_0 \quad (55)$$

then the components of the vector $\mathbf{z}^{(N)}/z_N^{(N)}$ are, in reverse order, the first $N + 1$ numbers in the sequence $\hat{z}_i$, $i = 0, 1, 2, \cdots$, generated by the relations

$$\hat{z}_0 = 1 \quad (56a)$$

$$\hat{z}_i = \frac{1}{\rho(1 - \theta_2)} \hat{z}_{i-1} \qquad i = 1, \cdots, k, \quad (56b)$$

$$\hat{z}_i = \frac{1}{\rho(1 - \theta_2)} \hat{z}_{i-1} + \frac{\rho(1 - \theta_1 - \theta_2)}{1 - \theta_2} \hat{z}_{i-k} - \frac{1 - \theta_1}{1 - \theta_2} \hat{z}_{i-k-1},$$
$$i > k. \quad (56c)$$

The first component equation in (55) then gives

$$\frac{1}{z_N^{(N)}} = \sum_{i=0}^{k} \pi_i \hat{z}_{N-i}, \qquad N > k, \quad (57)$$

where $\pi_0, \cdots, \pi_k$ are the leading $(k + 1)$ entries in the first row of $(\mathbf{I} - \rho\mathbf{C} - \rho^2\mathbf{D})$. (The remaining components of this row are null.) For $N > 2k$, each term on the right-hand side of (57) satisfies the recursion (56c). Therefore $1/z_N^{(N)}$ satisfies the same recursion. From (54), since $G^{(N)}$ is proportional to $z_N^{(N)}$ we find that $1/G^{(N)}$ also satisfies the same recursion, i.e., for $N > 2k$,

$$\frac{1}{G^{(N)}} = \frac{1}{\rho(1-\theta_2)} \cdot \frac{1}{G^{(N-1)}} + \frac{\rho(1-\theta_1-\theta_2)}{1-\theta_2} \frac{1}{G^{(N-k)}} - \frac{1-\theta_1}{1-\theta_2} \frac{1}{G^{(N-k-1)}} \cdot \quad (58)$$

It can additionally be shown that the above recursion holds for $2k \geqq N > 1$, by direct substitution of the initial values of $G^{(N)}$.

4.2 Mean time for first passage within a burst

For a buffer of size greater than N, let t denote the time required for the buffer content to first exceed N within a burst. Let $H^{(N)}$ denote the expectation of t conditional to the hypothesis that the level N is indeed exceeded within the burst. (Equivalently, $H^{(N)}$ is the mean time taken by a buffer of size N to first overflow within a burst, given that an overflow does occur.) Clearly

$$\begin{aligned} H^{(N)} &= \sum_{t=0}^{\infty} (t+1) s_N(t) \cdot [\text{Prob. that burst length} \geqq t+1]/G^{(N)} \\ &= \sum_{t=0}^{\infty} (t+1) s_N(t) \rho^t / G^{(N)} \\ &= \mathbf{e}_N^T \sum_{t=0}^{\infty} (t+1) \rho^t \mathbf{s}(t)/G^{(N)}. \end{aligned} \quad (59)$$

A comparison of (51) and (59) shows that

$$H^{(N)} = \frac{d}{d\rho} (\rho G^{(N)}) \cdot \frac{1}{G^{(N)}} \cdot \quad (60)$$

Multiplying (53) or (54) by ρ and differentiating with respect to ρ we can get closed expressions for $H^{(N)}$ for arbitrary initial state and for the buffer initially empty. The resulting expressions are rather unwieldy.

We can also use (60) to get a recursion for $H^{(N)}$. Thus let

$$V^{(N)} \triangleq \frac{1}{\rho G^{(N)}} \cdot \quad (61)$$

Then

$$\begin{aligned} H^{(N)} &= \frac{d}{d\rho} \left(\frac{1}{V_N} \right) \cdot \frac{1}{G^{(N)}} \\ &= -\frac{\rho U^{(N)}}{V^{(N)}}, \end{aligned} \quad (62)$$

where $U^{(N)} \triangleq (d/d\rho)V^{(N)}$. Here $V^{(N)}$ satisfies the recursion (58), and $U^{(N)}$ satisfies a recursion obtained by differentiating the recursion for $V^{(N)}$. Thus

$$V^{(N)} = \frac{1}{\rho(1-\theta_2)} V^{(N-1)} + \frac{\rho(1-\theta_1-\theta_2)}{1-\theta_2} V^{(N-k)} - \frac{1-\theta_1}{1-\theta_2} V^{(N-k-1)}$$

and

$$U^{(N)} = \frac{1}{\rho(1-\theta_2)} U^{(N-1)} + \rho \frac{(1-\theta_1-\theta_2)}{1-\theta_2} U^{(N-k)} - \frac{1-\theta_1}{1-\theta_2} U^{(N-k-1)} - \frac{1}{\rho^2(1-\theta_2)} V^{(N)} + \frac{(1-\theta_1-\theta_2)}{1-\theta_2} V^{(N-k)}. \quad (63)$$

V. ASYMPTOTIC BEHAVIOR

In this section we discuss the behavior as $N \to \infty$ of sequences generated by the recursion

$$\varphi_N - \frac{1}{\mu(1-\theta_2)} \varphi_{N-1} - \frac{\mu(1-\theta_1-\theta_2)}{1-\theta_2} \varphi_{N-k} + \frac{1-\theta_1}{1-\theta_2} \varphi_{N-k-1} = \xi_N, \quad (64)$$

with $N = k+1,\ k+2,\ \cdots$ and $0 < \theta_1 < 1,\ 0 < \theta_2 < 1$, and $0 < \mu \leqq 1$ the parameter ranges.

Every recursion derived in this paper can be put into the canonical form (64) by simple manipulations; furthermore, all but the recursion (63), Section 4.2, correspond to the homogeneous form of (64), i.e., $\xi_N \equiv 0$. In formulas for infinitely long burst (Sections 3.1, 3.2, 3.3) the parameter $\mu = 1$; in formulas for geometrically distributed bursts (Sections 4.1, 4.2) $0 < \mu = \rho < 1$.

Due to the linear, time-independent nature of the recursions in (64), the behavior of the solutions is determined by the sequence $\{\xi_N\}$ and the roots, λ_i, of the characteristic polynomial:

$$C(\lambda, \mu) \triangleq \mu(1-\theta_2)\lambda^{k+1} - \lambda^k - \mu^2(1-\theta_1-\theta_2)\lambda + \mu(1-\theta_1). \quad (65)$$

For the special case $\mu = 1$ the relevant properties of the roots were derived in Ref. 2. Here we derive the properties for arbitrary μ in the range $0 < \mu \leqq 1$. These properties are summarized in the following

Lemma: For the range of parameters specified above, (a) $C(\lambda, \mu)$ has exactly two positive real zeros λ_1 and λ_2 which lie in the ranges $[\mu(1-\theta_1)]$

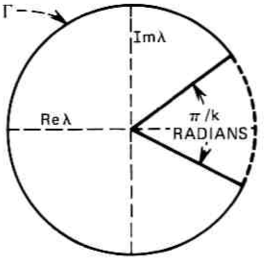


Fig. 2—Proof of lemma.

$< (\lambda_1)^k \leqq \mu$ and $1/\mu \leqq \lambda_2 < 1/\mu(1-\theta_2)$ (the equality signs are unnecessary unless $\mu = 1$); (b)[*] *the remaining zeros all satisfy $|\lambda_i|^k < \mu$.*

Proof: (a) Regardless of the sign of $(1-\theta_1-\theta_2)$ there are two sign reversals in the coefficients of $C(\lambda, \mu)$. By Descartes' rule, therefore, $C(\lambda, \mu)$ has at most two positive real zeros. On the other hand, successively setting $\lambda = 0$, $\lambda^k = \mu(1-\theta_1)$, $\lambda^k = \mu$, $\lambda = 1/\mu$, $\lambda = 1/\mu(1-\theta_2)$ we find that $C(\lambda, \mu)$ takes on the respective values $\mu(1-\theta_1)$, $\mu^2\theta_1\theta_2[\mu(1-\theta_1)]^{1/k}$, $-\mu\theta_1(1-\mu^{(k+1)/k})$, $-\theta_2(\mu^{-k}-\mu)$, and $\mu\theta_1\theta_2/(1-\theta_2)$. Also $C(\lambda, \mu) \to +\infty$ as $\lambda \to +\infty$. For $0 < \mu < 1$, therefore, there are exactly two zeros in the respective ranges asserted. For $\mu = 1$ further examination is required to decide whether one or both of these zeros become exactly equal to 1. Noticing that $C(1, 1) = 0$ and $(\partial/\partial\lambda) C(1, 1) = \theta_1 - k\theta_2$, it follows that when $\mu = 1$, either λ_1 or λ_2 or both become equal to 1 according as $\theta_1 - k\theta_2 < 0$, > 0, or $= 0$. (b) We will prove the stronger result that the remaining zeros lie strictly within the contour Γ (Fig. 2) defined by the following segments in the complex λ plane:

$$\lambda = Re^{j(\pi/k)} \qquad 0 \leqq R \leqq \mu^{1/k} \tag{66a}$$

$$= \mu^{1/k}e^{j\theta} \qquad \frac{\pi}{k} \leqq \theta \leqq 2\pi - \frac{\pi}{k} \tag{66b}$$

$$= Re^{-j(2\pi-\pi/k)} \qquad 0 \leqq R \leqq \mu^{1/k}. \tag{66c}$$

To prove this let us define

$$C_1 \triangleq \mu\lambda[(1-\theta_2)\lambda^k - \mu(1-\theta_1-\theta_2)] \tag{67a}$$

$$C_2 \triangleq \lambda^k - \mu(1-\theta_1) \tag{67b}$$

[*] We are tacitly assuming $k > 1$. For $k = 1$, $C(\lambda, \mu)$ becomes a quadratic with both roots positive and real in the ranges given in (a).

so that

$$
\begin{aligned}
C(\lambda, \mu) &= C_1 - C_2 \\
&= C_2 \left(\frac{C_1}{C_2} - 1 \right). \qquad (68)
\end{aligned}
$$

We will show that $\mathrm{Re}[C_1/C_2 - 1] < 0$ for all λ on the contour Γ. Then by an obvious modification of Rouche's theorem,[10] it follows that $C(\lambda, \mu)$ and C_2 each have the same number of zeros within Γ. As C_2 has $k - 1$ zeros within Γ, this proves the lemma.

To show that $\mathrm{Re}\,(C_1/C_2 - 1) < 0$ for all λ on Γ, let us consider separately the circular arc defined by (66b) and the radial lines defined by (66a) and (66c).

(*i*) On the circular arc (66b) straightforward manipulation gives

$$
\begin{aligned}
&|C_1|^2 - \mu^{2+2/k}|C_2|^2 \\
&\qquad = -2\theta_2\mu^2(2 - \theta_1 - \theta_2)(1 - \cos k\theta) \leqq 0. \qquad (69)
\end{aligned}
$$

For $\mu < 1$, therefore, $|C_1/C_2| < 1$, hence $\mathrm{Re}\,(C_1/C_2 - 1) < 0$. If $\mu = 1$, this argument remains valid except at points where $\cos k\theta = 1$, for then $|C_1/C_2| = 1$. However, if $\cos k\theta = 1$ and $\mu = 1$, we find that $C_1/C_2 - 1 = e^{j\theta} - 1$, whose real part <0 for $\pi/k \leqq \theta \leqq 2\pi - \pi/k$.

(*ii*) On the radial lines (66a) and (66c),

$$
\begin{aligned}
&\mathrm{Re}\left(\frac{C_1}{C_2} - 1 \right) \\
&\qquad = \mu R \left(1 - \theta_2 - \frac{\mu\theta_1\theta_2}{R^k + \mu(1 - \theta_1)} \right) \cos\frac{\pi}{k} - 1, \qquad (70)
\end{aligned}
$$

which is obviously <0 for $R^k \leqq \mu$.

All the recursions of this paper except (63) correspond to the homogeneous form of (64), i.e., $\xi_i \equiv 0$. Solutions of all such recursions are of the form

$$
\varphi_N = \sum_{i=0}^{k} \beta_i \lambda_i^N, \qquad (71)
$$

and therefore the asymptotic behavior is governed by λ_2, and λ_1 when it is equal to 1. (In the special case $\rho = 1$ and $\theta_1 = k\theta_2$, the dominant root is repeated and the usual modification must be made.) Dropping the subscript i from λ_i and β_i we give below an expression for the latter in terms of the initial conditions of the recursion, namely

$(\varphi_o, \cdots, \varphi_k)$:

$$\beta = a \cdot \left[\frac{-1}{\lambda^{k+1}} \cdot \frac{(1-\theta_1)}{(1-\theta_2)} \cdot \varphi_o + \left\{ 1 - \frac{1}{\lambda\rho(1-\theta_2)} \right\} \sum_{i=1}^{k-1} \frac{\varphi_i}{\lambda_i} + \frac{\varphi_k}{\lambda^k} \right], \quad (72a)$$

where

$$a = \rho(1-\theta_2)\lambda^{k+1}/[\lambda^k + k\rho^2(1-\theta_1-\theta_2) - (k+1)\rho(1-\theta_2)]. \quad (72b)$$

Thus, for example, the recursion for the probability of overflow [eq. (58)], with $\tau = 0$, in the canonical form (64) has the initial conditions $\varphi_0 = 1$, $\varphi_i = 1/[\rho(1-\theta_2)]^i$, $i = 1, 2, \cdots, k$. Also, in this case the dominant root of the characteristic polynonial λ_2 is the only root outside the unit circle in the complex plane. Therefore,

$$\frac{1}{G^{(N)}} \sim \beta\lambda_2^N, \quad (73)$$

where β is obtained from (72) for the appropriate values of $\varphi_o, \cdots, \varphi_k$. It can be easily shown that $\beta > 0$. In (73) (and similarly throughout this section) we use the notation $1/G^{(N)} \sim \beta\lambda_2^N$ to mean that $|1/G^{(N)} - \beta\lambda_2^N| < \epsilon^N$, for sufficiently large N, and $\epsilon < 1$.

In a manner similar to the derivation of (73) we can show that the probability of a transmission fault (Section 3.2) has the following asymptotic behavior

$$\frac{1}{T^{(N)}} \sim \alpha_1\lambda_2^N + \frac{\theta_1+\theta_2}{\theta_1 - k\theta_2} \quad \text{when } \theta_1 < k\theta_2 \quad (74a)$$

$$\sim \left[\frac{2\theta_2(N+1)}{2-\theta_1-\theta_2} + \frac{1-\theta_1-\theta_2}{1-\theta_1} \right] \frac{\theta_1+\theta_2}{\theta_1} \quad \text{when } \theta_1 = k\theta_2 \quad (74b)$$

$$\sim \frac{\theta_1+\theta_2}{\theta_1 - k\theta_2} \quad \text{when } \theta_1 > k\theta_2. \quad (74c)$$

In (74), α_1 is obtained from the generic formula (72a). We have shown that $\alpha_1 > 0$ and, of course, $1 < \lambda_2 < 1/1 - \theta_2$. Likewise, the mean first passage time (see Section 3.3) is, asymptotically,

$$F^{(N)} \sim \alpha_2\lambda_2^N - \frac{(\theta_1+\theta_2)N}{k\theta_2 - \theta_1}, \qquad \alpha_2 > 0, \qquad \text{when } \theta_1 < k\theta_2$$

$$\sim \frac{(\theta_1+\theta_2)N^2}{k(2-\theta_1-\theta_2)} + \alpha_3 N + \alpha_4, \qquad \text{when } \theta_1 = k\theta_2$$

$$\sim \frac{(\theta_1+\theta_2)N}{\theta_1 - k\theta_2} + \alpha_5, \qquad \text{when } \theta_1 > k\theta_2.$$

Finally,

$$H^{(N)} \sim \alpha_6 + \alpha_7 N, \qquad \alpha_7 > 0. \tag{76}$$

VI. COMPUTATIONS

We have written computer programs to recursively compute the quantities $T^{(N)}$, $F^{(N)}$, $G^{(N)}$, $H^{(N)}$ as functions of N for specified values of θ_1, θ_2, k, and ρ. Figures 3 through 6 are sample illustrations generated by these programs for $\theta_1 = 0.2$ and $\theta_2 = 0.1$. The asymptotic behavior

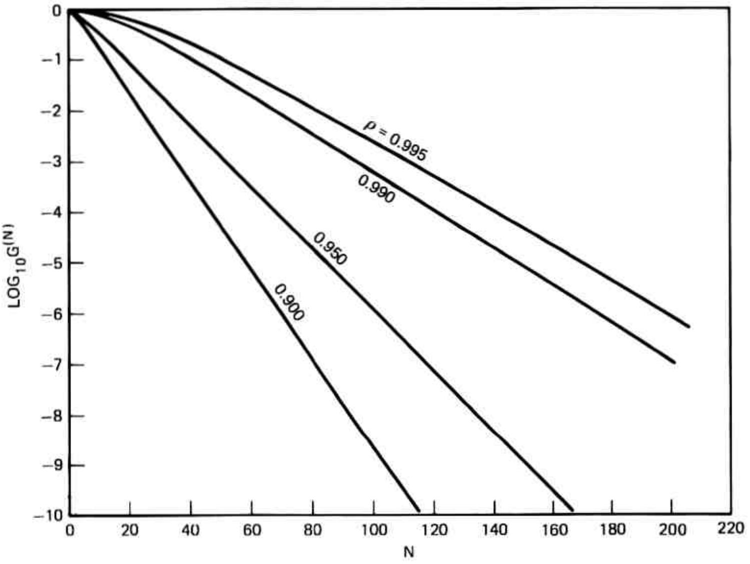


Fig. 3—Probability of overflow in a burst vs level ($\theta_1 = 0.2$, $\theta_2 = 0.1$, $k = 5$).

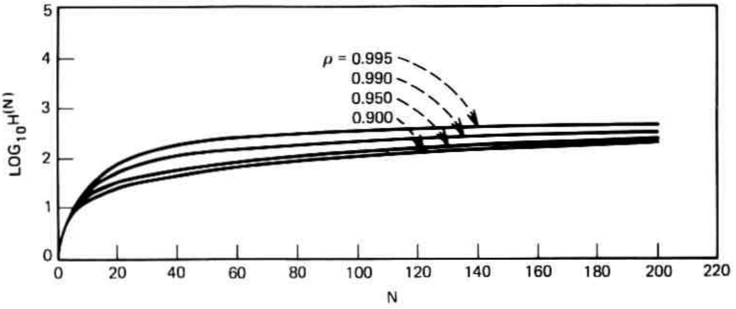


Fig. 4—Mean time for first passage conditional on overflow vs level ($\theta_1 = 0.2$, $\theta_2 = 0.1$, $k = 5$).

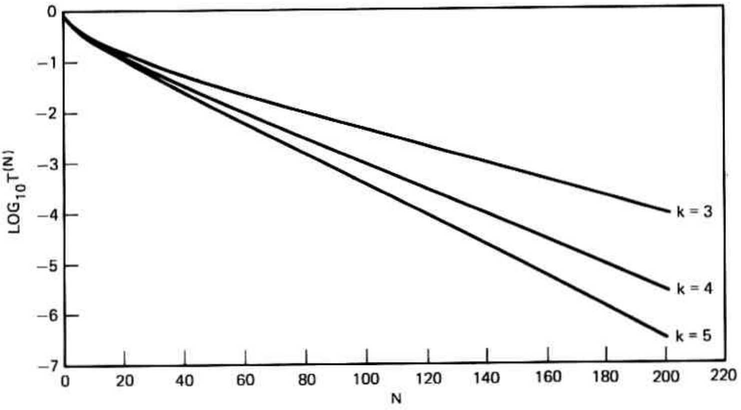


Fig. 5—Steady-state probability of transmission fault vs buffer size ($\theta_1 = 0.2$, $\theta_2 = 0.1$).

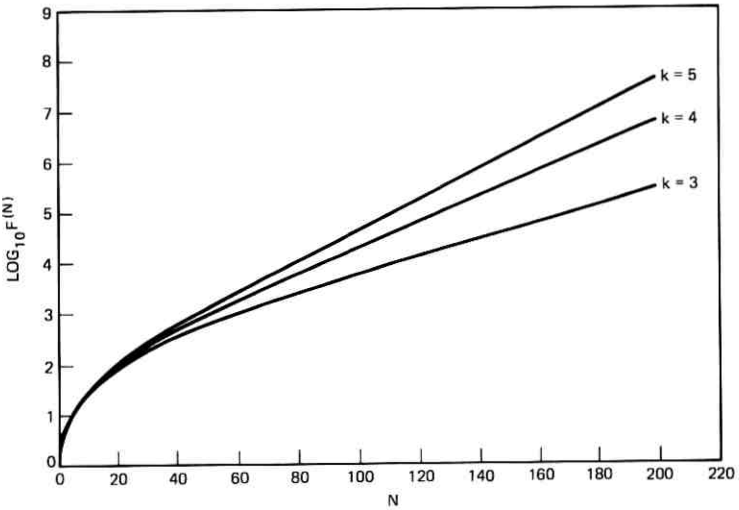


Fig. 6—Mean time for first passage in infinitely long bursts ($\theta_1 = 0.2$, $\theta_2 = 0.1$).

of the various quantities is seen to be in accord with that given by eqs. (73)–(76) of the previous section. The dependence on the parameters ρ and k also is intuitively reasonable.

APPENDIX

(a) We prove the assertion made in the text [immediately following eq. (37)] that the eigenvalues of the matrix (16) all lie strictly within the unit circle. Let

$$\mathbf{M} \triangleq \begin{bmatrix} (1-\theta_1)\mathbf{B} - \lambda\mathbf{I} & \theta_2\mathbf{A} \\ \theta_1\mathbf{B} & (1-\theta_2)\mathbf{A} - \lambda\mathbf{I} \end{bmatrix}, \tag{77}$$

where **I** is the identity matrix of order $N + 1$. Then we must show that

$$\det \mathbf{M} \neq 0, \qquad \text{for } |\lambda| \geqq 1. \tag{78}$$

From the defining equations (14) and (18), we notice that the last column of A is identically zero. Thus

$$\det \mathbf{M} = -\lambda \det \mathbf{M}', \tag{79}$$

where $\mathbf{M}'$ is obtained from **M** by deleting its last row and column. Let m_{ij}, $i, j = 0, \cdots, 2N + 1$, denote the elements of $\mathbf{M}'$. Then a theorem of Hadamard[11] states that $\det \mathbf{M}' \neq 0$ provided $\mathbf{M}'$ is irreducible and

$$|m_{jj}| \geqq P_j = \sum_{i=0,\, i\neq j}^{2N+1} |m_{ij}|, \tag{80}$$

for all j, with strict inequality for at least one j. The irreducibility condition as stated in Ref. 11 is satisfied. To show (80) we note that

$$|m_{jj}| = |\lambda| \quad \text{and } P_j = 1, \quad \text{for } j = 1, \cdots, 2N + 1, \tag{81}$$

and

$$|m_{00}| = |1 - \theta_1 - \lambda|, \qquad P_0 = \theta_1. \tag{82}$$

Thus except at $\lambda = 1$, we find that (80) is true with strict inequality for $j = 0$. This proves the assertion (78) except for the point $\lambda = 1$. However, for $\lambda = 1$, $\det \mathbf{M} = \theta_1(1 - \theta_2)^N$ which, by assumption, $\neq 0$.

(b) Following eq. (52) we made the assertion that $\mathbf{I} - \rho\mathbf{C} - \rho^2\mathbf{D}$ is a nonsingular matrix for $\rho \leqq 1$. The proof is as follows. Let

$$\mathbf{M} \triangleq \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}, \tag{83}$$

with $\mathbf{M}_{11} = (1 - \theta_1)\mathbf{B} - \lambda\mathbf{I}$, $\mathbf{M}_{21} = \theta_1\mathbf{B}$, etc. As $\mathbf{M}_{21}$ commutes with $\mathbf{M}_{11}$, an identity of Schur[12] states that

$$\det \mathbf{M} = \det [\mathbf{M}_{11}\mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{12}]. \tag{84}$$

However, straightforward manipulation of the right side of (84) shows that

$$\det \mathbf{M} = \det (\lambda^2\mathbf{I} - \lambda\mathbf{C} - \mathbf{D}). \tag{85}$$

Then the assertion follows from (78).

REFERENCES

1. B. Gopinath, Debasis Mitra, and M. M. Sondhi, "Formulas on Queues in Burst Processes—I," B.S.T.J., *52*, No. 1 (January 1973), pp. 9–33.
2. D. Mitra and B. Gopinath, "Buffering of Data Interrupted by Source with Priority," Proc. Fourth Asilomar Conf. on Circuits and Systems, 1970.
3. M. R. Schroeder and S. L. Hanauer, "Interpolation of Data with Continuous Speech Signals," B.S.T.J., *46*, No. 8 (October 1967), pp. 1931–1933.
4. D. N. Sherman, "Data Buffer Occupancy Statistics for Asynchronous Multiplexing of Data in Speech," Proc. Intl. Conf. on Communications, I.E.E.E., 1970, San Francisco.
5. P. T. Brady, "A Technique for Investigating On-Off Patterns of Speech," B.S.T.J., *44*, No. 1 (January 1965), pp. 1–22.
6. P. T. Brady, "A Model for Generating On-Off Speech Patterns in Two-Way Conversations," B.S.T.J., *48*, No. 7 (September 1969), pp. 2445–2472.
7. J. O. Limb, "Buffering of Data Generated by the Coding of Moving Images," B.S.T.J., *51*, No. 1 (January 1972), pp. 239–259.
8. B. Haskell, private communication.
9. S. Karlin, *A First Course in Stochastic Processes*, New York: Academic Press, 1966.
10. M. Marden, "Geometry of Polynomials," Mathematical Surveys, *3*, American Mathematical Society, Providence, Rhode Island, 1966, pp. 2–3.
11. Marden, pp. 140–141.
12. F. R. Gantmacher, *The Theory of Matrices*, vol. 1, New York: Chelsea Publishing Co., 1960, p. 46.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 3, March 1974
Printed in U.S.A.

Adaptive Coding for Coherent Detection of Digital Phase Modulation

By C. L. RUTHROFF and W. F. BODTMANN

(Manuscript received October 3, 1972)

Although coherent phase-shift keying (CPSK) is an efficient means of transmitting digital signals over carrier systems, it has not enjoyed widespread use at microwave and millimeter wavelengths because of the difficulty of recovering an accurate reference carrier for coherent detection.

In this paper, a system is described which requires only a narrow-band phase-locked-oscillator filter for reference carrier recovery. This is accomplished by block-coding and decoding the pulse sequence at the terminals; the recovery of a baseband timing wave is also facilitated by the coding process. It is also shown that: (i) for an arbitrary random input sequence, accurate carrier recovery cannot be achieved with just a narrow-band filter, (ii) for the system described, any input pulse sequence is acceptable, and (iii) there is a maximum error in the phase of the recovered reference carrier which can be controlled by choosing the number of pulses in the coding block and the bandwidth of the recovery filter.

I. INTRODUCTION

Coherent phase-shift keying (CPSK) is one of the most efficient means of modulation for the transmission of digital information over carrier systems. In particular, CPSK is at least as efficient as frequency-shift keying or differentially coherent phase-shift keying.[1-3] Equally important from the point of view of hardware realization, CPSK is suited to operation with amplifiers which operate most efficiently in a nonlinear regime; this class of amplifiers includes those using traveling-wave tubes, varactor up-converters, and tunnel or IMPATT diodes used as power amplifiers or as injection-locked oscillator amplifiers.

Traveling-wave-tube amplifiers have been proposed for use in satellite repeaters, and there is considerable current work directed toward the application of millimeter-wave integrated circuit injection-locked oscillator amplifiers in digital radio and waveguide transmission systems.[4–16] The CPSK method described here is suitable for those applications.

For the type of operation envisaged for these systems, the statistics of the digital sources are usually unknown. To achieve maximum operational flexibility, it was assumed at the outset that the system must operate with any input pulse sequence. With this arrangement, the statistics of the signal source need not be restricted.

In the system to be described, the recovery of the reference carrier phase—a major problem in CPSK transmission—is accomplished with the aid of a block-coding and decoding of the pulse sequence at the terminals. This coding allows recovery of the reference carrier with a narrow-band phase-locked-oscillator filter.

Another important problem in digital systems with unrestricted pulse sequences is the recovery of the timing wave for use in the regeneration process. The same coding process which affords reference carrier recovery for all sequences also assures timing wave recovery for all sequences.

In this paper, the block-coding, the reference carrier recovery, and the timing wave recovery are described for binary and multilevel CPSK systems.

II. COHERENT DIGITAL PHASE MODULATION

2.1 The baseband and modulated carrier signal formats

A diagram of the block-coded CPSK carrier system is shown in Fig. 1. Ignoring the block coder for the moment, the input to the radio system is a baseband sequence of discrete amplitudes as illustrated by a binary sequence of ones and zeros in Fig. 2a. The ones are coded as positive pulses and the zeros as negative pulses as shown in Fig. 2b; this sequence is the input to the phase modulator. Although raised-cosine pulses are used for illustration, other pulse shapes can also be used.

An m-level baseband sequence of raised-cosine pulses shown in Figure 2b is written

$$v(t) = V_o \sum_n a_n p(t - nT), \tag{1}$$

where T is the pulse interval, V_o the peak pulse amplitude, $a_n = \pm 1$, and

$$p(t) = \begin{cases} \dfrac{1}{2}\left(1 + \cos \dfrac{2\pi t}{T}\right), & |t| \leqq \dfrac{T}{2} \\[2ex] 0, & |t| > \dfrac{T}{2}. \end{cases}$$

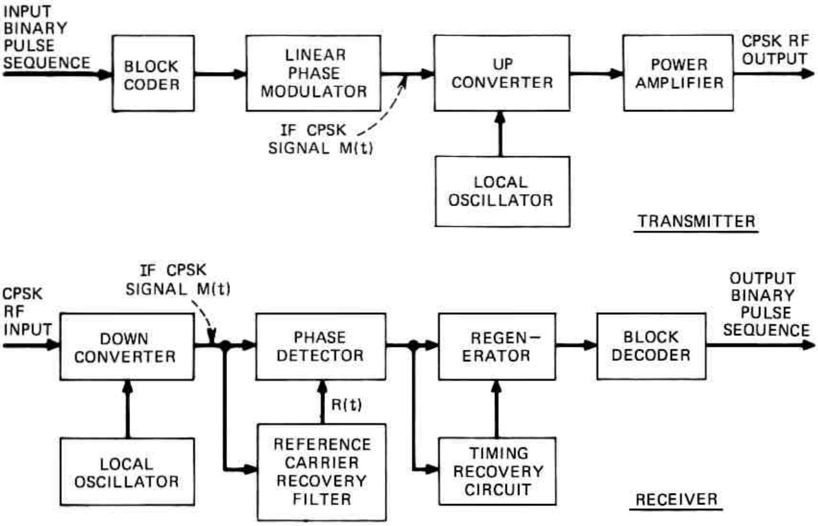


Fig. 1—Block-coded CPSK carrier terminals.

This baseband signal is used to phase modulate a sinusoidal carrier. The output of the phase modulator is

$$M(t) = A_c \cos \left[\omega_c t + \sum_n a_n p(t - nT)\right], \tag{2}$$

where $a_n = k\pi/m$, $k = \pm 1, \pm 3, \cdots, \pm(m-1)$. The pulse sequence of Fig. 2b represents both the baseband signal voltage of (1) and the phase modulation in (2). The peak baseband voltage V_o produces a peak phase deviation of $\pi/2$ radians for the binary case illustrated.

A vector representation of the modulated signal is shown in Fig. 3a. The carrier amplitude is A_c and the unmodulated phase of the carrier

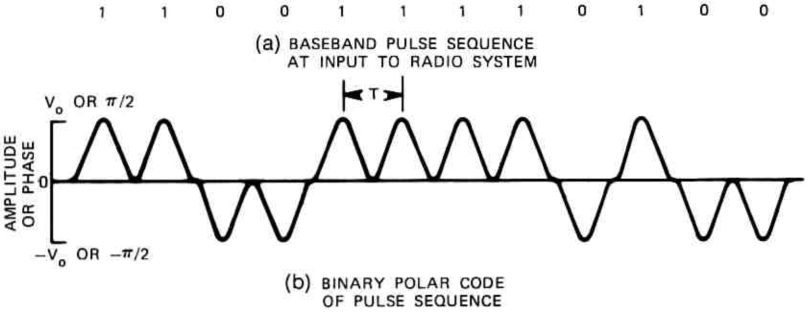

1 1 0 0 1 1 1 1 0 1 0 0
(a) BASEBAND PULSE SEQUENCE AT INPUT TO RADIO SYSTEM
T
V₀ OR π/2
0
−V₀ OR −π/2
AMPLITUDE OR PHASE
(b) BINARY POLAR CODE OF PULSE SEQUENCE

Fig. 2—Binary polar signal format.

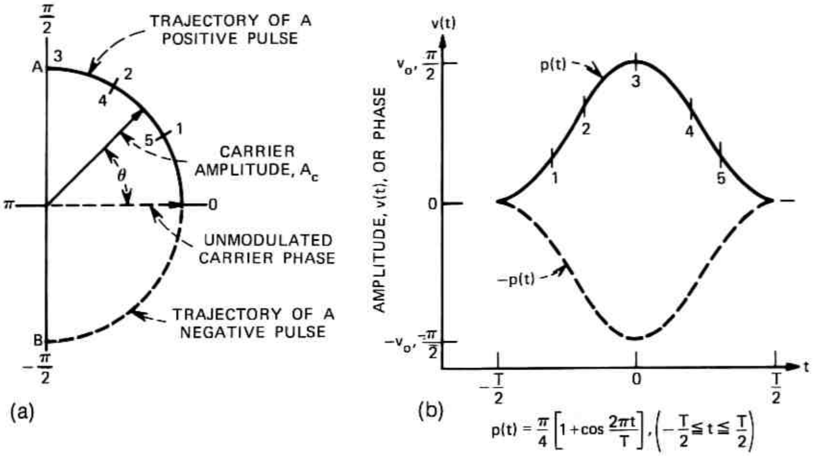

$\frac{\pi}{2}$
TRAJECTORY OF A POSITIVE PULSE
A
3
2
4
1
5
CARRIER AMPLITUDE, A_c
θ
π
0
UNMODULATED CARRIER PHASE
TRAJECTORY OF A NEGATIVE PULSE
B
$-\frac{\pi}{2}$
(a)
v(t)
$v_o, \frac{\pi}{2}$
AMPLITUDE, v(t), OR PHASE
p(t)
3
2
4
1
5
0
−p(t)
$-v_o, \frac{-\pi}{2}$
$-\frac{T}{2}$
0
$\frac{T}{2}$
t
(b)
$p(t) = \frac{\pi}{4}\left[1 + \cos\frac{2\pi t}{T}\right], \left(-\frac{T}{2} \leq t \leq \frac{T}{2}\right)$

Fig. 3—Phase plane representation of a binary signal.

is zero. When pulses modulate the phase of the carrier the amplitude remains constant and the phase follows the modulating signal voltage. Trajectories for the positive and negative raised-cosine pulses of Fig. 3b are shown in Fig. 3a for the binary case.

It is worth noting that double-sideband suppressed-carrier modulators and switched delay-line modulators are sometimes regarded as phase modulators. A justification for this interpretation is that the pulses are sampled at the receiver only when the phase is at the peak value. However, they differ from phase modulators in the amount of amplitude modulation that is generated. The trajectory of the double-sideband suppressed-carrier modulation is the vertical axis in Fig. 3a between the points A and B; the trajectory of the switched delay-line modulation may be intermediate between the vertical trajectory and the circular trajectory of a phase modulator. Since future systems are expected to have power amplifiers operating in the region of saturation, the distortion caused by large variations in amplitude can be avoided by restricting consideration to phase modulators. Phase modulators suitable for this purpose are described elsewhere.[15]

2.2 A description of coherent phase detection

Let the input to the phase detector of the receiver in Fig. 1 be the phase-modulated signal $M(t)$. The phase detector requires a local reference signal with the proper phase. This reference signal is written

$$R(t) = -2A_R \sin[\omega_c t + \epsilon(t)], \tag{3}$$

where $\epsilon(t)$ is any error in the reference phase. The output of the phase detector is the low-frequency part, $V_R(t)$, of the product of the input signal and the reference signal.

$$V_R(t) = A_C A_R \sin\left[\sum_n a_n p(t - nT) - \epsilon(t)\right]. \tag{4}$$

If the phase error, $\epsilon(t)$, is zero, and if the output of the phase detector is sampled at times $t = nT$, the output will be $\pm A_C A_R$ accordingly as $a_n = \pm 1$ and the transmitted pulse sequence is recovered.

If the reference phase error is not zero, the signal output amplitude will be reduced by the factor $\cos \epsilon$. For example, if $\epsilon = \pi/4$, the baseband pulse amplitude will be reduced 3 dB. An important function of the system to be described is to recover the reference phase in such a manner that $\epsilon(t)$ is small.

III. REFERENCE CARRIER RECOVERY WITH A PHASE-LOCKED OSCILLATOR

The reference carrier recovery filter is assumed to be a phase-locked oscillator with a locking bandwidth much smaller than the bandwidth of the modulating pulse sequence. The analysis presented here applies to an injection-locked oscillator or a first-order phase-locked loop; the noiseless case will be considered.*

Let the input signal be

$$M(t) = A_c \cos\left[\omega_c t + \theta(t)\right]. \tag{5}$$

The differential equation describing the locking behavior of a negative resistance sine-wave oscillator has been derived in several forms.[17–19] With the present notation the equation is

$$\frac{d\epsilon(t)}{dt} = (\omega_o - \omega_c) - \Delta \sin\left[\epsilon(t) - \theta(t)\right], \tag{6}$$

where ω_o is the unlocked oscillator frequency, $|\omega_o - \omega_c| \ll \omega_c$, 2Δ is the locking bandwidth, and $\epsilon(t)$ is the reference phase error.

Since the oscillator is being used to recover the reference phase, the phase error, $\epsilon(t)$, should be as small as possible. For this reason it is necessary that the locking bandwidth of the locked oscillator be much smaller than the bandwidth of the signal, $\theta(t)$. This assumption, in its most useful form, means that $\Delta T \ll 1$. Following Adler, expression (6) is rearranged as follows:

$$\frac{d\epsilon(t)}{dt} = \Delta K - \Delta \sin\left[\epsilon(t) - \theta(t)\right], \tag{7}$$

* Eisenberg has presented a related analysis which includes additive thermal noise.[16]

where $K = (\omega_o - \omega_c)/\Delta$. The term K represents any initial difference between the free-running frequency of the oscillator and the input frequency; in the region of interest $|K| < 1$.

We are interested in deriving an unambiguous reference carrier for multilevel digital modulation in which each pulse is time-limited to a single interval of duration T.

$$\theta(t) = \sum_n a_n p(t - nT). \tag{8}$$

During a single pulse the variation in $\theta(t)$ is much larger than the variation in $\epsilon(t)$ because $\Delta T \ll 1$. Therefore, the phase error at the end of the nth pulse can be found by integrating (7) over the nth pulse with $\epsilon(t)$ held constant at the value of the phase error at the beginning of the nth pulse. Writing $\epsilon_n = \epsilon(nT)$, we have, from (7) and (8),

$$\begin{aligned}\epsilon_{n+1} - \epsilon_n &= \Delta \int_{nT}^{(n+1)T} \{K - \sin[\epsilon_n - \theta(t)]\} dt \\ &= \Delta T \left[K - \frac{\sin \epsilon_n}{T} \int_{nT}^{(n+1)T} \cos \theta(t) dt \right. \\ &\qquad \left. + \frac{\cos \epsilon_n}{T} \int_{nT}^{(n+1)T} \sin \theta(t) dt \right]. \end{aligned} \tag{9}$$

Each pulse is nonzero in a single interval of duration T so we have

$$\int_{nT}^{(n+1)T} \cos \theta(t) dt = \int_{nT}^{(n+1)T} \cos a_n p(t - nT) dt = \int_0^T \cos |a_n| p(x) dx,$$

and

$$\begin{aligned}\int_{nT}^{(n+1)T} \sin \theta(t) dt &= \int_{nT}^{(n+1)T} \sin a_n p(t - nT) dt \\ &= \frac{a_n}{|a_n|} \int_0^T \sin |a_n| p(x) dx.\end{aligned}$$

Simplifying the notation, we write

$$b_n \equiv a_n/|a_n|, \qquad C_n \equiv \frac{1}{T} \int_0^T \cos |a_n| p(x) dx,$$

and

$$S_n \equiv \frac{1}{T} \int_0^T \sin |a_n| p(x) dx. \tag{10}$$

Expression (9) becomes

$$\epsilon_{n+1} - \epsilon_n = \Delta T [K - C_n \sin \epsilon_n + b_n S_n \cos \epsilon_n]. \tag{11}$$

As shown in (10), C_n and S_n are functions of the shape of the pulse. For the digital signal described by (8) the peak deviation is less than π radians for any number of levels and, for the class of pulse shapes of interest, S_n is positive. This is not true of C_n—it can be positive, negative, or zero. Eisenberg[16] has derived C and S for several pulse shapes of interest. The sign of C_n has an important effect upon the phase of the recovered reference carrier; in order that the recovered reference carrier have an unambiguous phase near zero degrees, it will be shown that a pulse shape must be used for which $C_n > 0$.

For the binary case, (11) can be written

$$\frac{\epsilon_{n+1} - \epsilon_n}{\Delta T} = K - \sqrt{C^2 + S^2} \sin \left(\epsilon_n - b_n \tan^{-1} \frac{S}{C} \right). \tag{12}$$

Let the probability that $b_n = +1$ be p and the probability that $b_n = -1$ be $(1 - p)$. The average phase, ϵ_o, will be such that the error due to p positive pulses is equal in amplitude and opposite in sign from the error due to $(1 - p)$ negative pulses. From (12) we get

$$p \left[K - \sqrt{C^2 + S^2} \sin \left(\epsilon_o - \tan^{-1} \frac{S}{C} \right) \right] = -(1-p)\left[K - \sqrt{C^2 + S^2} \sin \left(\epsilon_o + \tan^{-1} \frac{S}{C} \right) \right],$$

and solving for the average phase, we get

$$\epsilon_o = \sin^{-1} \frac{K}{\sqrt{C^2 + (2p-1)^2 S^2}} + \tan^{-1} (2p - 1) \frac{S}{C}. \tag{13}$$

Under the best circuit adjustment, $K = 0$ and the average phase error is given by the second term in (13). When C is positive, the average phase is in the first or fourth quadrant and when $p = \frac{1}{2}$ the average phase is zero. On the other hand, when C is negative, the average phase is in the second or third quadrant and when $p = \frac{1}{2}$ the average phase is π.

A switch of the reference carrier phase from near zero to near π can happen in a multilevel system. Consider a 4-level system with rectangular pulses and peak phase deviations of $\pm\pi/4$ and $\pm 3\pi/4$. Suppose that for a time the pulses alternate between $\pm\pi/4$. From (10), $C = \cos \pi/4 = 1/\sqrt{2} > 0$ and the average phase is zero. If the pulse sequence then changes to alternate between $\pm 3\pi/4$, $C = \cos 3\pi/4 = -1/\sqrt{2} < 0$ and the average phase becomes π. This very undesirable situation can be avoided by using pulse shapes for which $C > 0$. In the rest of this paper we assume that, in all cases, a pulse shape is chosen for which $C_n > 0$.

It is highly desirable that the average phase of the reference carrier be near zero. This means that in addition to requiring that $C_n > 0$, it is also necessary that $K \approx 0$ and $p \approx \frac{1}{2}$ as may be seen from (13). The parameter K can be kept near zero by setting the rest frequency of the phase-locked oscillator equal to the signal carrier frequency. The system is required to operate with any input sequence so it is unlikely that p will always be near one-half. The input sequence can be coded—by a block coder to be described in Section V—into a transmitted sequence with $p = \frac{1}{2}$, thus insuring $\tan^{-1} (2p - 1)S/C \approx 0$. Even with coding, however, the reference phase will fluctuate about zero and it is necesssary to insure that these fluctuations do not cause substantial degradation in performance relative to the performance which would be obtained with a perfect reference carrier. It has often been thought that the problem in the recovery of reference phase is that the occurrence of long sequences of identical pulses drives the recovered phase beyond reasonable limits and that if the sequences of pulses were sufficiently random this problem would go away. Random sequences are therefore of great interest. In the next section the variance of the reference phase error is derived and the results illustrated by an example.

IV. REFERENCE CARRIER PHASE ERROR FOR RANDOM SEQUENCES

The differential equation which describes the phase-locked oscillator is nonlinear and therefore difficult to solve. We begin by noting that (7) can be solved exactly on a pulse-by-pulse basis if the pulses are rectangular. For pulses with other shapes an equivalent rectangular pulse can be derived. Then, linearizing the equation, and recognizing that the phase error is approximately normally distributed, the variance can be estimated for the equivalent rectangular pulses. The binary case is considered.

Let the pulses be rectangular with peak deviation $\pm\theta_n$. Then, rearranging (7) and setting $\theta(t) = \theta_n$, we have

$$\int_{nT}^{(n+1)T} \frac{d\epsilon(t)}{-K + \sin\left[\epsilon(t) - \theta_n\right]} = \int_{nT}^{(n+1)T} - \Delta dt.$$

The solution to the integral on the left can be found in many tables of integrals. After some algebra the result can be written

$$\epsilon_{n+1} = \theta_n + 2 \tan^{-1} \times \left[\frac{R \tan \frac{\epsilon_n - \theta_n}{2} + \left(K - \tan \frac{\epsilon_n - \theta_n}{2}\right) \tanh\left(\frac{\Delta T}{2} R\right)}{R + \left(1 - K \tan \frac{\epsilon_n - \theta_n}{2}\right) \tanh\left(\frac{\Delta T}{2} R\right)} \right], \tag{14}$$

where $\epsilon_n = \epsilon[(n+1)T]$ and $R = \sqrt{1-K^2}$. Equation (14) is exact for rectangular pulses and, if an input sequence of rectangular pulses is specified, the exact phase error can be computed. We will estimate the variance for a linearized version of (14). When $K = 0$, (14) can be written

$$\epsilon_{n+1} = \theta_n + 2\tan^{-1}\left(e^{-\Delta T}\tan\frac{\epsilon_n - \theta_n}{2}\right). \tag{15}$$

Approximating the tangent by its argument,

$$\epsilon_{n+1} \approx \theta_n(1 - e^{-\Delta T}) + \epsilon_n e^{-\Delta T}. \tag{16}$$

Applying (16) repeatedly we get

$$\epsilon_{n+k} \approx \epsilon_n e^{-k\Delta T} + (1 - e^{-\Delta T}) \sum_{m=0}^{k-1} \theta_{n+m} e^{-(k-1-m)\Delta T}.$$

When k is large, the phase error is independent of n and becomes

$$\epsilon_k \approx (1 - e^{-\Delta T}) \sum_{m=0}^{k-1} \theta_m e^{-m\Delta T}, \tag{17}$$

where the pulses have been rearranged to simplify the notation.

From (12) it may be seen that the error due to a shaped pulse when $\epsilon_n \approx 0$ is given by

$$\epsilon \approx \Delta T\sqrt{C^2 + S^2}\sin\left(\tan^{-1}\frac{S}{C}\right), \tag{18}$$

where a pulse of positive polarity is assumed. Comparison of (18) with the error in (17) due to the pulse θ_o suggests that the equivalent rectangular pulse is obtained by letting

$$\theta_m = b_m\sqrt{C^2 + S^2}\sin\left(\tan^{-1}\frac{S}{C}\right) \tag{19}$$

in (17) where $b_m = +1$ with probability p, $b_m = -1$ with probability $(1-p)$.

The variance of (17) can be found by straightforward means;[20] the mean, from (13), and the variance for shaped pulses are:

$$\begin{aligned} \mu &\approx \tan^{-1}(2p-1)\frac{S}{C} \\ \sigma^2 &\approx 2\left[\sqrt{C^2+S^2}\sin\left(\tan^{-1}\frac{S}{C}\right)\right]^2 \Delta T p(1-p). \end{aligned} \tag{20}$$

The reference error is approximately normally distributed and the probability that the reference phase error will exceed a specified value

ϵ_s is[21]

$$P(|\epsilon| \geqq \epsilon_s) \approx Q\left(\frac{\epsilon_s - m}{\sigma}\right) + Q\left(\frac{\epsilon_s + m}{\sigma}\right). \qquad (21)$$

These results will be illustrated by an example. Let the pulse rate be 100 megabits per second and the locking bandwidth 0.5 MHz. For raised-cosine pulses with a peak deviation $\pm\pi/2$,

$$\mu \approx \tan^{-1}(2p - 1)$$
$$\sigma^2 \approx 2\Delta T S^2 p(1 - p)$$

with $S = 0.6021947$. For $p = \frac{1}{2}$, $\mu = 0$ and $\sigma \approx 0.0213$ radian rms. The fraction of time that the phase error exceeds 0.1 radian is

$$P(|\epsilon| \geqq 0.1) \approx 2Q(4.697) = 2.75 \times 10^{-6}.$$

In some applications, changes in pulse pattern density may occur which are reflected in fluctuations in p. In this event the probability of a pulse being positive will not be constant at $p = \frac{1}{2}$ but will wander slowly about this value. Suppose, in the foregoing example, p increases by five percent from $p = \frac{1}{2}$ to $p = 0.525$. Then, $\mu = 0.05$ and

$$P(|\epsilon| \geqq 0.1) \approx Q(0.235) + Q(7.04) \approx 0.41.$$

Two important conclusions are illustrated by this example.

(*i*) For practical circuit parameters, the probability of exceeding reasonable phase errors is uncomfortably large even for well-behaved random input sequences.

(*ii*) Small variations in pulse pattern density can cause large increases in the probability of exceeding reasonable phase errors.

It should be understood that the above example is but one of many possible examples; however, the filter bandwidth assumed is a practical value for systems operating at millimeter wavelengths. It should also be noted that the pulse sequences assumed in the example are highly idealized and may or may not approximate sequences from real sources.

V. THE BINARY BLOCK CODER

The coder described in this section is a digital adaptation of a coder invented by F. K. Bowers.[22]

The operation of the block coder will be described with the aid of Fig. 4. The block counter is an up/down counter which counts each

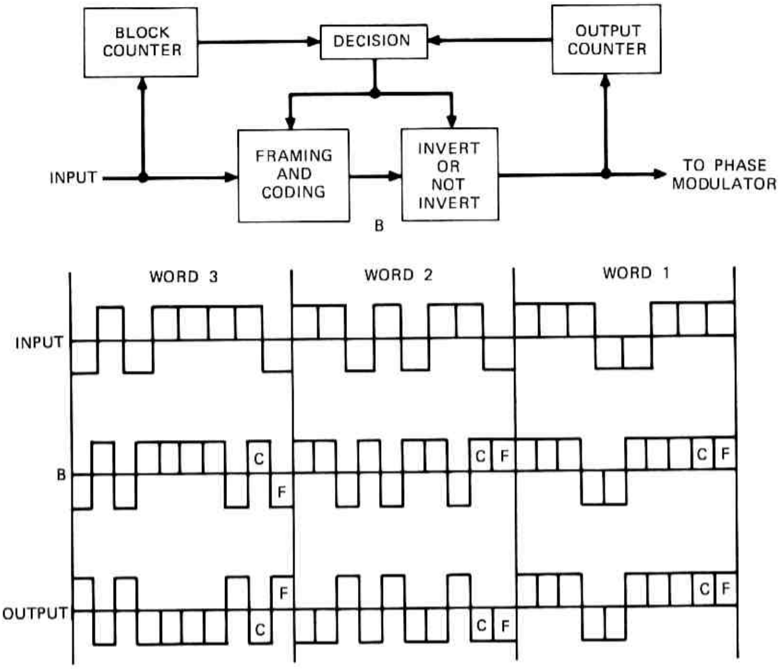


Fig. 4—Binary coder.

successive block of M pulses in the input sequence and indicates on its output terminal whether that block contains more positive than negative pulses. M is an even integer. The output counter is also an up/down counter which counts all pulses transmitted and indicates whether a surplus of positive or negative pulses has been transmitted since the start of transmission. The outputs of the two counters are used in the decision circuit to invert or not invert the block of M pulses just counted, the decision always being made to equalize the number of positive and negative pulses transmitted.

In addition to the framing pulses, a coding pulse is added to each block of M pulses and is used in the receiver to re-invert those blocks which were inverted at the transmitter. Figure 4 shows the operations of a block coder on a sequence of binary input pulses. In this configuration a framing pulse and a coding pulse have been added to each block of M input pulses. While a coding pulse is necessary for each block of M input pulses, fewer framing pulses can be used if desired.

The decoding process at the receiving terminal is illustrated in Fig. 5. The position of the coding pulses in the input sequence are known

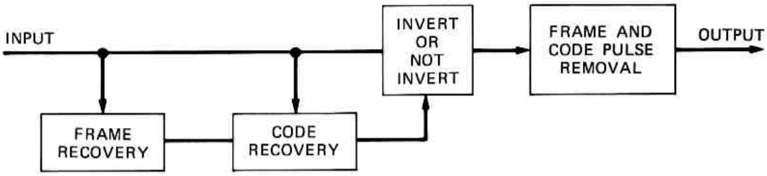


Fig. 5—Binary decoder.

relative to the position of the framing pulse. When framing is established, the coding pulses can be detected and the proper block inversions made so that the output sequence will be identical to the input sequence at the transmitting terminal.

A detailed analysis of the coding operation reveals the following results for M even.

(*i*) The output count, and hence the sum of the output sequence, cannot exceed $\pm(1 + 3M/2)$.

(*ii*) At the end of each frame of $M + 1$ output pulses the output count cannot exceed $\pm(M + 1)$; whatever the count at the end of a frame, the count at the end of the next frame will have moved in the direction of zero by a count of at least one.

(*iii*) The maximum number of pulses between a zero in the output counter and the next zero is $(M + 1)(M + 2)$.

(*iv*) The maximum number of identical pulses is $2 + 5M/2$ and the output count at the end of such a sequence is $\pm(1 + 3M/2)$.

In deriving these properties it is necessary to adopt a convention as to the output indicated by the output counter when the count is zero. If the count approached zero from the negative side it will indicate that a surplus of negative pulses has been sent and the converse is true if the zero count is approached from the positive side. Suppose the output counter indicates that a surplus of positive pulses has been transmitted. The coding pulse is counted as a positive pulse at the input counter making an odd number of pulses counted. At the end of the frame the input counter indicates that a surplus of positive or negative pulses is contained in the block. The block is inverted or not so that the output count goes toward zero. Since there is always a surplus of at least one pulse in each block, the output counter counts toward zero at least one count at the end of every frame; the output count can pass through zero in this process. Now suppose that the output count is zero and that this count was approached from the negative side. The output counter indicates that a surplus of negative pulses has been transmitted. If the next block has all positive pulses, the block will not be inverted and the output count will go to $(M + 1)$. This is the maximum count which can occur at the end of a frame since it has already been shown that at the end of the next frame the count must go toward zero by at least one. Property (*ii*) has therefore been demonstrated.

The example can be continued to demonstrate property (*i*). Let the count be $(M + 1)$ at the end of a frame. At the end of the next frame the count cannot exceed M so that the maximum number of positive pulses that can be added to the count during the frame is $M/2$. Thus, the maximum count is $M + 1 + M/2 = 1 + 3M/2$ and this is property (*i*).

The maximum number of pulses between zeros of the output count is found by achieving the maximum count of $(M + 1)$ in the first block and reducing the count by the minimum of one in successive blocks until zero is reached. There are just $(M + 2)$ blocks necessary to reach the next zero and $(M + 1)$ pulses per block so the maximum number of pulses between zeros is $(M + 1)(M + 2)$. This is property (*iii*). Finally property (*iv*) is achieved by letting the count at the end of a frame be $-(M + 1)$. The next frame has all pulses positive which brings the output counter to zero from the negative direction. The next $(1 + 3M/2)$ pulses can be positive bringing the total number of successive positive pulses to $(M + 1) + (1 + 3M/2) = 2 + 5M/2$ as stated.

When the coder is in operation the transmitted sequence contains equal numbers of positive and negative pulses. The resulting average phase is given by (13) with $p = \frac{1}{2}$.

$$\epsilon_o = \sin^{-1} \frac{K}{C}.$$

The fluctuations about ϵ_o can be determined from (17). The number of pulses between zeros is $(M + 1)(M + 2)$ and if ΔT is sufficiently small that $(M + 1)(M + 2)\Delta T \ll 1$, the exponential terms in (17) are approximately unity. Then, since the maximum sum of the output sequence is $1 + 3M/2$, an upper bound on the phase error results.

$$|\epsilon_{\max}| \leqq \sin^{-1} \frac{|K|}{C} + \Delta T \left(1 + \frac{3M}{2}\right) \theta_m, \tag{22}$$

where θ_m is the equivalent rectangular pulse as given in (19).

A graphic example of the effect of coding is given in Fig. 6. A sequence of 200 random pulses from the Rand table of random numbers

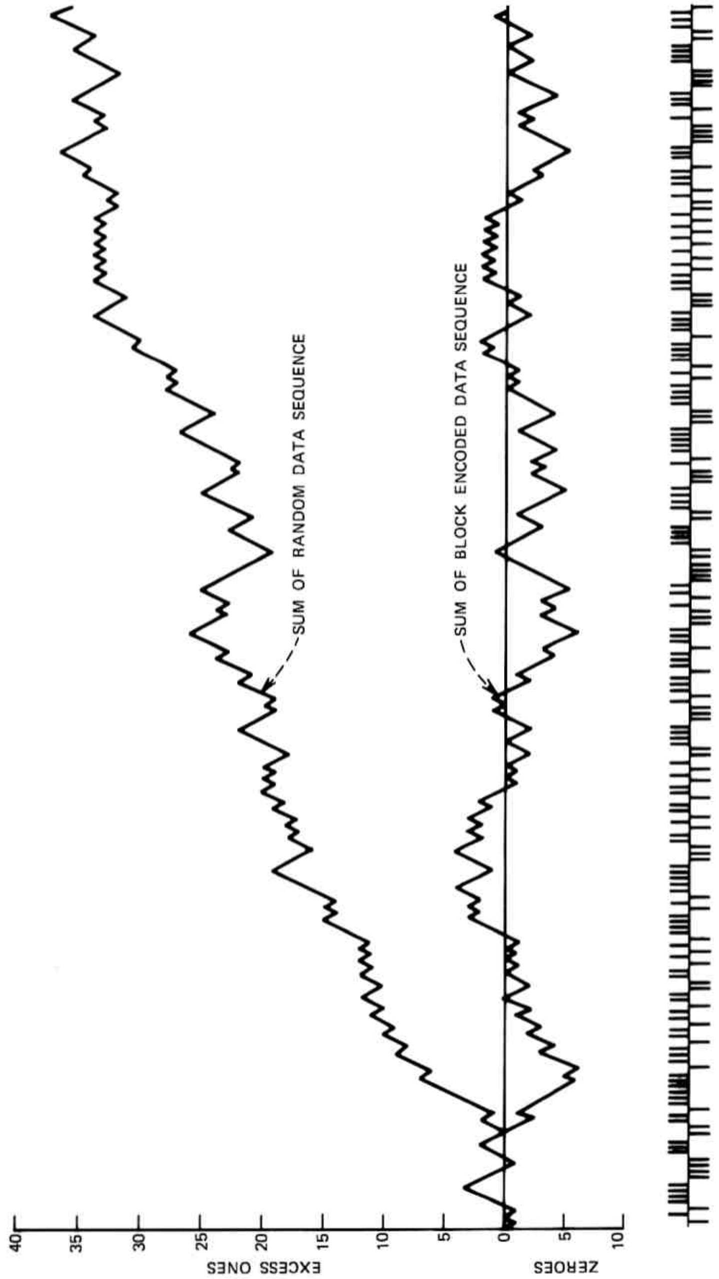


Fig. 6—The effect of coding a random sequence.

is shown at the bottom of the figure.[23] The sum of these digits, $\sum_n b_n$, is the upper plot. Note that although there are many transitions between positive and negative pulses the sum remains above zero most of the time. The slow drift of this sum illustrates the manner in which the phase error wanders.

The same input pulse sequence is shown after coding in a block coder with $M = 8$. The framing and coding pulses are not present. In the uncoded sequence the maximum error for $\Delta T = 0.01$ is 0.36 radian (20.6 degrees) for rectangular pulses with $\pi/2$ radian deviation, whereas the maximum phase error in the coded sequence is 0.06 radian (3.4 degrees).

The original sequence in Fig. 6 is not a rare case. As shown, $\sum_n b_n$ reaches 36 and the probability of this is

$$\begin{aligned} P_{200}(|\sum_n b_n| \geqq 36) &\approx 2Q\left(\frac{36}{\sqrt{200}}\right) \\ &\approx 2Q(2.54) = 0.011. \end{aligned}$$

Thus, about one out of a hundred sequences of 200 pulses each has a sum at least as great as the one shown in Fig. 6.

The price paid for the recovery of the reference carrier with a small phase error is an increase in the transmission rate by the factor $(M + 1)/M$.

VI. TIMING WAVE RECOVERY

It has been shown that, by coding the transmitted pulse sequence, the reference carrier can be recovered accurately. As shown in Fig. 1, the reference carrier is used to drive the phase detector in which the baseband pulse sequence is recovered. In a self-timed system, such as the one depicted in Fig. 1, it is necessary to recover a timing wave for use in regenerating the pulse sequence. Block-coding also helps in this process.

Bennett has shown that a timing wave can be recovered by suitable nonlinear operations even if a spectral line at the timing frequency does not exist; the method requires a suitable number of transitions between signal polarities.[24] But the block coding discussed in Section V insures that the largest number of pulses between signal transitions is $2 + 5M/2$. Therefore, the block coding insures the recovery of both the reference carrier and the timing wave for any sequence of pulses whatever.

For the sequence of pulses shown in Fig. 6 it is instructive to note that, although the sum $\sum_n b_n$ fails to cross the axis for a string of 186 pulses, there are frequent transitions between signal states and for each transition the timing recovery filter will receive a timing pulse.[24] There are 98 transitions in all and the maximum number of identical pulses between transitions is six. This is typical of the behavior of random sequences and is the reason that the recovery of the reference phase is usually more difficult than the recovery of the timing wave.

VII. MULTILEVEL BLOCK-CODED CPSK

The binary coding scheme described in Section V can be extended to 4, 8, 16, and higher numbers of levels. In each case the coder operates to equalize the numbers of pulses with equal amplitudes and conjugate phases. For example, the 4-level coder illustrated in Fig. 7 equalizes the numbers of pulses with $\pi/4$ radian peak deviation and opposite signs, and equalizes the numbers of pulses with $3\pi/4$ radian peak deviation and opposite signs. The 4-level decoder is shown schematically in Fig. 8.

Because the multilevel coder equalizes the numbers of positive and negative pulses for each pair of level's the computation of bounds on the phase error reduces to the binary case. A bound can be computed for each pair of levels and the largest bound applies; this will usually be the bound computed for the pair of levels with the largest deviation.

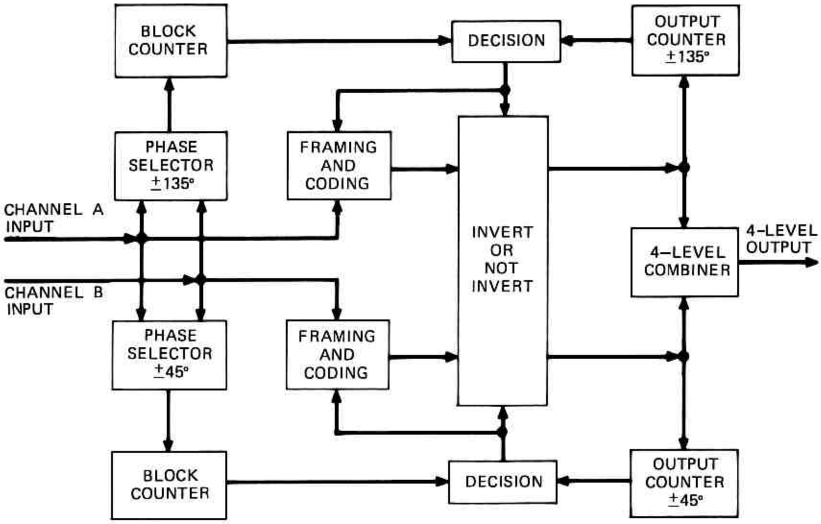


Fig. 7—Four-phase block encoder.

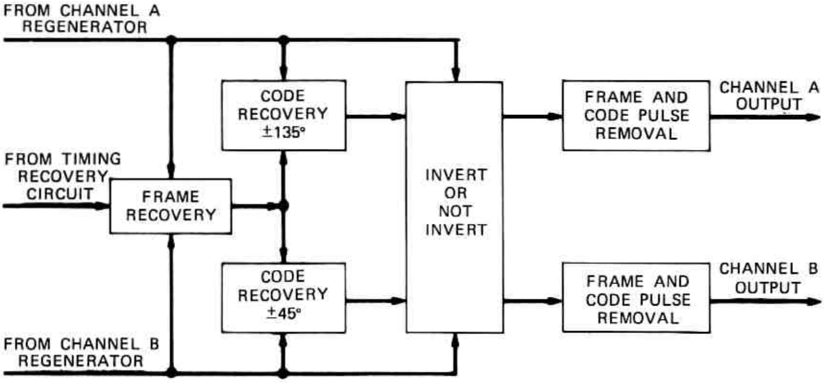

FROM CHANNEL A REGENERATOR
CODE RECOVERY ±135°
INVERT OR NOT INVERT
FRAME AND CODE PULSE REMOVAL
CHANNEL A OUTPUT
FROM TIMING RECOVERY CIRCUIT
FRAME RECOVERY
CODE RECOVERY ±45°
FRAME AND CODE PULSE REMOVAL
CHANNEL B OUTPUT
FROM CHANNEL B REGENERATOR

Fig. 8—Four-phase block decoder.

VIII. CONCLUSION

The CPSK system which results from block-coding the input digital sequence as described in this paper has the following properties:

(*i*) The system places no restrictions on the pulse sequence accepted from the source; any sequence whatever can be transmitted.

(*ii*) Recovery of the reference carrier at repeater points is accomplished with a narrow-band filter.

(*iii*) A timing wave can be recovered for any sequence of pulses.

(*iv*) Any pulse shaping required can be done at baseband.

(*v*) The phase-modulated carrier is suited to operation with nonlinear amplifiers; in some applications RF filters are not required to shape the spectrum.

The costs of providing these features are:

(*i*) A block coder must be supplied at the transmitting terminal and a decoder at the receiving terminal.

(*ii*) The transmission rate is increased by the factor $(M + 1)/M$ where M is the number of pulses in the coding block. In principle, M can be very large; in practice, it will be limited by the frequency stabilities of the RF oscillators used in the system.

(*iii*) The error rate is increased by the factor $2(1 - P_e)$ because an error in a coding pulse causes M errors in the signal sequence. This increase in error rate is of little practical importance.

IX. ACKNOWLEDGMENTS

Some of the ideas in this paper grew out of discussions with B. S. Glance; his papers and those of M. V. Schneider, W. W. Snell, and R. F. Trambarulo form the foundation for this work.

We are also indebted to L. J. Greenstein and V. K. Prabhu for valuable critical comments; V. K. Prabhu derived the increase in error rate.

REFERENCES

1. C. R. Cahn, "Performance of Digital Phase-Modulation Communication Systems," IRE Trans. Commun. Syst., *CS-7*, May 1959, pp. 3–6.
2. W. R. Bennett and J. R. Davey, *Data Transmission*, New York: McGraw-Hill, Inc., 1965, pp. 225–229.
3. V. K. Prabhu, "On the Performance of Digital Modulation Systems That Expand Bandwidth," B.S.T.J., *49*, No. 6 (July-August 1970), pp. 1033–1057.
4. M. V. Schneider, B. S. Glance, and W. F. Bodtmann, "Microwave and Millimeter Wave Hybrid Integrated Circuits for Radio Systems," B.S.T.J., *48*, No. 6 (July-August 1969), pp. 1703–1726.
5. B. S. Glance, "Power Spectra of Multilevel Digital Phase-Modulated Signals," B.S.T.J., *50*, No. 9 (November 1971), pp. 2857–2878.
6. B. S. Glance, "Microstrip Impatt Oscillator with High Locking Figure of Merit," Proc. IEEE, *57*, No. 11 (November 1969), pp. 2052–2053.
7. C. L. Ruthroff, "Injection-Locked-Oscillator FM Receiver Analysis," B.S.T.J., *47*, No. 8 (October 1968), pp. 1653–1661.
8. T. L. Osborne and C. H. Elmendorf, "Injection-Locked Avalanche Diode Oscillator FM Receiver," Proc. IEEE, *57*, No. 2 (February 1969), pp. 214–215.
9. B. S. Glance, "Digital Phase-Demodulator," B.S.T.J., *50*, No. 3 (March 1971), pp. 933–949.
10. M. V. Schneider, "Microstrip Lines for Microwave Integrated Circuits," B.S.T.J., *48*, No. 5 (May-June 1969), pp. 1421–1444.
11. M. V. Schneider, "Dielectric Loss in Integrated Microwave Circuits," B.S.T.J., *48*, No. 7 (September 1969), pp. 2325–2332.
12. M. V. Schneider, "A Scaled Hybrid Integrated Multiplier from 10 to 30 GHz," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1933–1942.
13. W. W. Snell, Jr., "Low-Loss Microstrip Filters Developed by Frequency Scaling," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1919–1931.
14. R. F. Trambarulo, "A 30-GHz Inverted-Microstrip Circulator," IEEE Trans. Microwave Theory and Techniques, *MTT-19*, No. 7 (July 1971), pp. 662–664.
15. C. L. Ruthroff and W. F. Bodtmann, "A Linear Phase Modulator for Large Baseband Bandwidths," B.S.T.J., *49*, No. 8 (October 1970), pp. 1893–1903.
16. M. Eisenberg, "Almost-Coherent Detection of Phase-Shift-Keyed Signals Using an Injection-Locked Oscillator," unpublished work.
17. R. Adler, "A Study of Locking Phenomena in Oscillators," Proc. IRE, *34*, No. 6 (June 1946), pp. 351–357.
18. R. V. Khoklov, "A Method of Analysis in the Theory of Sinusoidal Self-Oscillations," IRE Trans. Circuit Theory, *CT-7*, No. 4 (December 1960), pp. 398–413.
19. A. J. Viterbi, *Principles of Coherent Communication*, New York: McGraw-Hill, Inc., 1966, pp. 10–75.
20. A. Papoulis, "Narrow-Band Systems and Gaussianity," IEEE Trans. Inform. Theory, *IT-18*, No. 1 (January 1972), pp. 20–27.
21. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, N.B.S., Washington, D. C.: U.S. Government Printing Office, 1964, pp. 966–977.
22. F. K. Bowers, U.S. Patent No. 2,957,947, issued October 25, 1960.
23. The Rand Corporation, *A Million Random Digits With 100,000 Normal Deviates*, New York: The Free Press, 1966, pp. 66, lines 27 through 30.
24. W. R. Bennett, "Statistics of Regenerative Digital Transmission," B.S.T.J., *37*, No. 6 (November 1958), pp. 1501–1542.

Copyright © 1974 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 3, March 1974
Printed in U.S.A.

Controlled Current Filaments in PNIPN Structures With Application to Magnetic Field Detection

By G. PERSKY and D. J. BARTELINK

(Manuscript received August 20, 1973)

Stable biasing of multiterminal PNIPN structures to support controlled current filaments is proposed. A filament forms when base layer spreading resistance is sufficiently high for lateral base voltage drops to shut off injection at all but a small interior portion of the structure. For elongated parallel stripe emitter-base configurations, application of a magnetic field normal to the current filament and stripe axes results in lateral displacement of the filament which is detectable through a change in the external circuit current flow pattern. This displacement can be significantly larger than that of a single-pass Hall deflection, yielding high sensitivity. Analysis of an ideal model confirms a substantial improvement in performance over that of conventional Hall devices, viz., a manyfold increase in the ratio of short circuit signal current to drive current, similar improvement in signal-to-offset ratio, and controllable high output impedance making large signal voltages available. Solutions for the ideal model are presented for carrier transport in the I region both without and with lateral diffusive spread. It is argued that departures of actual device behavior from this model are not apt to be important. Possible circuit connections and a sample calculation of parameter values for a realizable structure are also given.

I. INTRODUCTION

The purpose of this paper is to show how PNPN structures can be biased stably to support controlled current filaments and to describe a sensitive magnetic field detector utilizing this principle in a PNIPN structure. PNPN devices are widely used as 2-terminal bistable switches[1] and as 3- and 4-terminal controlled switches,[2] and have also been utilized in 4-terminal operation as a linear amplifier.[3] The multiterminal circuit operation of the PNIPN structure described here forces nearly equal base and emitter currents and thereby suppresses these switching and amplifying effects. Stable filament formation properties are introduced when significant spreading resistance is incorporated in each base layer. For the operating conditions considered, the central junction remains in reverse bias and supports counterflowing confined streams of both electrons and holes. The shape and position of this filament are controlled by fully characterized device and circuit parameters, in contrast with previously reported filamentary instabilities.[4]

Magnetic field sensing is made possible because a magnetic field applied perpendicular to the filament displaces it laterally and thereby produces a signal in the external circuit. The displacement can be many times larger than the Hall displacement of either carrier species for a single transit of the I region. The I region is incorporated in the structure for the purpose of increasing filament length and hence its interaction with the magnetic field. The analysis will show that the sensitivity of the device can markedly exceed that of an ideal Hall effect detector of similar dimensions. Improved sensitivity is permitted because the compensating electron and hole streams prevent the buildup of a net Hall voltage. For moderate magnetic fields, detection is linear, yielding field polarity as well as magnitude. This behavior differs strongly from that of previously reported filamentary magnetic sensors in which detection is related to precipitous disruption of the filament when the field reaches a sufficient magnitude.[5]

Section II explains how stable multiterminal operation of the PNIPN structure can be achieved and how base resistance leads to the formation of a controlled current filament. An intuitive picture of the magnetic response is then developed. Sections III and IV present an analytical treatment of the filament characteristics and the magnetic response, respectively. Two cases are considered, transport in the intrinsic region without lateral spread and with diffusive spread. Section V assesses various effects that may cause actual device behavior to depart from the ideal operation predicted in Sections III and IV, shows possible circuit connections for the device, and presents theoretical performance characteristics for a realizable structure. Section VI summarizes the main features of the analysis. Preliminary experimental results are presented elsewhere.[6]

II. GENERAL CONSIDERATIONS

Figure 1a shows an elementary circuit which causes the emitter currents to equal the base currents in an idealized one-dimensional

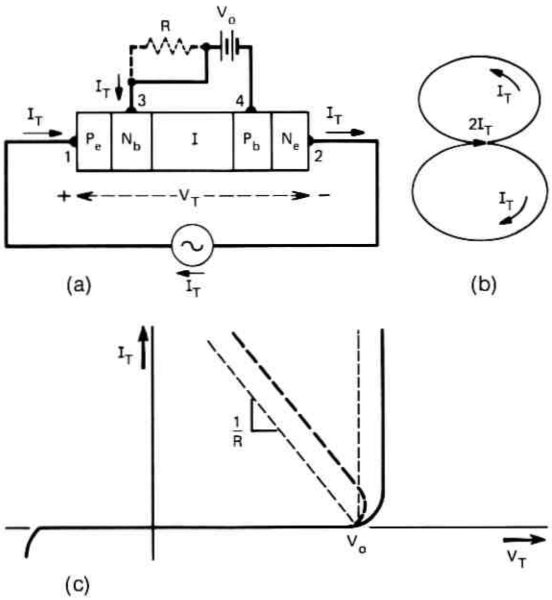


Fig. 1—(a) Four-terminal connection of PNIPN structure. (b) Current loop with figure-8 configuration. (c) Terminal 1–2 $I_T - V_T$ characteristics.

symmetric PNIPN structure with infinite current gain in each emitter-base configuration. We assume in addition that there is no significant recombination in the I region and that electrons and holes have identical properties apart from the charge sign. The current, I_T, supplied by the constant current source in Fig. 1a, follows a figure-8 path as shown in Fig. 1b. Upon entering emitter P_e, the current is injected as hole current through base N_b and region I. It arrives on base P_b where, as a stream of majority carriers, it can exit only through contact 4 to battery V_o. Simultaneously, electrons are injected by emitter N_e to arrive at N_b where, as majority carriers, their only path is to close the loop through contact 3. It is the direct external connection through battery V_o that permits stable conduction of the current $2I_T$ in the I region. Interruption of this external current would force the central junction to become forward biased, corresponding to the "on" state of the switching mode. With battery V_o in place, a typical terminal characteristic between contacts 1 and 2 is shown in Fig. 1c. It is single-valued and consists of the characteristic of a battery V_o and two diodes, all connected in series. Clearly, any finite impedance source connected between these terminals will give dc stable operation. The fact that stable 4-terminal operation of PNPN devices is possible has recently been demonstrated.[3]

With the addition of a resistance R in series with battery V_o, the voltage across the whole structure is reduced by $I_T R$, causing the characteristic eventually to bend back into a negative resistance region as indicated by the dashed curve in Fig. 1c. Stable operation will then require a source impedance greater than R. Note that this type of voltage turnback is consistent with common-base current gain α maintained at or near unity for each emitter-base configuration, throughout the negative resistance portion of the characteristic except near zero voltage. We have operated a commercial 4-terminal PNPN device, as well as an Ebers equivalent pair of transistors, in this circuit and have observed a stable negative resistance as depicted in Fig. 1c.

Formation of a stable current filament is brought about by base spreading resistance in the otherwise ideal PNIPN structure. The filament formation mechanism can be understood qualitatively with reference to the schematic illustration given in Fig. 2. We retain the assumption that the central diode is everywhere in reverse bias and that $\alpha = 1$ for each emitter-base configuration. This structure is explicitly 2-dimensional, having a stripe geometry, and there is assumed to be no functional dependence on the third coordinate. With the end terminals of each base layer shorted together as shown, the current filament will locate itself along the center line of the structure. We now trace the temporal evolution toward this state, starting from an initial distribution of hole current which is assumed to be uniform. Upon arrival on P_b, the hole current flow is divided between the base

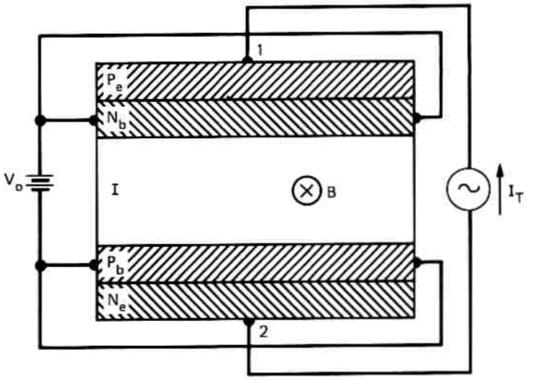


Fig. 2—Filament forming structure with multiterminal circuit connection.

contacts and, consistent with uniform spreading resistance of the base, produces a parabolic voltage profile with its maximum at the center. The total injection of electrons from N_e must correspond to a current I_T, but the base voltage profile will not permit this injection to be uniform. Because the base-emitter voltage is a maximum at the center and because the law governing electron injection is a highly nonlinear function of this voltage, the electron current density will peak sharply at the center. With only moderate lateral spreading in the I region, which is readily attainable,[7] electrons arrive at N_b with a distribution still peaked at the center. Since the average electron must now cross a greater length of resistive base than the average hole did in the uniform distribution, a greater maximum base voltage will be developed and the voltage gradient at points away from the contacts will be enhanced. This sharper voltage profile will in turn lead to an injected hole distribution more sharply peaked than the incident electron distribution. The analysis will show that, after a steady state is reached, the injected distribution of electrons and holes becomes identical. Because of the exponential injection law, this steady-state profile will become progressively sharper as I_T is increased. In particular, when the voltage from base center to base contact is 1 V, the ratio of current density at the center to that at the edge is exp $(qV/kT) \sim e^{40}$. When the filament is highly localized at the center, it is clear that the base resistance acts very much like the resistor R in series with battery V_o in Fig. 1a, and that negative resistance from terminals 1 to 2 in Fig. 2 will similarly result.

The sharpest filament profile occurs when the I region is made extremely thin to eliminate the lateral diffusive and/or space-charge spread. Although a thick I region is needed for good magnetic field sensitivity, previous work on confined electron beams in Si[7] demonstrates that highly localized distributions of electrons and holes arriving at the base layers can still be expected. Accordingly, in this paper it is assumed that space-charge spreading is negligible for reasons of low beam current or electron-hole charge compensation, and diffusion will be used to characterize the lateral spread.

When a magnetic field is applied into the plane of Fig. 2, the filament will move some distance to the right of center, producing an observable current unbalance in the external circuit. Such bodily displacement of the filament is brought about by the Lorentz force, which by virtue of the counterstreaming motion of the electrons and holes causes a Hall displacement to the right for both carrier species. If there were no effects tending to return the filament to center, the interjection of a unidirectional Hall displacement into each pass of the regenerative particle flow loop would translate the filament indefinitely to the right, in the manner depicted in Fig. 3a.

However, when the filament is shifted off-center, a "restoring force" is produced. This force is proportional to the displacement of the filament from the center, while the Lorentz force remains constant. Therefore, an equilibrium position is attained for which the return injection maximum is displaced back toward center by an amount equaling the single-pass Hall displacement, as indicated in Fig. 3b. Further insight into the nature of this equilibrium state can be gained from a study of Fig. 3c, which illustrates the relationship between the arriving hole current distribution, $J_p(x)$, and voltage profile $V_b(x)$ in base P_b. Since the distribution $J_p(x)$ is displaced to the right of center, it sends more current to the right-hand contact than to the left-hand contact because the resistance is less looking to the right. The point in the $J_p(x)$ profile which divides the leftward from the rightward flowing currents must therefore lie to the left of the centroid of $J_p(x)$.

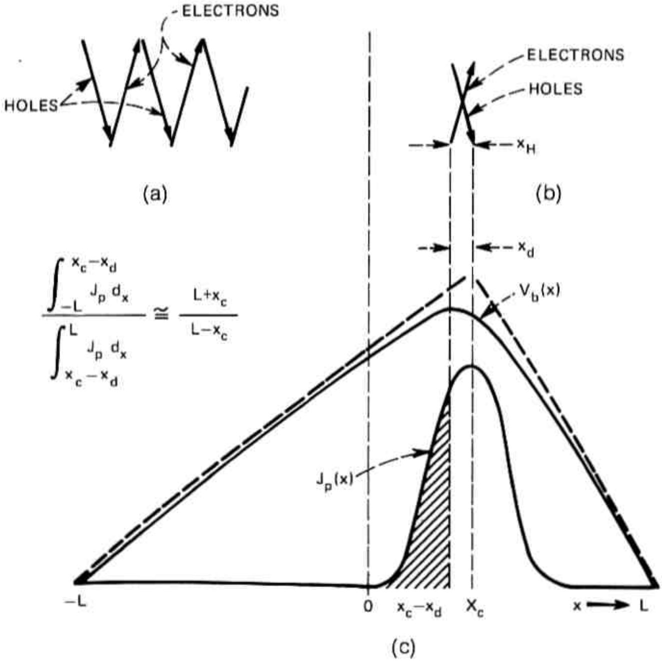


Fig. 3—(a) Representation of multipass displacement in the absence of a restoring force. (b) Representation as in (a) with restoring force. (c) Illustration of relation between hole current profile $J_p(x)$ and base voltage $V_b(x)$ for a displaced filament.

This division point is, of course, the point of maximum base voltage since it is the electric field in the base which causes current conduction toward the contacts. In the vicinity of each contact, far from the filament, the magnitude of the slope of the $V_b(x)$ curve must correspond to the total current at that contact. The ratio of these slopes is, for the present discussion, adequately characterized by the assumption that the straight-line extrapolations intersect at the centroid position x_c, as depicted in the figure, and thus correspond to a ratio $(L + x_c)/(L - x_c)$. The leftward displacement of the maximum of $V_b(x)$ from the $J_p(x)$ centroid is therefore determined by the requirement that the areas under the $J_p(x)$ curve to the right and left of the division point be in the ratio $(L + x_c)/(L - x_c)$. The significance of the leftward displacement x_d is that the return injection profile of the electrons peaks at the voltage maximum and is therefore displaced leftward from the centroid of the arriving distribution by this amount. Equilibrium occurs when x_d is equal to the rightward single-pass Hall displacement x_H.

It is apparent that, to within the above approximations, filament displacement in the magnetic field must be linear since $x_c \propto x_d$ and $x_d = x_H$. Furthermore, the sensitivity increases with drive current because this increase narrows the filament, requiring a larger off-center displacement x_c to bring x_d into equality with x_H. For narrow filaments it is possible for the displacement to be many times larger than x_H, resulting in a signal current greatly exceeding that of a Hall device of similar dimensions. As a practical matter, the short circuit signal current of devices typified by Fig. 2 will saturate at perhaps ten times that of a Hall device, because the sharpness of the profile eventually becomes diffusion-limited. However, this does not appear to be a fundamental limitation on device sensitivity, as is shown by the example at the end of Section IV.

III. DERIVATION OF CURRENT PROFILE AND TERMINAL CHARACTERISTICS IN THE ABSENCE OF A MAGNETIC FIELD

This section presents the calculation of the filament profile in the absence of a magnetic field, as well as the device terminal characteristics. It is shown that the shape of the filament can be characterized directly in terms of the device parameters in both the absence and presence of diffusion. We first consider, in Section 3.1, the highly idealized model introduced in the last section, and neglect diffusion as well. In Section 3.2 we take into account diffusion, which is the most important additional effect present in a real situation.

3.1 Fully regenerative solution

Here we develop the mathematical solution relating the filament current profile to the structure parameters and the drive current I_T. The various voltages and currents entering the analysis are shown in Fig. 4, where it must be remembered that the two contacts on each base are shorted together as in Fig. 3. The procedure followed starts with a consideration of the lower base layer. Employing the continuity equation and the base resistance per unit length, r, we derive a general relation between the hole current per unit length $J_{pi}(x)$ incident on base P_b, and the base voltage $V_b(x)$ developed with respect to the base contacts. From $V_b(x)$ and the terminal voltage V_{eL}, we find the emitter-base voltage profile and, through the junction law, the injected return electron distribution $J_{nr}(x)$. V_{eL} is ultimately determined by the requirement that the total emitter current is I_T. We can write a similar relation, for the upper base, between the incident electron profile $J_{ni}(x)$ and return hole profile $J_{pr}(x)$. In general, the complete set of self-consistent equations is then obtained by introducing the appropriate connection between the incident and return profiles of each species. For a symmetrical structure and in the absence of diffusion, $J_{nr}(x)$ can be directly equated to $J_{pi}(x)$. A single equation immediately results.

The functional dependence of $V_b(x)$ on $J_{pi}(x)$ can be written in the form

$$V_b(x) = \int_{-L}^{L} Z(x, x')J_{pi}(x')dx', \tag{1}$$

where the transfer impedance function $Z(x, x')$ is the voltage response at x to a δ-function of current incident at x'. It is easy to verify that $Z(x, x')$ is given by

$$\begin{aligned} Z(x, x') &= r(L - x)(L + x')/2L, \qquad x \geqq x' \\ &= r(L + x)(L - x')/2L, \qquad x \leqq x'. \end{aligned} \tag{2}$$

The other equations required to complete the description of the lower emitter-base configuration are the voltage balance equation

$$V_e(x) = V_b(x) + V_{eL} \tag{3}$$

and the junction law

$$J_{nr}(x) = J_s \exp\left[qV_e(x)/kT\right], \tag{4}$$

where the constant J_s has dimensions of current per unit length. Equation (4) assumes large injection, i.e., net saturation current is

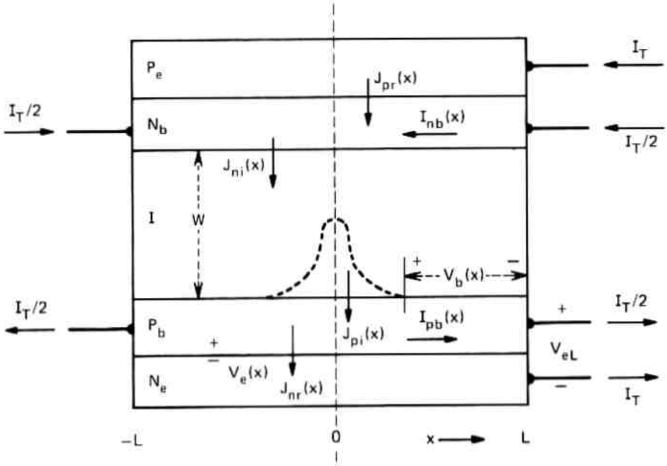


Fig. 4—Definition of variables.

negligible. Although with filamentary conduction this cannot be the case everywhere along the junction, the errors involved are unimportant when high-level injection is achieved in the vicinity of the device center.

Combining Eqs (1) to (4) produces the relation between $J_{pi}(x)$ and $J_{nr}(x)$:

$$I_{\text{reg}} \ln \left[J_{nr}(x)/J_s\right] = \int_{-L}^{x} (L - x)(L + x')J_{pi}(x')dx' \frac{1}{2L^2} + \int_{x}^{L} (L + x)(L - x')J_{pi}(x')dx' \frac{1}{2L^2} + \frac{V_{eL}}{rL}, \quad (5)$$

where

$$I_{\text{reg}} \equiv \frac{kT}{qrL} \quad (6)$$

is a structural, regenerative current constant and is the amount of current necessary to produce a voltage drop kT/q when flowing from base center to either base contact. Differentiating (5) yields

$$-\frac{I_{\text{reg}}}{J_{nr}(x)} \frac{dJ_{nr}(x)}{dx} = \frac{1}{2L^2}\left[\int_{-L}^{x} (L + x')J_{pi}(x')dx' - \int_{x}^{L} (L - x')J_{pi}(x')dx'\right], \quad (7)$$

the right-hand side of which may be identified as $1/L$ times the rightward flowing current $I_{pb}(x)$ in P_b. Accordingly, we rewrite (7) as

$$-I_{\text{reg}}\frac{dJ_{nr}(x)}{dx} = J_{nr}(x)I_{pb}(x)/L, \tag{8}$$

where

$$\begin{aligned} I_{pb}(x) &= \frac{1}{2L}\left[\int_{-L}^{x}(L+x')J_{pi}(x')dx' - \int_{x}^{L}(L-x')J_{pi}(x')dx'\right] \\ &= \frac{1}{2}\left[\int_{-L}^{x}J_{pi}(x')dx' - \int_{x}^{L}J_{pi}(x')dx' \right. \\ &\qquad\qquad\left. + \frac{2}{L}\int_{-L}^{L}x'J_{pi}(x')dx'\right]. \end{aligned} \tag{9}$$

Note that

$$J_{pi}(x) = \frac{dI_{pb}(x)}{dx}. \tag{10}$$

A single equation in one unknown is obtained by invoking the assumptions of symmetry and lack of diffusion:

$$\left.\begin{aligned} J_{ni}(x) &= J_{pi}(x) \\ J_{nr}(x) &= J_{pr}(x) \end{aligned}\right\} \text{ by symmetry,} \tag{11a}$$

$$\left.\begin{aligned} J_{ni}(x) &= J_{nr}(x) \\ J_{pi}(x) &= J_{pr}(x) \end{aligned}\right\} \text{ by diffusion} = 0. \tag{11b}$$

Clearly, all currents are equal. In particular, $J_{nr}(x) = J_{pi}(x)$, so that from (8) and (10) we obtain

$$-I_{\text{reg}}L\frac{d^2I_b(x)}{dx^2} = I_b(x)\frac{dI_b(x)}{dx}. \tag{12}$$

In (12) and thereafter we drop the superfluous subscripts; variable $I_b(x)$ still refers to the rightward flowing current in P_b, and also gives the leftward flowing current in N_b.

The nonlinear second-order differential Eq. (12) can be solved as follows. Rewriting (12) as

$$-2LI_{\text{reg}}\frac{d^2I_b(x)}{dx^2} = \frac{dI_b^2(x)}{dx}$$

and integrating from 0 to x yields

$$-2LI_{\text{reg}}\left[\frac{dI_b(x)}{dx} - \frac{dI_b(0)}{dx}\right] = I_b^2(x) - I_b^2(0). \tag{13}$$

By symmetry about the center line of the structure,

$$I_b(0) = 0. \tag{14}$$

Upon introducing the maximum value of the current profile,

$$J_o \equiv J(0) = \frac{dI_b(0)}{dx}, \tag{15}$$

eq. (13) therefore becomes

$$\frac{dI_b}{dx} = J_0 - \frac{I_b^2}{2LI_{\text{reg}}}. \tag{16}$$

Putting (16) into the form

$$\frac{dI_b}{2LJ_oI_{\text{reg}} - I_b^2} = \frac{dx}{2LI_{\text{reg}}}, \tag{17}$$

and integrating from $-x$ to x results in

$$\frac{1}{\sqrt{2LJ_oI_{\text{reg}}}}\left[\tanh^{-1}\left(\frac{I_b(x)}{\sqrt{2LJ_oI_{\text{reg}}}}\right) - \tanh^{-1}\left(\frac{I_b(-x)}{\sqrt{2LJ_oI_{\text{reg}}}}\right)\right] = \frac{2x}{2LI_{\text{reg}}}. \tag{18}$$

Again, by symmetry about the center line,

$$I_b(-x) = -I_b(x). \tag{19}$$

Using (19) and the property that $\tanh^{-1}(y)$ is an odd function of y, we obtain, after some rearrangement,

$$I_b(x) = \sqrt{2LJ_oI_{\text{reg}}}\tanh\left(\sqrt{\frac{J_oL}{2I_{\text{reg}}}}\cdot\frac{x}{L}\right). \tag{20}$$

From Fig. 4,

$$I_b(L) = \frac{I_T}{2}. \tag{21}$$

Therefore, the current profile peak J_o can be determined from the externally imposed drive current I_T with the relation

$$I_T = 2\sqrt{2LJ_oI_{\text{reg}}}\tanh\left(\sqrt{\frac{J_oL}{2I_{\text{reg}}}}\right). \tag{22}$$

Using this J_o in (20) gives the functional dependence of the base current on position in terms of the drive current and known parameters of the structure. In Fig. 5a, $2I_b(x)/I_T$ is plotted vs. x/L for various values of the dimensionless regeneration parameter $I_T/4I_{\text{reg}}$.

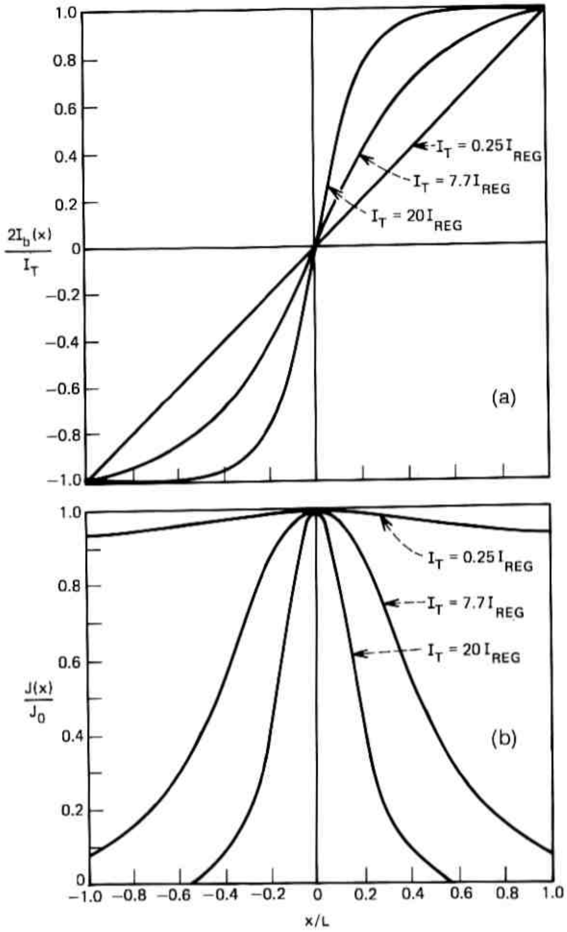


Fig. 5—(a) Position dependence of normalized base current for various values of regeneration parameters. (b) Filament current profile normalized to unity peak value for the same regeneration parameter values as in (a).

For sufficiently large drive currents such that this parameter is much greater than unity, (22) reduces to

$$I_T \approx 2\sqrt{2LJ_oI_{\text{reg}}}, \tag{23}$$

and hence

$$\frac{I_T}{4I_{\text{reg}}} = \frac{qI_TrL}{4kT} \approx \sqrt{\frac{J_oL}{2I_{\text{reg}}}}. \tag{24}$$

Since the right-hand side of (24) is just the argument of the tanh function in (22), the regeneration parameter is properly approximated by (24) for contours of which the slope is small in the vicinity of $x = L$, i.e., the filament does not touch the boundaries. The contours are adequately described by

$$I_b(x) \approx \frac{1}{2} I_T \tanh\left(\frac{I_T}{4I_{\text{reg}}} \cdot \frac{x}{L}\right), \tag{25}$$

which follows upon substituting (23) into (20).

The filament profile itself is obtained simply by differentiating (20) or, for larger I_T, (25). We find, respectively,

$$J(x) = J_o \operatorname{sech}^2\left(\sqrt{\frac{J_o L}{2I_{\text{reg}}}} \cdot \frac{x}{L}\right) \tag{26}$$

$$J(x) = \frac{I_T^2}{8I_{\text{reg}}} \operatorname{sech}^2\left(\frac{I_T}{4I_{\text{reg}}} \cdot \frac{x}{L}\right). \tag{27}$$

Plots of $J(x)/J_0$ vs. x/L for the values of $I_T/4I_{\text{reg}}$ used in Fig. 5a are displayed in Fig. 5b. It is evident that, for large values of the regeneration parameter, highly localized current flow is obtained. Equation (24) shows that this parameter is made large through increase of I_T, r, or L. However, $I_{\text{reg}}L$ is independent of L so that from (27) one sees that the absolute width of the filament is unchanged by variations of L for fixed r. For a given I_T the only way to sharpen the filament is to increase r. With a high degree of control, the current path is self-contained within an interior portion of the structure. Confined current flows without benefit of physical nonuniformity and is furthermore independent of overall dimension L. The parameter values necessary to produce a well-localized filament can be realized in a practical structure, as is demonstrated by the example in Section V.

We now proceed to calculate the terminal characteristics. The most straightforward approach consists of relating $J(L)$ to I_T with (26) and (22) and using the junction law (4) to relate $J(L)$ to V_{eL}. Recalling, however, that (4) applies only at high-level injection, which may not be satisfied at $x = L$, a more trustworthy method must be employed. Since (4) is reliable at $x = 0$, we may utilize it to find $V_e(0)$ from J_o, and relate J_o to I_T with (22). Then V_{eL} is determined from (3) and (1), where (26) is used in the integral in (1). It is clear that, whatever the junction law, $J(L)$ follows V_{eL}, as impressed through the voltage balance described above, even if V_{eL} is negative. Hence, with this method the errors in calculating V_{eL} are no greater than those in obtaining J_o and $V_b(0)$ with the large injection assumption. When J_o greatly exceeds the saturation current, these errors are small.

From (3) and (4),

$$V_{eL} = \frac{kT}{q} \ln (J_o/J_s) - V_b(0), \tag{28}$$

where, in accordance with (1),

$$V_b(0) = \int_{-L}^{L} Z(0, x')J(x')dx'. \tag{29}$$

Substitution of (2) and (26) into (29) yields

$$\begin{aligned} V_b(0) &= \frac{1}{2} J_o r \int_{-L}^{0} (L + x') \operatorname{sech}^2 \left(\sqrt{\frac{J_0 L}{2I_{\text{reg}}}} \cdot \frac{x'}{L} \right) dx' \\ &\qquad + \frac{1}{2} J_o r \int_{0}^{L} (L - x') \operatorname{sech}^2 \left(\sqrt{\frac{J_0 L}{2I_{\text{reg}}}} \cdot \frac{x'}{L} \right) dx' \\ &= J_o r \int_{0}^{L} (L - x') \operatorname{sech}^2 \left(\sqrt{\frac{J_o L}{2I_{\text{reg}}}} \cdot \frac{x'}{L} \right) dx' \\ &= 2 \frac{kT}{q} \ln \cosh \left(\sqrt{\frac{J_o L}{2I_{\text{reg}}}} \right). \end{aligned} \tag{30}$$

Substitution into (28) results in

$$V_{eL} = \frac{kT}{q} \ln (J_o/J_s) - \frac{2kT}{q} \ln \cosh \left(\sqrt{\frac{J_o L}{2I_{\text{reg}}}} \right), \tag{31}$$

where the definition (6) of I_{reg} was used. From (31)

$$V_{eL} = \frac{kT}{q} \ln \left[J_o \Big/ J_s \cosh^2 \sqrt{\frac{J_o L}{2I_{\text{reg}}}} \right]. \tag{32}$$

Together with (22), (32) specifies the terminal characteristics. It is usually reliable only for $J_o \gg J_s$ because of the large injection assumption. When the regeneration parameter is large, (23) may be introduced into (32), yielding V_{eL} directly in terms of I_T.

$$V_{eL} = \frac{2kT}{q} \ln \left[I_T \Big/ 2\sqrt{2LJ_s I_{\text{reg}}} \cosh \left(\frac{I_T}{4I_{\text{reg}}} \right) \right]. \tag{33}$$

The terminal voltage V_T developed by current source I_T, as in Fig. 1c, is

$$\begin{aligned} V_T &= V_o + V_{\text{built-in}} + 2V_{eL} \\ &= V_o + V_{\text{built-in}} + \frac{4kT}{q} \ln \left[I_T \Big/ 2\sqrt{2LJ_s I_{\text{reg}}} \cosh \left(\frac{I_T}{4I_{\text{reg}}} \right) \right]. \end{aligned} \tag{34}$$

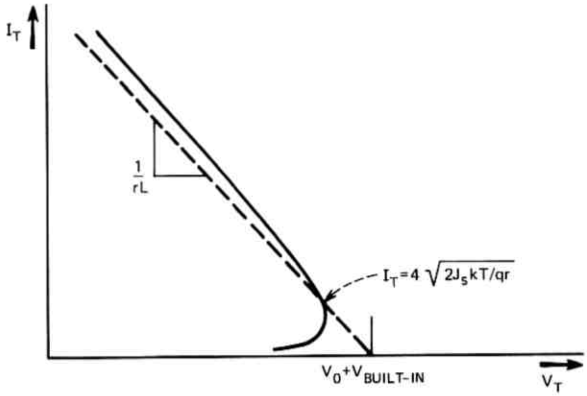


Fig. 6—Form of terminal $I_T - V_T$ characteristic based on fully regenerative solution.

For large argument the cosh function can be approximated by an exponential, permitting (34) to be rewritten as

$$V_T = V_o + V_{\text{built-in}} + \frac{4kT}{q} \ln \left[\frac{I_T}{4\sqrt{2J_s kT/qr}} \right] - rLI_T. \qquad (35)$$

The I_T vs V_T characteristic is displayed in Fig. 6. It should be observed that the asymptotic negative resistance is essentially the same as in the structure of Fig. 1. This is evidence of the fact that, with large regeneration and a filament strongly confined to the center of the structure, regions of the base away from the filament have an effect indistinguishable from external series resistors.

3.2 Diffusion limited solution

While the regenerative solution of Section 3.2 may well be applicable to PNPN structures with a narrow central junction, we must take into account the diffusive spread of the carrier streams in the wide I region of a PNIPN magnetic field sensor. In the presence of diffusion, eqs. (8) and (9) are still valid, but the equality (11b) between the incident and return currents no longer holds. For example, the stream of holes $J_{pr}(x)$ injected by emitter P_e spreads under the action of diffusion while crossing the I region, to arrive at P_b with a new broader profile $J_{pi}(x)$. An initially spike-like or Gaussian profile arrives as a Gaussian, and any other localized distribution also tends toward a

Gaussian. This relation can be expressed mathematically by

$$J_{pi}(x) = \int_{-L}^{L} G(x, x')J_{pr}(x')dx', \tag{36}$$

where $G(x, x')$ is the diffusion Green's function[8]

$$G(x, x') = \frac{\alpha_D}{\sqrt{\pi}} \exp\left[-\alpha_D^2(x - x')^2\right] \tag{37}$$

and α_D is the diffusive spreading parameter. Here α_D is given by

$$\alpha_D^2 \equiv v_d/4D_\perp W, \tag{38}$$

where W is the I region width and v_d and $D_\perp$ are the drift velocity and transverse diffusion coefficient of the carriers traversing it. It is, of course, assumed here that the diffusive spread is insufficient to cause the carrier stream to contact the boundaries at $x = \pm L$.

Utilizing the symmetry relations (11a) and substituting (36) into (9), we obtain from (8) the equation in one unknown, $J_r(x)$,

$$-I_{\text{reg}} \frac{dJ_r(x)}{dx} = \frac{1}{2L^2} J_r(x) \left[\int_{-L}^{x} (L + x')dx' \int_{-L}^{L} G(x', x'')J_r(x'')dx'' - \int_{x}^{L} (L - x')dx' \int_{-L}^{L} G(x', x'')J_r(x'')dx'' \right], \tag{39}$$

where the species subscript has been dropped. In view of the complexity of (39), we attempt only an approximate solution. It is evident that such a solution would be most difficult in the parameter range for which the diffusive spread and regenerative filament width are comparable. In the limit of small diffusion, which we shall not consider, perturbation theory could be used to find the slight modification produced in the completely regenerative solution. At the other extreme, large diffusion, the regenerative mechanism is largely interrupted and the incident current profile tends toward a diffusion-controlled Gaussian.

In the case of large diffusion, where the incident current profile is Gaussian, we may solve (39) approximately by also parameterizing $J_r(x)$ as a Gaussian, but with a different spreading parameter. This procedure can be justified in the following way. If we had a uniform incident current profile, the base voltage developed would be a parabolic function of x. Then, with the assumed exponential junction law, the injected return current is fortuitously Gaussian. This return profile will remain Gaussian whatever the form of $J_i(x)$ in the regions external to $J_r(x)$, as long as $J_i(x)$ is reasonably uniform within the region of

$J_r(x)$. Therefore, in situations where the return profile is much narrower than the incident profile, $J_r(x)$ is always well approximated by a Gaussian. In the diffusion-controlled case, this narrow Gaussian return profile diffusively spreads into the broad Gaussian incident on the opposite base, thereby closing the self-consistent loop.

We assume that

$$J_r(x) = J_o \exp(-\alpha_r^2 x^2) \tag{40}$$

with α_r the return profile spreading parameter. Then, after inserting (40) and (39) into (36), integration yields

$$\begin{aligned} J_i(x) &= \frac{\alpha_D J_o}{\sqrt{\pi}} \int_{-L}^{L} \exp[-\alpha_D^2(x - x')^2] \exp(-\alpha_r^2 x'^2)dx' \qquad (41) \\ &= \frac{\alpha_i}{\alpha_r} J_o \exp(-\alpha_i^2 x^2), \end{aligned}$$

where

$$\alpha_\imath \equiv \frac{\alpha_r \alpha_D}{\sqrt{\alpha_r^2 + \alpha_D^2}} \tag{42}$$

is the spreading parameter of the incident Gaussian. In performing the integration, it has been assumed that $\alpha_D L \gg 1$ and $\alpha_r L \gg 1$, so the limits may be taken at infinity. We insert the form (41) for (36) into the bracket on the right-hand side of (39) and integrate again. The result is

$$\begin{aligned} \int_{-L}^{x} (L + x')J_i(x')dx' - \int_{x}^{L} (L - x')J_i(x')dx' &= 2L \int_{0}^{x} J_i(x')dx' \\ &= \frac{\sqrt{\pi} L}{\alpha_r} J_o \operatorname{erf}(\alpha_i x). \qquad (43) \end{aligned}$$

Substitution of (40) and (43) into (39) yields

$$4I_{\text{reg}}\alpha_r^3 x = \frac{\sqrt{\pi}}{L} J_o \operatorname{erf}(\alpha_i x), \tag{44}$$

which clearly cannot be satisfied at all x for any spreading parameter values. The necessary approximation consists of replacing the error function by its first-order power series expansion term valid for small $\alpha_i x$. We obtain

$$2I_{\text{reg}}\alpha_r^3 = \frac{J_o}{L} \alpha_i \equiv \frac{J_o}{L} \alpha_r \alpha_o / \sqrt{\alpha_r^2 + \alpha_D^2}. \tag{45}$$

Using the normalization of (40),

$$J_o = \frac{I_T \alpha_r}{\sqrt{\pi}}, \tag{46}$$

(45) becomes

$$\alpha_r^4 + \alpha_D^2\alpha_r^2 - \frac{(I_T\alpha_D/2LI_{\text{reg}})^2}{\pi} = 0, \tag{47}$$

of which the meaningful root is

$$\alpha_r^2 = \frac{I_T\alpha_D}{2LI_{\text{reg}}\sqrt{\pi}}\left[\sqrt{1+\left(\frac{\alpha_D LI_{\text{reg}}}{\sqrt{\pi}I_T}\right)^2} - \frac{\alpha_D LI_{\text{reg}}}{\sqrt{\pi}I_T}\right]. \tag{48}$$

For (48) to be accurate requires that the return distribution fall to a negligible amplitude at values of x such that the next expansion term in erf $(\alpha_i x)$ beyond the first makes an insignificant contribution in (44). Thus, setting $x = 1/\alpha_r$, for which

$$\text{erf}\left(\frac{\alpha_i}{\alpha_r}\right) = \frac{2}{\sqrt{\pi}}\left(\frac{\alpha_i}{\alpha_r}\right)\left[1 - \left(\frac{\alpha_i}{\alpha_r}\right)^2 \Big/ 3\cdots\right], \tag{49}$$

the criterion is easily seen to be

$$\left(\frac{\alpha_i}{\alpha_r}\right)^2 \ll 3. \tag{50}$$

This is not really very stringent, because it indicates about 3 percent accuracy when the incident distribution is only three times wider than the return distribution. A simpler expression for α_r^2 than (48) may be obtained when the inequality (50) is well satisfied. We may estimate the magnitude of the dimensionless ratio $\alpha_D LI_{\text{reg}}/\sqrt{\pi}I_T$ by applying (50) to (48), together with the relation $\alpha_D \sim \alpha_i$, which follows from (42) and (50) and is used to eliminate α_i. Neglecting the departure from unity of the bracketed expression in (48), we see that there results the condition

$$\frac{\alpha_D LI_{\text{reg}}}{\sqrt{\pi}I_T} \ll \frac{3}{2}. \tag{51}$$

Therefore, in the diffusion-controlled regime, α_r^2 is well approximated by

$$\alpha_r^2 \approx \frac{I_T\alpha_D}{2\sqrt{\pi}LI_{\text{reg}}}. \tag{52}$$

Surprisingly, the Gaussian parameterization of $J_r(x)$ yields a solution which, in the absence of diffusion [$\alpha_D \rightarrow \infty$ in (47)], departs only moderately from the fully regenerative solution (27). Figure 7 compares these solutions for the same value of I_T and shows that the Gaussian approximation overestimates the peak amplitude by 28 percent and is correspondingly narrower. Although the Gaussian therefore only

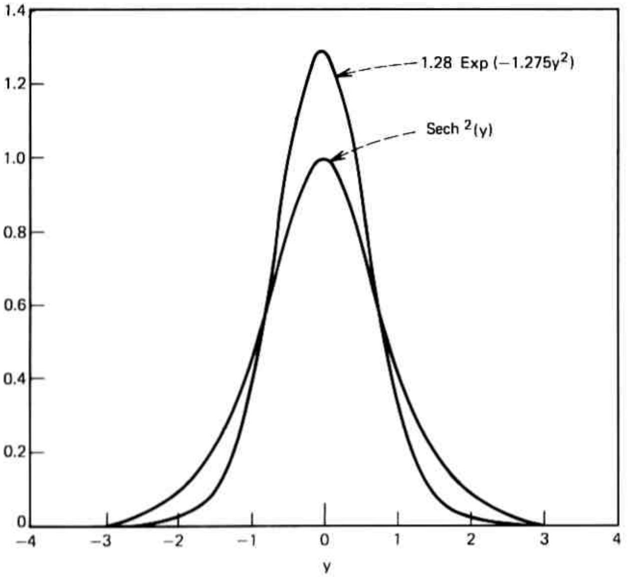


Fig. 7—Comparison of Gaussian approximation in the diffusionless case with fully regenerative solution at one value of I_T.

approximately represents the true solution, it demonstrates that we can carry the large diffusion approximation well outside its intended range of validity without a precipitous drop in accuracy.

The terminal characteristics in the diffusion-controlled case may be found with the same procedure employed for the fully regenerative solution. As long as the diffusion-controlled filament remains narrow compared to $2L$, the asymptotic negative resistance is reduced by the factor $(1 - 1/\sqrt{\pi}\alpha_i L)$ which is close to unity.

IV. RESPONSE TO A MAGNETIC FIELD

Section II gave a qualitative explanation of the magnetic response of the PNIPN structure. It was shown that unequal leftward and rightward base currents resulted. Here we calculate this current unbalance in the limit of linear response. We define the signal current I_S as the increase in current flowing out of the right-hand contact of base P_b. Small signal calculation of I_S is simplified because it presupposes that the magnetic driving force is negligibly perturbed by the magnetically produced changes in current profile. Thus, the terminal response is obtained by perturbation theory without a recalculation of the filament shape. Again, we neglect and consider the effect of diffusion in Sections 4.1 and 4.2, respectively.

4.1 Fully regenerative case

With the magnetic field directed into the page in Fig. 2, both the downward flowing holes and upward flowing electrons are deflected to the right by the Hall displacement

$$x_H = \mu BW, \tag{53}$$

which is the same for both carrier species, assuming equal mobilities. As a consequence of this deflection, relations (11b) become

$$\begin{aligned} J_{ni}(x + x_H) &= J_{nr}(x) \\ J_{pi}(x + x_H) &= J_{pr}(x), \end{aligned} \tag{54}$$

which is applicable as long as the current profiles do not contact the boundaries. The symmetry of the structure preserves relations (11a), which, together with (54), yield from (8) and (9)

$$\begin{aligned} -I_{\text{reg}} \frac{d}{dx} J_i(x + x_H) = \frac{1}{2L^2} J_i(x + x_H) \biggl[\int_{-L}^{x} (L + x')J_i(x')dx' \\ - \int_{x}^{L} (L - x')J_i(x')dx' \biggr], \end{aligned} \tag{55}$$

where we have dropped the species subscript. By changing variables, (55) can be rewritten

$$\begin{aligned} -I_{\text{reg}} \frac{d}{dx} J_i(x) &= \frac{1}{2L^2} J_i(x) \biggl[\int_{-L}^{x-x_H} (L + x')J_i(x')dx' \\ &\qquad - \int_{x-x_H}^{L} (L - x')J_i(x')dx' \biggr] \\ &= \frac{1}{2L^2} J_i(x) \biggl[\int_{-L}^{x} (L + x')J_i(x')dx' \\ &\qquad - \int_{x}^{L} (L - x')J_i(x')dx' + 2L \int_{x}^{x-x_H} J_i(x')dx' \biggr]. \end{aligned} \tag{56}$$

We recognize from (9) that $1/2L$ times the first two terms in the last bracket is $I_b(x)$. The last term, furthermore, can be written to first order in the magnetic field as

$$2L \int_{x}^{x-x_H} J_i(x')dx' \approx -2Lx_H J_i(x). \tag{57}$$

Therefore, (56) becomes

$$-I_{\text{reg}}\frac{d}{dx}J_i(x) = \frac{1}{L}J_i(x)I_b(x) - \frac{x_H}{L}J_i^2(x), \tag{58}$$

which, by (10), can be written

$$-I_{\text{reg}}\frac{d}{dx}J_i(x) = \frac{1}{2L}\frac{d}{dx}I_b^2(x) - \frac{x_H}{L}J_i^2(x). \tag{59}$$

Integration of (59) from $-L$ to L, together with the vanishing of $J_i(\pm L)$, yields

$$\tfrac{1}{2}[I_b^2(L) - I_b^2(-L)] = x_H\int_{-L}^{L} J_i^2(x)dx. \tag{60}$$

From the definition of I_S

$$\begin{aligned} I_b(L) &= \frac{I_T}{2} + I_S \\ I_b(-L) &= -\frac{I_T}{2} + I_S. \end{aligned} \tag{61}$$

Therefore, to first order in I_S, (60) becomes

$$\frac{I_S}{I_T} = x_H\int_{-L}^{L} J_i^2(x)dx/I_T^2. \tag{62}$$

Since the right-hand side of (62) is by virtue of x_H already linear in the magnetic field, the unperturbed filament profile may be used for $J_i(x)$. We can see from this equation that I_S/I_T will increase for fixed x_H when the filament profile $J_i(x)$ is made sharper. Evaluation for the fully regenerative profile (27) results in

$$\frac{I_S}{I_T} = \frac{x_H}{12L}\cdot\frac{I_T}{I_{\text{reg}}}. \tag{63}$$

Substitution for I_{reg} from (6) and for x_H from (53) gives

$$\begin{aligned} \frac{I_S}{I_T} &= \frac{\mu BW}{2L}\left(\frac{q}{kT}\frac{rLI_T}{6}\right) \\ &\approx \frac{\mu BW}{2L}\cdot\left(\frac{qV_b(0)}{3kT}\right), \end{aligned} \tag{64}$$

where $\mu BW/2L$ is the short circuit current ratio of an ideal Hall device of similar dimensions and $qV_b(0)/3kT$ is a convenient measure of the enhancement of the sensitivity with regeneration. $V_b(0) \approx rLI_T/2$ is the center-to-edge base voltage in the absence of the magnetic field and can be on the order of volts, leading to enhancement factors in the range 10 to 100.

4.2 Diffusion limited case

Putting (9) into (8) and using only the symmetry relation (11a) we have, upon dropping the species subscripts,

$$-I_{\text{reg}}\frac{d}{dx}J_r(x) = J_r(x)\frac{1}{2L^2}\left[\int_{-L}^{x}(L+x')J_i(x')dx' - \int_{x}^{L}(L-x')J_i(x')dx'\right]. \tag{65}$$

In contrast with the procedure followed in Section 4.1, it is convenient here to integrate (65) from $-L$ to $+L$ at once, to obtain

$$0 = \int_{L}^{L} J_r(x)d\int_{-L}^{x}(L+x')J_i(x')dx' - \int_{-L}^{L}J_r(x)dx\int_{x}^{L}(L-x')J_i(x')dx'. \tag{66}$$

Again, we have assumed the vanishing of the filament profile at the boundaries, i.e., $J_r(\pm L) = 0$. Upon introducing $I_e(x)$ defined by

$$J_r(x) = \frac{dI_e(x)}{dx}, \tag{67}$$

integration by parts of (66) yields

$$0 = [I_e(L) + I_e(-L)]LI_T + [I_e(L) - I_e(-L)] \times \int_{-L}^{L} xJ_i(x)dx - 2L\int_{-L}^{L} I_e(x)J_i(x)dx. \tag{68}$$

From the second form of (9) and from (61)

$$\frac{2}{L}\int_{-L}^{L} xJ_i(x)dx = I_b(L) + I_b(-L) = 2I_S. \tag{69}$$

In analogy with (61), we define I'_S by

$$I_e(\pm L) = \pm\frac{I_T}{2} + I'_S. \tag{70}$$

I_e is a construct which can be interpreted as the lateral emitter current if the emitter, like the base, had contacts at $\pm L$. I'_S is the magnetically produced unbalance in I_e. Substitution of (69) and (70) into

(68) results in

$$0 = I_T I'_S + I_T I_S - \int_{-L}^{L} I_e(x) J_i(x) dx. \tag{71}$$

This equation is merely a simplified version of the integral of (65).

To proceed further, it is necessary to introduce explicitly the simultaneous diffusive spreading and lateral magnetic displacement of the carrier stream as it crosses the intrinsic region. Combining (36) and (54) leads to the general relation between J_i and J_r,

$$J_i(x) = \int_{-L}^{L} G(x, x') J_r(x' - x_H) dx', \tag{72}$$

where $G(x, x')$ is the diffusion Green's function (37). Expanding (72) to first order in x_H yields

$$\begin{aligned} J_i(x) &\approx \int_{-L}^{L} G(x, x') J_r(x') dx' - x_H \int_{-L}^{L} G(x, x') \frac{d}{dx} J_r(x') dx' \\ &= \int_{-L}^{L} G(x, x') J_r(x') dx' - x_H \int_{-L}^{L} \frac{d}{dx} G(x, x') J_r(x') dx', \end{aligned} \tag{73}$$

where the second form has been obtained through an integration by parts with the boundary condition $J(\pm L) = 0$, and the relation $dG/dx' = - dG/dx$. Upon substituting (73) into the integral in (71), the first term of (73) gives rise to an integral of the form

$$\mathcal{G} = \int_{-L}^{L} I_e(x) dx \int_{-L}^{L} G(x, x') J_r(x') dx'. \tag{74}$$

It is possible to show by successive integration by parts that

$$\mathcal{G} = I_e(L) \int_{-L}^{L} I_e(x) G(x, L) dx - I_e(-L) \int_{-L}^{L} I_e(x) G(x, -L) dx. \tag{75}$$

The vanishing of J_r in the vicinity of the boundaries corresponds to a nearly constant value of $I_e(x)$ in the boundary regions where $G(x, \pm L)$ has a significant magnitude. By noting the normalization

$$\int_{-L}^{L} G(x, \pm L) dx = \frac{1}{2}, \tag{76}$$

we obtain

$$\begin{aligned} \mathcal{G} &= \tfrac{1}{2}[I_e^2(L) - I_e^2(-L)] \\ &= I_T I'_S. \end{aligned} \tag{77}$$

Therefore, substitution of (73) into (71) eliminates the I'_S term, leaving

$$I_T I_S = - x_H \int_{-L}^{L} I_e(x) dx \int_{-L}^{L} \frac{d}{dx} G(x, x') J_r(x') dx'. \tag{78}$$

Interchanging the order of integration, integrating by parts with respect to x, and utilizing (67) give

$$\begin{aligned} I_T I_S = -x_H\Big[I_e(L)\int_{-L}^{L} G(L, x')J_r(x')dx' \\ - I_e(-L)\int_{-L}^{L} G(-L, x')J_r(x')dx' \\ - \int_{-L}^{L}\int_{-L}^{L} J_r(x)G(x, x')J_r(x')dx\, dx'\Big]. \qquad (79) \end{aligned}$$

Because $J_r(x)$ and $G(\pm L, x)$ do not overlap, the first two integrals in (79) vanish, yielding the final result

$$\frac{I_S}{I_T} = x_H \int_{-L}^{L}\int_{-L}^{L} J_r(x)G(x, x')J_r(x')dx\, dx'/I_T^2. \qquad (80)$$

In the limit of no diffusion $G(x, x') \rightarrow \delta(x - x')$ and (80) reduces to expression (62), but (80) is valid for arbitrary diffusive spreading. For $J_r(x)$ parameterized as a Gaussian according to (40) and using (37) and the normalization (46), (80) becomes

$$\frac{I_S}{I_T} = x_H \cdot 2\alpha_r^2 L I_{\text{reg}}/I_T\sqrt{1 + 2\alpha_D^2/\alpha_r^2}. \qquad (81)$$

In the diffusion-controlled regime characterized by α_r^2 as given in (52), the radical in (81) is approximated by unity, and we find

$$\begin{aligned} \frac{I_S}{I_T} &= x_H\alpha_D/\sqrt{\pi} \\ &= x_H\sqrt{v_d/4\pi D_\perp W}. \qquad (82) \end{aligned}$$

The result (82) can also be obtained from (80) by letting $J_r(x) \rightarrow I_T\delta(x)$ for which

$$\frac{I_S}{I_T} = x_H G(0, 0) = x_H\alpha_D/\sqrt{\pi}. \qquad (83)$$

The equality of (82) and (83) demonstrates that, in the diffusion-controlled regime in which α_i/α_r need only satisfy (50), the structure nevertheless responds to a magnetic field as if the return current profile were a very sharp spike. The absence of I_T on the right-hand side of (82) indicates that diffusion saturates the magnetic sensitivity and, unlike (63), the signal is now only linearly proportional to the drive current I_T. To compare the diffusion-controlled detector with a Hall effect device, we substitute for x_H from (53), define the voltage across the I region at the center by

$$V_B = V_o - 2V_b(0), \tag{84}$$

and introduce the transverse noise temperature of the carriers defined by the Einstein relation

$$kT_n = qD_\perp/\mu. \tag{85}$$

Thus (82) becomes

$$\frac{I_S}{I_T} = \frac{\mu BW}{2L}\left(\frac{L}{W}\sqrt{\frac{qV_B}{\pi kT_n}}\right). \tag{86}$$

The expression in parentheses is the sensitivity enhancement factor for this case, which should be compared with (64), derived in the absence of diffusion. Equation (86) shows that the sensitivity of the diffusion-controlled detector is improved by increasing the central bias voltage until carrier heating predominates. At 8 V, the radical has a value of approximately 10 for W sufficiently large that $T_n \sim T$.

Equation (86) seems to suggest that large sensitivity enhancement with respect to Hall devices can be achieved by making L/W very large. This improvement is, however, illusory because it merely creates an unfavorable geometry for the Hall device. A fair comparison is possible when the device configurations are nearly square. Although in this case an enhancement factor involving only qV_B/kT_n is indicated, this should not be construed as an ultimate limitation imposed by diffusion, but rather as a structural limitation. The following example will illustrate how, for fixed W and L, the fully regenerative enhancement factor can be obtained within the constraints imposed by diffusion. An analysis has been carried out for a structure in which the emitters are contacted at $\pm L$ and have resistances per unit length approaching but less than that of the base layers. It has been found that, in the absence of diffusion, emitter resistance broadens the filament but does not diminish its off-center displacement or signal current when a magnetic field is applied. Since a broader filament is less subject to diffusive spreading when diffusion is taken into account, the effect of sufficient emitter resistance is to carry the filament formation and magnetic response out of the diffusion-controlled regime back into the fully regenerative regime. Therefore, the diffusion limit given by (86) would appear to be appropriate only to the structure analyzed in detail, rather than to be fundamental.

V. PRACTICAL MAGNETIC DETECTORS

The previous sections of this paper have established the fundamental principles according to which controlled filaments might be produced in PNIPN structures and have analyzed their magnetic sensitivity. Certain idealizations were made in order to develop a coherent theory. One purpose of this section is to give at least a preliminary account of the effect of removing these idealizations, so that we may relate the theory to practical devices. Since magnetic response has heretofore been characterized solely in terms of the short circuit signal current I_S, it is also necessary to analyze the behavior of the magnetic detector in an actual circuit which presents a finite impedance to the detector output. Several realizable circuits are considered. Finally, practical design parameters of a particular detector are given and performance predictions are made. Because filament formation in these devices requires that they be biased into the negative resistance range, there may be a tendency for ac instability, notwithstanding their apparent stability at dc. The dependence of oscillatory behavior on parasitics suggests that, at the outset, only experimental resolution of the stability question is feasible.

5.1 Removal of idealizations

The model developed thus far has been based on the explicit assumptions of (1) complete structural and electrical symmetry, (2) high level injection, (3) infinite current gain, and (4) lateral carrier stream spreading in the I region by diffusion only. It has also been implicit in the analysis that it is permissible to neglect the effects of lateral electric fields in the I region, filament position pinning resulting from structural imperfections, and possible modulation of base width and conductivity. While a detailed investigation of all these effects is beyond the scope of this paper, we shall explain why they are not apt to modify greatly the operation described in the previous sections.

In view of the regenerative nature of the filament, the assumption of infinite current gain might appear questionable. In actual fact, it is easily shown that for finite, but reasonably large, values of common emitter current gain β, device performance is only slightly degraded. We consider first the fully regenerative case, i.e., no diffusion. In the absence of a magnetic field we recall from (11a) and (11b) that $J_i(x) = J_r(x)$. When $\beta \to \infty$, the base current, and hence the base voltage $V_{b\infty}(x)$, are produced entirely by $J_i(x)$ as given by (9) and (1), respectively. For finite β, there is an additional base current component produced similarly by a current profile $J_r(x)/\beta (= J_i(x)/\beta)$ which is subtractive, and hence reduces the base voltage drop to $V_b(x) = (1 - 1/\beta) V_{b\infty}(x)$. This voltage reduction is the same as would be caused by retaining infinite β and reducing r from the original value

r_∞ to

$$r = \left(1 - \frac{1}{\beta}\right) r_\infty. \tag{87}$$

Assuming now that an increase in the actual base resistance is made to compensate for this effect, no modification results in the filament profile if the current through the battery is maintained unchanged. To do so with finite β requires an increase in emitter current by a factor $(\beta + 1)/(\beta - 1)$. It is clear that the filament disappears for $\beta < 1$, but that for $\beta \gg 1$ there need only be a small degradation.

In the diffusion-controlled regime there can be additional significant effects of finite β. When the incident profile is much broader than the return profile, we have $J_r(0)/J_i(0) = \alpha_r/\alpha_i > 1$. Therefore, in the vicinity of the origin, the injection process will give rise to subtractive base current components comparable to those produced by $J_i(x)$, unless β is sufficiently larger than α_r/α_i. The presence of such subtractive components lowers the base voltage at the origin, broadening the return profile and self-consistently lowering $J_r(0)$ until $\beta > J_r(0)/J_i(0)$ is suitably satisfied. Clearly, in the diffusion-controlled regime, finite current gain places a limit on the sharpness of the return profile which cannot be improved by increase of base resistance, i.e., $\alpha_r < \beta\alpha_D$ if the approximation of a Gaussian return profile is retained. Because the magnetic sensitivity is only weakly dependent on the return profile width if (50) is satisfied, as shown by the comparison of (82) and (83), it should only be slightly affected by finite current gain as long as $\beta \gg \sqrt{3}$.

We now briefly consider several effects that can modify filament formation and translation through localized departure from the simple theory. Lateral fields in the I region, brought about by the base layer voltage, can cause deflection[9] of the carrier streams not taken into account in the filament analysis. As a result of the symmetry of the two emitter-base configurations, there is electrical symmetry about the plane midway between the bases. Therefore, the electric field streamlines in the I region may converge near midplane, but still connect, in 1-to-1 fashion, points on the two base layers lying equidistant from filament center. Consequently, although the filament may tend to neck in at the center, this effect will not by itself give rise to additional lateral spreading. Similarly, when the filament is displaced off-center by a magnetic field, these lateral fields will not cause a net restoring force toward device center.

Filamentary instabilities characteristically occur at the particular cross-sectional location where breakdown is most easily initiated.[5] We have shown that in the present controlled filament formation mechanism, nucleation takes place at the center of the structure. It is still possible, however, that at other locations pinning points may exist for the filament because of structural inhomogeneities such as, for example, a locally enhanced injection efficiency. It is convenient to classify such inhomogeneities according to their size relative to the filament width. Large-scale inhomogeneities, which we shall assume to be reasonably weak, should result in only mild distortion of filament shape and position. In using the structure as a magnetic field sensor, this effect would produce a dc "offset voltage," but not otherwise interfere with the magnetic response. On the other hand, intense small-scale parameter fluctuations would provide distinct filament pinning points. However, in the diffusion-controlled regime this effect should be much reduced. Not only does the diffusion introduce an averaging over dimensions larger than the inhomogeneity, but the accompanying interruption of the feedback loop serves to damp down the multipass gain fluctuations. Because of the filament centering force inherent in the simple theory, pinning the filament becomes progressively more difficult at points away from device center. Ultimately, however, the importance of filament pinning will have to be determined experimentally.

In contrast with structurally associated departures from ideal behavior, localized parameter variations may occur self-consistently induced by the filament itself. Under conditions of high current density, transport in the base may be modified by increased base width or conductivity. It is well known that for transistors operated at high currents the base tends to widen. A similar effect here would lead to a decrease in the base resistance per unit length r. When there is a perfectly compensated filament of electrons and holes in the collector, however, one would expect this effect to disappear but, if there is diffusive spread of carrier streams, locally perfect compensation is absent and some base widening may still occur. A similar local decrease in r would result directly from the conductivity modulation produced by the injected carriers. This effect is readily minimized by making the base layer thin, while keeping the same sheet resistance. With a thinner base the minority carrier density for a given current is lower, while majority carrier concentration is higher. In any event, a local reduction in r will broaden the filament, but one would expect the change in shape to be more pronounced than the actual change in width. Similar modification of the filament profile can be anticipated from the falloff of injection efficiency at extremely high injection levels.[10]

We now examine the assumptions of structural and electrical symmetry. Structural asymmetries, an example of which is an inequality of base resistance, is subject to technological control and can probably be made small. Such asymmetry will invalidate (11a), resulting in inequivalent electron and hole profiles, but if reasonably small it is unlikely to affect the average filament properties or magnetic response. In contrast, the electrical asymmetry is mostly governed by the disparity of the electron and hole mobilities which is not controllable and may be quite large. An immediate and important consequence of such a mobility ratio is inequality of the electron and hole Hall displacements. It might appear that, because of this inequality, a magnetic field would disrupt the filament by pulling apart the electron and hole streams. Indeed, it has been proposed that the magnetic response of a GaAs double injection diode can be explained by such a mechanism.[5] In the present system, this phenomenon may occur at very high magnetic fields but should normally be avoidable, since the filament is broader than the single-pass Hall displacement and there is no strongly nonlinear pinning point. We have made an analysis based on a rigid displacement of the electron and hole current profiles in the fully regenerative case which indicates that no strong disruption is to be expected. The results show that the coordinate difference between centroids of the return distributions is just one-half the difference between their Hall displacements and is therefore much less than the off-center displacement. A quantitative measure of the unbalance can be obtained from the ratio of the unbalance of the signal currents in the two base layers to their average:

$$\frac{I_{Sn} - I_{Sp}}{I_S} = \frac{3I_{\text{reg}}}{I_T}\left(\frac{x_{Hn} - x_{Hp}}{x_H}\right), \tag{88}$$

where I_{Sn} and I_{Sp} are the signal currents in N_b and P_b, x_{Hn} and x_{Hp} are the Hall displacements of electrons and holes, and I_S and x_H are the average signal current and Hall displacement. The factor $I_T/3I_{\text{reg}}$ is recognized from (64) as twice the enhancement factor, and the right-hand side of (88) is therefore much less than unity.

Another effect of the mobility ratio is the destruction of the inherent filament space-charge neutrality, with the result that there will be increased lateral space-charge spreading. Qualitatively, the effects of space-charge spreading are not greatly different from those of diffusion and therefore the simple diffusion theory should account for its main features. Since, unlike diffusion, space-charge repulsion scales with filament current, it can be minimized by increasing the base resistances so that the necessary base voltage drops can be achieved at low current. Another approach is to use a circuit that equalizes the carrier densities by equating the electron-to-hole emitter current ratio to the mobility ratio, thereby restoring a nearly neutral filament.

5.2 Magnetic-detector circuit connections

Up to this point, the response to a magnetic field has been characterized only in terms of a signal current I_S. Here we consider the interconnection of the detector with a finite load impedance. In the circuit of Fig. 2, I_S could have been detected only by a perfect ammeter. Figure 8 shows a straightforward circuit modification which provides terminals for the connection of load resistors, R_L. In the absence of magnetic field, the voltage and current of the six device terminals, and therefore the filament profile, are completely unaltered by the addition of the external resistors R_{ex}, provided battery V_{00} has the value

$$V_{00} = V_0 + I_T R_{\text{ex}}. \tag{89}$$

The magnetic response is most easily understood by adopting an alternative view, in which resistors R_{ex} are considered part of extended base layers having total effective resistance $2R_{\text{eff}} = 2R_{\text{ex}} + 2rL$. If the filament remains sufficiently confined to fall well within the actual device boundaries, the whole configuration behaves as if it has a base of effective length $2L_{\text{eff}}$ related to R_{eff} by $2rL_{\text{eff}} = 2R_{\text{eff}}$, so that

$$L_{\text{eff}} = L + \frac{R_{\text{ex}}}{r}. \tag{90}$$

When the load terminals are open circuited, i.e., $R_L \to \infty$, the signal current for both the fully regenerative and diffusion-controlled case, given by (63) and (86) respectively, are unchanged by the change from L to L_{eff}. In (63) the product LI_{reg}, and hence I_S, is independent of L by (6), while in (86) I_S is explicitly independent of L as long as battery V_{00} has been increased in accordance with (89). The open circuit voltage V_{L0} is therefore

$$V_{L0} = 2R_{\text{ex}} I_S \tag{91}$$

and has the polarity given in Fig. 8(a). When the terminals are short circuited ($R_L = 0$), I_S is still that given by (63) or (86) and now flows

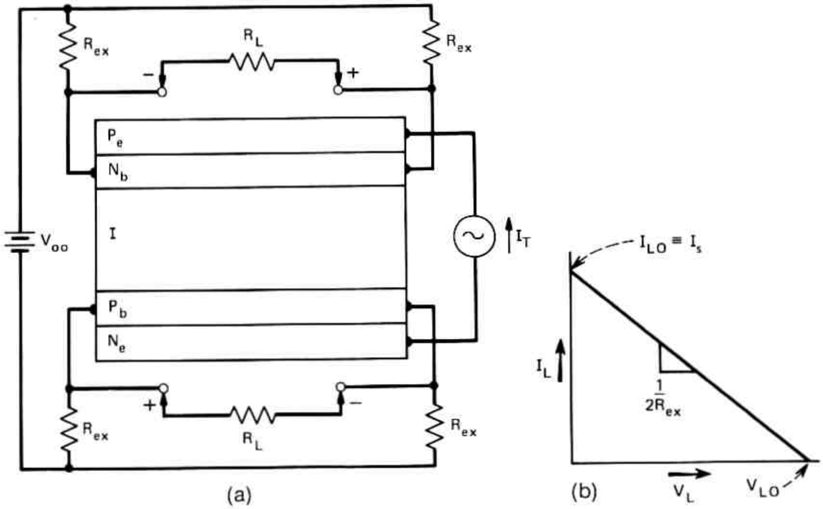


Fig. 8—(a) Magnetic field detector circuit with provision for load R_L. (b) Load line for (a).

completely through the short as I_{L0}. Within the small signal approximation, the device is linear and we obtain the load line given in Fig. 8(b). It is an interesting feature that the output impedance, $2R_{\text{ex}}$, is given solely by the magnitude of external resistors. The apparent ability to obtain an indefinite increase in open circuit voltage, by increase of R_{ex}, is just a reflection of the fact that battery V_{00} is correspondingly increased in accordance with (89). It is worth noting that, although I_S has the same value in both the open and short circuited conditions, the off-center displacement of the filament, x_c, is unequal in the ratio L_{eff}/L, reflecting the stronger centering force in the case of the short circuit.

A problem encountered with all magnetic detectors is that structural nonuniformities result in "offset voltages." If the present structure had only a single base layer, the filament would locate itself at the electrical center and there would be no offset voltage. It is expected that in the actual structure the electrical centers of the two base layers will not exactly coincide so that the filament will seek an intermediate position. The result will be an offset voltage for each base layer. It should be clear that the position of this new electrical center is determined only by structural imperfections and will therefore not depend on the enhancement factor. Consequently, the ratio of signal-to-offset voltage for this device should exceed that for the equivalent Hall effect

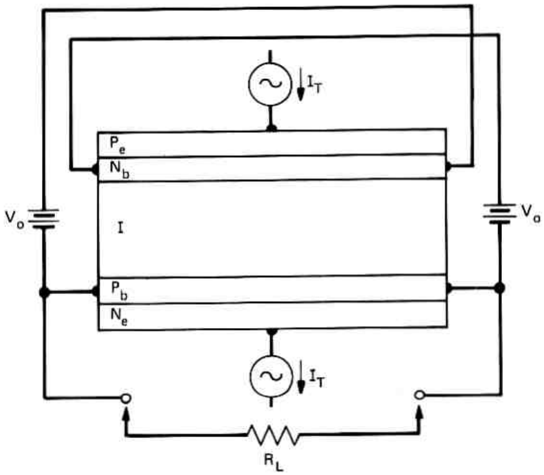


Fig. 9—Signal-summing offset-nulling magnetic detector circuit.

device by this enhancement factor. Furthermore, it is possible to envision a circuit connection, as shown in Fig. 9, in which the signal currents of the two base layers are additive, while their offset voltages are cancelled at least to first order. This circuit has a cross connection of the two base layers by means of two batteries V_o. It also has the interesting feature of displaying terminal characteristics of a nearly ideal magnetically controlled current source $2I_S$.

5.3 Sample device parameters

Figure 10 shows a realizable configuration of the magnetic detector. It is a planar structure formed on a nearly intrinsic substrate. The largest areas are base layers N_b and P_b. Application of reverse bias V_o between N_b and P_b depletes the substrate in the intervening region. Heavily doped emitters N_e and P_e are shaped to be completely on top of the base layers. This structure, with the dimensions shown, can readily be fabricated with current technology and therefore constitutes a reasonable choice for initial experiments. It is also assumed that a base sheet resistance of 10 $k\Omega/\square$ is attainable. With these constraints the structure is far from optimum, but the performance characteristics shown below nevertheless compare favorably with other magnetometers.

With a base sheet resistance of 10 $k\Omega/\square$ and a base width of 12.5 μm, we find a resistance per unit length $r = 800\ \Omega/\mu$m. Equation (6), with $L = 100\ \mu$m, then yields $I_{\text{reg}} = 0.312\ \mu$A. For a drive current $I_T = 10$

μA, the filament profile in the fully regenerative case is found from (27) to be

$$J(x) = 0.4 \operatorname{sech}^2 \left(8 \frac{x}{L} \right) \mu\text{A}/\mu\text{m}. \tag{92}$$

Referring to Fig. 7, the half amplitude points fall at $x_w = \pm 0.85L/8 \simeq \pm 11$ μm, so that the filament is indeed much narrower than the length of the base. Using (30) the corresponding voltage from base center to edge, $V_b(0)$, is 0.366 V. This result may be compared with the value 0.4 V obtained by assuming a perfectly sharp profile for which $I_T/2$ flows through a resistance rL, and indicates that the finite filament width gives rise to a less than 10 percent voltage reduction.

Two considerations enter the choice of the battery voltage V_o. First, it must be sufficient to fully deplete the substrate material between P_b and N_b. Assuming a bulk resistivity of 5 kΩ-cm or better after the necessary processing steps, 5 V would be enough to deplete a plane parallel structure 50 μm across. Allowing for some extra width necessitated by the plane configuration and some margin for being well swept out, a voltage $V_o = 11$ V, corresponding to a drop of ~ 10 V at $x = 0$, should be just adequate. The second consideration is the diffu-

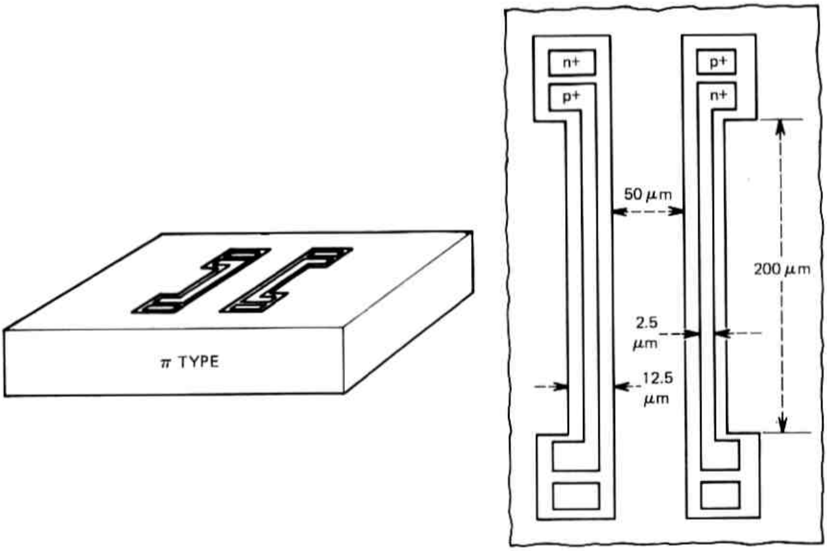


Fig. 10—Illustrative example of realizable magnetic detector.

sive spread. At $x = 0$, the average field in the I region will therefore be in the neighborhood of 2 kV/cm. This field is insufficient to greatly heat the carriers, so that it is justified in (38) to use $D_\perp$ expressed by (85), with $kT_n/q \simeq kT/q = 0.025\ eV$. The resulting value of α_D is 0.2 μm^{-1}, for which the half-amplitude half-width of the Green's function (37) is 4.16 μm. This value is small compared with the value $x_w = \pm 11\ \mu$m for the filament and works out to an additional spread of only about 12 percent. It is therefore proper to use the fully regenerative solution to calculate the magnetic response.

The regenerative enhancement factor defined in (64) for the above parameters works out to a value of 5. This value only specifies the enhancement of the short circuit signal current I_S over that of an equivalent Hall device. The full available output voltage when the device is used in the circuit of Fig. 8, however, still depends, by (91), on the choice of R_{ex}. Choosing R_{ex} arbitrarily to be 1 MΩ, adjusting V_{00} according to (89), and using (63) and (91) leads to

$$V_{\text{L0}} = 22BI_T \text{ volts},$$

which corresponds to a figure of merit of 22 V/GA. This figure of merit is of the same order of magnitude as that reported for other sensitive magnetometers.[11] It is expected that considerable improvement can result from proper design.

VI. SUMMARY

We have shown that spreading resistance in the base layers of a stripe geometry PNIPN structure, with cross section and circuit as shown in Fig. 2, leads to a localized current density profile, i.e., a filament. If lateral spread of the carrier streams in the I region can be neglected, the current density profile is adequately represented by eq. (27). A plot of this function appears in Fig. 5b, which shows that as the drive current I_T is increased, a sharpening of the filament occurs. The relevant parameter is the ratio of I_T to I_{reg}, where I_{reg}, defined in (6), is the amount of base current that would have to flow from device center to a base contact to produce a voltage drop kT/q. The numerical example given in Section V shows that for a realizable structure a typical value of I_{reg} is $\sim 0.3\ \mu$A, so that for $I_T \sim 10\ \mu$A a highly confined filament is obtained. When carrier transport in the I region is characterized by significant lateral diffusion, the ultimate sharpness of the filament becomes limited. For sufficiently large I_T it becomes a good approximation to represent both the return and incident current density profiles by Gaussians: (40) and (41), respectively. Although, as shown by (52), the return profile Gaussian continues to narrow with increasing I_T, the incident profile saturates to a width determined solely by diffusion, i.e., for $\alpha_r \rightarrow \infty$ we have $\alpha_i \rightarrow \alpha_D$, where α_D is given by (38). Using at filament center $v_d = \mu E = \mu V_B/W$ and the definition (85) of the transverse noise temperature, we find $\alpha_D^2 = qV_B/4kT_nW^2$. Therefore, the width of the diffusion controlled filament is independent of parameters characterizing the lateral extent of the structure.

The small signal linear analysis of the magnetic response of the PNIPN structure suggests that it may be regarded as a magnetically controlled current source. The principal result of the paper, eq. (62), relates the magnitude of the magnetic signal current I_S to the drive current I_T, the single-pass Hall deflection x_H, and the incident current density profile in the absence of diffusion. Noting that $I_T/2L$ represents the average current density $\langle J_i(x)\rangle$ and that for any nonuniform function $\langle J_i^2(x)\rangle > \langle J_i(x)\rangle^2$, we see from (62) that I_S/I_T will always be larger than $x_H/2L$, with the inequality increasing for progressively sharper filaments. Since $X_H/2L$ is just the ratio of short circuit signal current to drive current for an ideal Hall detector of dimensions W and $2L$, a clear advantage is indicated. A convenient measure of the enhancement is given by the factor $qV_b(0)/3kT$ in (64), where $V_b(0)$ is the center-to-edge base voltage in the absence of the magnetic field. This factor can be in the range 10 to 100. When lateral diffusion in the I region is important, (80) must be used in place of (62). Equation (80) involves the return profile $J_r(x)$ because $J_i(x)$ is explicitly related to $J_r(x)$ by the diffusion Green's function. The sensitivity enhancement still depends on the sharpness of the current density profile, but now, as shown by (83), an infinitely sharp return profile $J_r(x)$ leads to only a finite enhancement factor, given in (86) in terms of the fundamental parameters. Depending on the device geometry, the enhancement factor can again be of order 10 or more. The parameters which enter it are those pertinent to the diffusion-controlled filament and do not include r or I_T. Although this limiting behavior follows directly from the assumption of an infinitely sharp return profile, the derivation of (82) and subsequent discussion makes clear that it is also descriptive of the sensitivity when the return current profile is only moderately sharper than the diffusion-broadened incident profile. Because, within the limits set forth in Section V, the PNIPN magnetic detector behaves as a magnetically controlled current source, its useful output voltage is determined solely by the circuit in which it is imbedded. For the circuit of Fig. 8, the device considered in the numerical calculation should have a sensitivity of 22 V/GA when driven at $I_T = 10\mu A$.

An important feature of the PNIPN structure is the possible reduction of the offset level which is so troublesome in magnetic sensors. There are various ways in which this reduction can be effected. Most directly, the offset current, being of geometric origin, is not subject to the enhancement factor experienced by the signal current, and the signal-to-offset ratio is correspondingly improved. Furthermore, the addition of matched external resistors, as in Fig. 8, permits external control of the offset because such resistors act as extensions of the base layers, increasing the effective length of the device and thereby making a percentage improvement in the tolerance. A quite different approach to offset reduction is represented by the circuit of Fig. 9, in which the device incidentally appears to function as a magnetic current source. Analysis indicates that in this circuit configuration the signal currents in the base layers will be summed in R_L, while the offset currents will be nulled to first order, i.e., to the extent that they are of the same magnitude in each base layer. While the circuit of Fig. 9 may not itself turn out to be practical, it illustrates that the device can provide enough output information to make at least a first-order distinction between the signal and offset.

REFERENCES

1. J. L. Moll, M. Tanenbaum, J. M. Goldey, and N. Holonyak, Proc. IRE *44* 1956, p. 1174.
2. I. M. Mackintosh, Proc. IRE, *46*, 1958, p. 1229; F. E. Gentry, F. W. Gutzwieler, N. H. Holonyak, and E. E. Von Zastrow, *Semiconductor Controlled Rectifiers*, Englewood Cliffs, N. J., Prentice-Hall 1964.
3. N. C. Voulgaris and Edward S. Young, IEEE Trans. Electron Devices, *ED-16*, 1969, p. 468; IEEE Journal of Solid State Circuits, *SC-5*, 1970, p. 146.
4. A. M. Barnett, IBM Journal of Res. and Dev., *13*, 1969, p. 522.
5. I. J. Saunders, Solid State Electr., *11*, 1968, p. 1165.
6. D. J. Bartelink and G. Persky, "Magnetic Sensitivity of a Distributed Si Planar PNPN Structure Supporting a Controlled Current Filament," to be published.
7. G. Persky and D. J. Bartelink, J. Appl. Phys., *42*, 1971, p. 4414.
8. Philip M. Morse and Herman Feshback, *Methods of Theoretical Physics*, New York: McGraw-Hill, 1953, p. 857.
9. D. J. Bartelink, G. Persky, and D. V. Speeney, Proc. IEEE, *59*, August 2, 1970, p. 318.
10. S. M. Sze, *Physics of Semiconductor Devices*, New York, Wiley-Interscience, 1969, p. 106.
11. J. B. Flynn, J. Appl. Phys., *41*, 1970, p. 2750.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 3, March 1974
Printed in U.S.A.

Design of Transmitter and Receiver Filters for Decision Feedback Equalization

By ANDRES C. SALAZAR

(Manuscript received August 14, 1973)

We present the constructive design of finite order equalizer filters for data transmission systems employing decision feedback equalization. Both transmitter design with power constraints and receiver design with ambient noise considerations are treated. Expressions for the filter tap settings which maximize a signal-to-noise ratio are found for both baseband pulse amplitude modulation and quadrature amplitude modulation (QAM) systems. Design examples are given in a passband equivalent (of QAM) formulation for an average toll telephone connection. Neglecting the possibility of error propagation, these examples demonstrate that decision feedback equalization requires fewer taps for acceptable system performance as compared to linear equalization. The problem of postcursor size in a decision feedback equalized response is treated and shown to diminish in importance when a hybrid equalization procedure is imposed on the linear tap adjustment. The price one pays for allowing the linear filter taps to reduce the postcursor sizes in this hybrid equalizer is a lower signal-to-noise ratio.

I. INTRODUCTION

The advantage of using a nonlinear device, referred to as a decision feedback equalizer, to cancel the tails of pulses whose amplitudes have already been estimated in a PAM system has long been recognized. Figure 1 depicts the typical system in which the decision feedback mechanism has always been envisioned to perform this task. Namely, by making decisions on a symbol-by-symbol basis and by knowing the channel response precisely, a data system would be designed so that postcursor (tails of preceding pulses) ISI could be eliminated without the ambient noise penalty that a linear filter or equalizer imposes. The tacit assumption being made in any decision feedback implementation is that the signal-to-noise ratio is high without

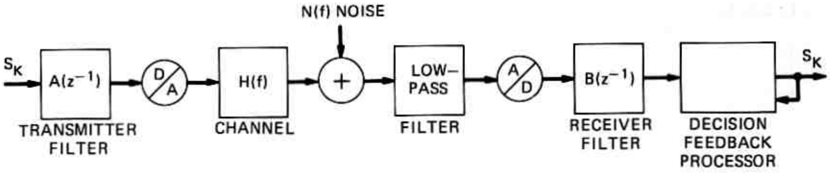


Fig. 1—Transmitter and receiver filter design.

equalization, and correct decisions are already being made with high probability.

In this paper we consider the design of finite order nonrecursive transmitting and receiving filters which counteract the two remaining sources of noise, the precursor tails which are the interfering samples of pulses whose amplitudes have not been decided upon, and ambient noise, everpresent in a communication system. In addition we seek the variation of the signal-to-noise ratio at the moment of decision when the sampling time is varied. Along with this variation of the criterion of system performance, we are also interested at each sampling time in the amount of postcursor ISI noise which the decision feedback mechanism is being asked to eliminate. This aspect of our investigation yields insight into the feasibility of decision feedback system implementation. It is, of course, possible to design the transmitting and receiving filters to achieve "hybrid" equalization between simple linear equalization and decision feedback. That is, some of the linear filter's degrees of freedom will be used to combat some postcursor ISI, although decision feedback is being used. The idea is to reduce the possibility that large postcursor tails will be produced by a linear filter whose sole job would otherwise be to reduce precursor ISI.

The system model we choose to work with is a sampled data or discrete one. In addition, the channel and the system's transmitting and receiving filters are assumed to be of finite nonrecursive type. Examples are discussed in a later section which involve voice-grade toll telephone channel spectra. These spectra have been reduced to a specified Nyquist equivalent bandwidth, both for baseband and passband applications. The timing involved in going from continuous waveforms to sampled data for these examples has been chosen to maximize a signal-to-noise criterion before any filtering is done at the receiver. Also, in the demodulation process for QAM, the carrier phase angle, if fixed, can be absorbed by the receiver's passband filter taps. (For more detail on the system model we use in the following sections, see Appendixes A and B.)

This paper follows two previous documents[1,2] that have dealt with asymptotic performance results concerning decision feedback equalization. In these previous works, the filters assumed in the decision feedback equalization scheme were of infinite length. In contrast, we focus our attention here on designing filters of finite, implementable length and study a channel which is modeled from transmission data taken from the 1969–70 Toll Connection Survey of the Bell System.

II. TRANSMITTER AND RECEIVER FILTER DESIGN (BASEBAND)

We begin by referring to Fig. 1 and denoting the channel response[†] by $\{h_n\}_0^M$. We are seeking nonrecursive filter tap weights $\{a_n\}_0^N$ and $\{b_n\}_0^N$, $N \ll M$ at the transmitter and receiver, respectively. We note the total response through the system is then

$$\{r_n\}_0^{M+2N+1} = \{a_n\}_0^N * \{h_n\}_0^M * \{b_n\}_0^N. \tag{1}$$

where $*$ denotes sequence convolution.

If we decide to sample at time τ and cancel[‡] r_k, $k > \tau$ through decision feedback, then we can define a signal-to-noise ratio

$$\rho(N, \tau, \mathbf{a}, \mathbf{b}) \triangleq \frac{r_\tau^2}{\sigma^2\|\mathbf{b}\|^2 + \sum\limits_{k<\tau} r_k^2} \tag{2}$$

representing the sampled signal in the numerator and two noise terms in the denominator. The first noise term consists of the ambient noise which is modified by the receiving filter. [We write $\mathbf{b}$ for $(b_0, b_1, \cdots, b_n)$ in E^{N+1} Euclidean space with $(\mathbf{a}, \mathbf{b})$ as the usual inner product and $\|\mathbf{b}\|^2 = (\mathbf{b}, \mathbf{b})$ the usual norm.] We have assumed that the noise samples are independent and of generalized variance[§] σ^2 and that the input binary stream of symbols is independently and fairly signed and of unit magnitude. The second denominator term is a measure of the precursor ISI.

2.1 Filter design by integral adjustment

If we assume that the transmitter filter is to be optimized independently from the receiver filter we are then concerned with the

[†] Appendix A explains our use of the sampled response $\{h_n\}$. We suppress the constant multiplier $1/T$ which converts the z^{-1} coefficients to time samples (where T is the time between samples).

[‡] We choose to cancel all postcursors. In practice, only a few are cancelled and others then become part of ISI term in (2).

[§] By generalized variance we imply that a constant multiplies the true noise sample variance. This constant takes into account the sampling speed at which we are measuring the signal-to-noise ratio.

response:

$$\{g_n\}_0^{M+N} = \{a_n\}_0^N * \{h_n\}_0^M. \tag{3}$$

Define $\mathbf{a} = (a_0, a_1, \cdots, a_N)$ and $\mathbf{h}_k = (h_k, h_{k-1}, \cdots, h_{k-N})$. We seek the maximum of

$$\rho(N, \tau, \mathbf{a}) = \frac{(\mathbf{h}_\tau, \mathbf{a})^2}{\sigma^2 + \sum_{k<\tau} (\mathbf{h}_k, \mathbf{a})^2} \tag{4}$$

subject to the constraint that $\|\mathbf{a}\|^2 = \mu^2$. That is, we place an average power constraint on the transmitter. Hence, by constructing the quadratic form induced[3] by the sum of bilinear forms $(\mathbf{h}_k, \mathbf{a})^2$, we reshape $\rho(N, \tau, \mathbf{a})$ into

$$\rho(N, \tau, \mathbf{a}) = \frac{(\mathbf{h}_\tau, \mathbf{a})^2}{\mu^{-2}(\mathbf{a}, \sigma^2 I\mathbf{a}) + (\mathbf{a}, Q\mathbf{a})}, \tag{5}$$

where I is the $(N + 1) \times (N + 1)$ identity matrix and Q is the $(N + 1) \times (N + 1)$ positive semidefinite matrix $(\sum_{k<\tau} h_{k-i}h_{k-j})$ $0 \leqq i, j \leqq N$ with $h_{-l} \equiv 0$, $l > 0$. By use of the Cauchy-Schwartz inequality we find readily that the maximum of $\rho(N,\tau,\mathbf{a})$ is achieved at

$$\mathbf{a}^* = \frac{\mu[\mu^{-2}\sigma^2 I + Q]^{-1}\mathbf{h}_\tau}{\|[\mu^{-2}\sigma^2 I + Q]^{-1}\mathbf{h}_\tau\|} \tag{6}$$

and the maximum is precisely

$$\max_{\|\mathbf{a}\|^2=\mu^2} \rho(N, \tau, \mathbf{a}) = \rho(N, \tau, \mathbf{a}^*) = (\mathbf{h}_\tau, (\mu^{-2}\sigma^2 I + Q)^{-1}\mathbf{h}_\tau). \tag{7}$$

We note that the sequence $(h_0, h_1, \cdots, h_\tau)$ is mapped by the vector $\mathbf{a}^*$ into a sequence $(\dot{g}_0, \dot{g}_1, \cdots, \dot{g}_\tau)$ which the receiver is now expected to process in forming the following signal-to-noise ratio:

$$\rho(N, \tau, \mathbf{a}^*, \mathbf{b}) = \frac{(\dot{\mathbf{g}}_\tau, \mathbf{b})}{\sigma^2\|\mathbf{b}\| + \sum_{k<\tau} (\dot{\mathbf{g}}_k, \mathbf{b})^2}, \tag{8}$$

where

$$\dot{\mathbf{g}}_k = (\dot{g}_k, \dot{g}_{k-1}, \cdots, \dot{g}_{k-N}).$$

Since $\rho(N, \tau, \mathbf{a}^*, \mathbf{b})$ is invariant to any scaling of $\mathbf{b}$, we choose to maximize the former with respect to $\|\mathbf{b}\| = 1$. By the same argument which led us to (6) and (7), we find

$$b^* = \frac{[\sigma^2 I + R]^{-1}\dot{\mathbf{g}}_\tau}{\|[\sigma^2 I + R]^{-1}\dot{\mathbf{g}}_\tau\|} \tag{9}$$

and

$$\rho(N, \tau, \mathbf{a}^*, \mathbf{b}^*) = \frac{1}{\|[\sigma^2 I + R]^{-1}\dot{\mathbf{g}}_\tau\|}(\dot{\mathbf{g}}_\tau, [\sigma^2 I + R]^{-1}\dot{\mathbf{g}}_\tau), \tag{10}$$

where R is the $(N + 1) \times (N + 1)$ matrix $(\sum_{k<\tau} \dot{g}_{k-i}\, \dot{g}_{k-j})$, $0 \leqq i, j \leqq N$. Hence, $\rho(N, \tau, \mathbf{a}^*, \mathbf{b}^*)$ in (10) represents the maximum signal-to-noise ratio achievable through the integral or independent adjustment of transmitter and receiver filters for a decision feedback system committed to sampling at time τ and constrained to use nonrecursive linear filters of length $N + 1$.

The difference between linear equalization and decision feedback can be seen readily by observing the denominator terms of the following signal-to-noise ratio:

$$\rho(N, \tau, \mathbf{a}, \mathbf{b}) = \frac{(\mathbf{g}_\tau, \mathbf{b})^2}{\sigma^2\|\mathbf{b}\|^2 + \sum_{k<\tau} (\mathbf{g}_k, \mathbf{b})^2 + \sum_{k>\tau} (\mathbf{g}_k, \mathbf{b})^2}. \tag{11}$$

For decision feedback systems, the last term in the denominator does not enter the picture because it is assumed it will be eliminated without noise penalty. However, in linear equalization, the filter $\mathbf{b}$ is expected not only to combat precursor ISI but postcursor ISI as well, with as little compromise to ambient noise as possible. We can rewrite (11) by assuming $\|\mathbf{b}\|^2 = 1$ (i.e., scaling irrelevant)

$$\rho(N, \tau, \mathbf{a}, \mathbf{b}) = \frac{(\mathbf{g}_\tau, \mathbf{b})^2}{[(\sigma^2 I + R_1 + R_2)\mathbf{b}, \mathbf{b}]}, \tag{12}$$

where R_1 and R_2 are, as usual, positive semidefinite channel response autocorrelation matrices. Here R_1 corresponds to precursor distortion while R_2 relates to postcursor ISI. We notice that, if we form for $0 \leqq \alpha \leqq 1$

$$\rho_\alpha(N, \tau, \mathbf{a}, \mathbf{b}) = \frac{(\mathbf{g}_\tau, \mathbf{b})^2}{[(\sigma^2 I + R_1 + \alpha R_2)\mathbf{b}, \mathbf{b}]}, \tag{13}$$

we can continuously vary $\rho_\alpha(N, \tau, \mathbf{a}, \mathbf{b})$ from the decision feedback formulation where $\alpha \equiv 0$ to the linear equalization case where $\alpha \equiv 1$. Thus, although we implement decision feedback equalization, it is possible to design the transmitting and receiving filters so that the amount of postcursor distortion is still mildly to strongly influential. Of course, a more general formulation of this "hybrid" design technique is possible by retracing our steps back to (11) and forming

$$\rho(N, \tau, \mathbf{a}, \mathbf{b}) = \frac{(\mathbf{g}_\tau, \mathbf{b})^2}{\sigma^2\|\mathbf{b}\|^2 + \sum_{k<\tau} (A_k\,\mathbf{g}_k, \mathbf{b})^2 + \sum_{k>\tau} (A_k\,\mathbf{g}_k, \mathbf{b})^2}, \tag{14}$$

where A_k are $(N+1) \times (N+1)$ diagonal matrices (obviously $A_k \equiv I$ for the linear equalizer case).[†]

2.2 Joint optimization of transmitter and filter design (baseband)

In the individual design of transmitter and receiver filters treated in the last section, we were able to find the optimal filters by a simple rearrangement of interference terms and applying the Cauchy-Schwartz inequality. We find that for joint filter optimization this procedure will be slightly modified and additional steps will be taken to arrive at the solution.

We recall that the total response of the system depicted in Fig. 1 is

$$\{r_n\}_0^{M+2N} = \{a_n\}_0^N * \{h_n\}_0^M * \{b_n\}_0^N \tag{15}$$

and the signal-to-noise ratio:

$$\rho(N, \tau, \mathbf{a}, \mathbf{b}) = \frac{(\mathbf{c}, \mathbf{h}_\tau)^2}{\sigma^2 \|\mathbf{b}\|^2 + \sum_{k<\tau} (\mathbf{c}, \mathbf{h}_k)^2}, \tag{16}$$

where $\mathbf{c}$ is the $2N+1$ dimensional vector formed from the sequence $\{a_n\}_0^N * \{b_n\}_0^N$ and $\mathbf{h}_k = (h_k, h_{k-1}, \cdots, h_{k-2N})$. Here again, $\rho(N, \tau, \mathbf{a}, \mathbf{b})$ is seen to be a continuous function of $\mathbf{a}$ and $\mathbf{b}$ and functionally invariant to the norm of $\mathbf{b}$. Hence, we constrain our search for the optimal $\mathbf{b}$ vector by imposing $\|\mathbf{b}\| = 1$.

The transmitter power constraint was imposed in Section 2.1 by $\|\mathbf{a}\| = \mu^2$. In practical situations, the constraint is more likely to be $\|\mathbf{a}\| \leqq \mu^2$. That is, we want to use only enough power to yield a sufficiently high signal-to-noise ratio at the receiver. For example, we constrain the receiver filter to be of unit norm since the norm is not going to contribute toward the enhancement of the signal-to-noise ratio at its output. Rather, it will be the transmitter filter power output which determines the output signal-to-noise ratio to a large extent. A way of solving the joint filter optimization problem with constraints, then, is by permitting the transmitter power level to be at that as-yet undetermined level so that the signal power through the transmitter, channel, and receiver will be at a prespecified ratio to that of the ambient noise. Hence, we have the following optimization problem:

$$\max_{\substack{\|\mathbf{h}*\mathbf{a}*\mathbf{b}\| = \eta \\ \|\mathbf{b}\| = 1}} \rho(N, \tau, \mathbf{a}, \mathbf{b}) = \max_{\substack{\|\mathbf{h}*\mathbf{a}*\mathbf{b}\| = \eta \\ \|\mathbf{b}\| = 1}} \frac{(\mathbf{a}*\mathbf{b}, \mathbf{h}_\tau)^2}{\sigma^2 + \sum_{k<\tau} (\mathbf{a}*\mathbf{b}, \mathbf{h}_k)^2}. \tag{17}$$

[†] Of course, some constraint must be put on A_k to make the maximization of ρ meaningful.

Proceeding as before, we obtain

$$(\mathbf{a}*\mathbf{b})^* = k_0\left(\frac{\sigma^2}{\eta^2}H^T H + R\right)^{-1}\mathbf{h}_\tau, \tag{18}$$

(where k_0 is determined from the constraint $\|\mathbf{h}*\mathbf{a}*\mathbf{b}\| = \eta$) with

$$\rho(N, \tau, \mathbf{a}^*, \mathbf{b}^*) = \left(\mathbf{h}_\tau, \left(\frac{\sigma^2}{\eta^2}H^T H + R\right)^{-1}\mathbf{h}_\tau\right), \tag{19}$$

where H is the $2M + 1 \times 2N + 1$ matrix such that $H(\mathbf{a}*\mathbf{b}) = \mathbf{h}*\mathbf{a}*\mathbf{b}$ where $\mathbf{h} = (h_0, h_1, h_2, \cdots, h_M)$. The matrix R is formed from the $\sum_{k<\tau} h_{k-i}h_{k-j}$ terms. Now (18) can be written in its z^{-1} transfer function representation.

$$\begin{aligned}(\mathbf{a}*\mathbf{b})^*(z^{-1}) &= k(1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \cdots + \alpha_{2N} z^{-2N})\\ A(z^{-1})B(z^{-1}) &= kQ_1(z^{-1})Q_2(z^{-1})\cdots Q_N(z^{-1}),\end{aligned} \tag{20}$$

where k is a determined constant and the Q's are quadratic factors with real coefficients. A choice of the quadratic factors for composing $A(z^{-1})$ and $B(z^{-1})$ exists. However, since $\|\mathbf{b}^*\| = 1$ we are then left with a determinable norm for $\mathbf{a}^*$. For example, we might choose

$$B^*(z) = \frac{Q_1(z^{-1})}{\|Q_1(z^{-1})\|}. \tag{21}$$

Hence,

$$A^*(z^{-1}) = \|Q_1(z^{-1})\| \cdot kQ_2(z^{-1}) \cdots Q_n(z^{-1}), \tag{22}$$

with norm

$$\|A^*(z^{-1})\| = k\|Q_1(z^{-1})\| \|Q_2(z^{-1}) \cdots Q_n(z^{-1})\|.$$

Regardless of how the quadratic factors are assigned, $B^*(z^{-1})$ is normalized and $A^*(z^{-1})$ is then left with some norm value which may be large or small. The total norm $\|\mathbf{a}*\mathbf{h}*\mathbf{b}\|$, however, was chosen to be η and for each receiver filter chosen from the quadratic factors of (20), a corresponding $\|A^*(z^{-1})\|$ results. It is of definite engineering interest to seek that quadratic factor combination which minimizes $\|A^*(z^{-1})\|$, but no obvious solution exists for this combinatorial problem. Other considerations may come into play at this point which would obviate the need for minimizing $\|A^*(z^{-1})\|$. For example, a minimum phase requirement for one of the two filters would delineate the two filters. Roundoff noise considerations for digital filter implementations might also contribute toward selecting one quadratic factor over another at the receiver. Cost considerations may warrant the splitting of the two filters into equal lengths (N even) so that the number of possible quadratic combinations is reduced considerably. In any case, this filter-splitting problem is akin to the quadratic factor placement problem in minimizing roundoff noise in digital filter implementations. In Appendix D we outline a technique for separating the transmitter and receiver filters.

Of course, it is possible to go through the same generalization on postcursor and precursor equalization that we did for the integral optimization problems of Section II. We obtain in that case

$$\rho(N, \tau, \mathbf{a}^*, \mathbf{b}^*) = \left[\mathbf{h}_\tau, \left(\frac{\sigma^2}{\eta^2}H^TH + R_1 + \alpha R_2\right)^{-1}\mathbf{h}_\tau\right], \tag{23}$$

where

$$\mathbf{a}*\mathbf{b} = k_0\left(\frac{\sigma^2}{\eta^2}H^TH + R_1 + \alpha R_2\right)^{-1}\mathbf{h}_\tau \tag{24}$$

and where $R_1(R_2)$ is the matrix corresponding to precursor (postcursor) interference terms and $0 \leqq \alpha \leqq 1$.

III. PASSBAND FORMULATION

It is possible to extend the results outlined in the previous sections to the passband equivalents of transmitter, channel, and receiver for a quadrature amplitude modulation (QAM) system.† The extension of results is not without complications, since QAM systems suffer from another form of distortion—co-channel interference (CCI). Thus, the transmitting and receiving filters will be expected to combat not only ambient noise and ISI but also co-channel intersymbol interference (CCISI).

3.1 Integral optimization

We begin by referring to Fig. 2 which illustrates the QAM system with decision feedback. We are interested in the transmitter and receiver filter designs so that a measure of transmission performance is maximized. Namely, we seek to maximize a sampled signal-to-generalized-noise ratio similar to that defined in (2). To define the terms which will appear in our performance measure, we note that the "in-phase" response at the receiver is

$$\begin{aligned}\{r_k^{(p)}\}_0^{M+2N} = {}& \{a_k^{(p)}\}_0^N*[\{h_k^{(p)}\}_0^M*\{b_k^{(p)}\}_0^N - \{h_k^{(q)}\}^M*\{b_k^{(q)}\}_0^N] \\ & - \{a_k^{(q)}\}_0^N*[\{h_k^{(q)}\}_0^M*\{b_k^{(p)}\}_0^N + \{h_k^{(p)}\}_0^M\{b_k^{(q)}\}_0^N],\end{aligned} \tag{25}$$

while the "quadrature" response at the receiver is

$$\begin{aligned}\{r_k^{(q)}\}_0^{M+2N} = {}& \{a_k^{(p)}\}_0^N*[\{h_k^{(p)}\}_0^M*\{b_k^{(q)}\}_0^N + \{h_k^{(q)}\}_0^M*\{b_k^{(p)}\}_0^N] \\ & + \{a_k^{(q)}\}_0^N*[\{h_k^{(p)}\}_0^M*\{b_k^{(p)}\}_0^N - \{h_k^{(q)}\}_0^M*\{b_k^{(q)}\}_0^N].\end{aligned} \tag{26}$$

† We will not concern ourselves with the problems of carrier acquisition and timing for the QAM system we consider here in discrete form (see Appendix A for a discussion of these items).

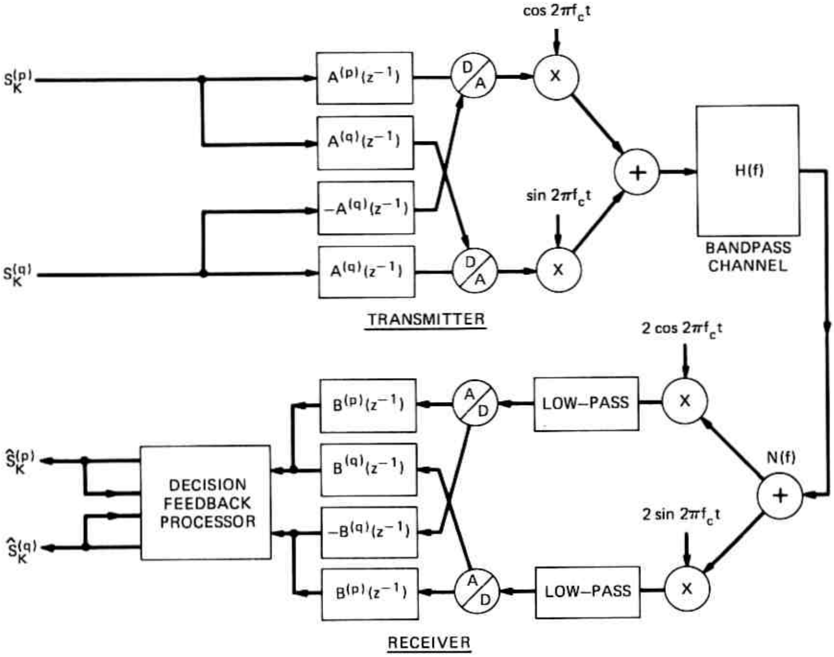


Fig. 2—QAM data system with decision feedback equalization.

Each channel response has two ISI components, an in-channel intersymbol interference term and the other due to co-channel interference. We notice that the CCI is completely eliminated if $\{b_k^{(p)}\} \equiv \{h_k^{(q)}\}$ and $\{b_k^{(q)}\} \equiv \{h_k^{(p)}\}$. However, in our considerations we will always assume $M \gg N$ so that our filters do not have a sufficient number of degrees of freedom to eliminate CCI (also, this action does constitute suboptimal filtering).

We form the in-phase resultant signal-to-noise ratio for independent input channels, independent symbols of unit magnitude with equal chance of occurrence and uncorrelated noise samples of variance σ^2. We first treat the case where the receiver filter is all pass (i.e., $\mathbf{b} = \mathbf{e}$, the identity vector in the algebra of convolution)

$$\rho_s(\underset{\mathbf{b}=\mathbf{e}}{N}, \tau, \mathbf{a}, \mathbf{b}) = \frac{[(\mathbf{a}^{(p)}, \mathbf{h}_\tau^{(p)}) - (\mathbf{a}^{(q)}, \mathbf{h}_\tau^{(q)})]^2}{\sigma^2 + \sum_{k<\tau} [(\mathbf{a}^{(p)}, \mathbf{h}_k^{(p)}) - (\mathbf{a}^{(q)}, \mathbf{h}_k^{(q)})]^2 + \cdots + \sum_{k\leq\tau} [(\mathbf{a}^{(q)}, \mathbf{h}_k^{(p)}) + (\mathbf{a}^{(p)}, \mathbf{h}_k^{(q)})]^2}. \qquad (27)$$

Now we define

$$\mathbf{a} = [\mathbf{a}^{(p)}, \mathbf{a}^{(q)}]$$

and $\mathbf{h}_\tau = [\mathbf{h}_\tau^{(p)}, -\mathbf{h}_\tau^{(q)}]$ and $\mathbf{h}_\tau^R = [\mathbf{h}_\tau^{(q)}, \mathbf{h}_\tau^{(p)}]$, vectors of $2N + 2$ unit length. Hence, we can rewrite (27) into

$$\rho_s(N, \underset{\mathbf{b}=\mathbf{e}}{\tau}, \mathbf{a}, \mathbf{b}) = \frac{(\mathbf{a}, \mathbf{h}_\tau)^2}{\sigma^2 + \sum_{k<\tau} (\mathbf{a}, \mathbf{h}_k)^2 + \sum_{k \leqq \tau} (\mathbf{a}, \mathbf{h}_k^R)^2}. \tag{28}$$

In (27) and (28) we have tacitly assumed that at the receiver each channel will "talk" to the other for the purpose of cancelling postcursor CCISI also.[†] That is, we have assumed a dual system of decision feedback equalization is being implemented.

The maximization of $\rho_s(N, \tau, \mathbf{a}, \mathbf{b})$ subject to $\mathbf{b} = \mathbf{e}$ and $\|\mathbf{a}\|^2 = \mu^2$ leads to a solution similar to that of (6) and (7):

$$\mathbf{a}^* = \frac{\mu[\mu^{-2}\sigma^2 I + Q + Q_c]^{-1}\mathbf{h}_\tau}{\|[\mu^{-2}\sigma^2 I + Q + Q_c]^{-1}\mathbf{h}_\tau\|} \tag{29}$$

$$\max_{\substack{\|\mathbf{a}\|^2 = \mu^2 \\ \mathbf{b}=\mathbf{e}}} \rho(N, \tau, \mathbf{a}, \mathbf{b}) = \rho(N, \tau, \mathbf{a}^*, \mathbf{e})$$

$$= \frac{\mu(\mathbf{h}_\tau,\ (\mu^{-2}\sigma^2 I + Q + Q_c)^{-1}\mathbf{h}_\tau)}{\|[\mu^{-2}\sigma^2 I + Q + Q_c]^{-1}\mathbf{h}_\tau\|}, \tag{30}$$

where Q and Q_c are respectively the in-phase and co-channel correlation matrices similarly formed, as was the Q matrix of (6). The $\mathbf{a}^*$ vector of (29) separates into $\mathbf{a}^{*(p)}$ and $\mathbf{a}^{*(q)}$ and the conditionally optimal transmitter bandpass filter is completely specified. Following the procedure in Section II, we now hold the transmitter design fixed at $\mathbf{a}^*$ and rewrite (25) as

$$\begin{aligned} \{r_k^{(p)}\}_0^{M+2N}\big|_{\mathbf{a}=\mathbf{a}^*} = {} & \{b_k^{(p)}\}_0^N * \{\{a_k^{*(p)}\}_0^N * \{h_k^{(p)}\}_0^M\} \\ & - \{b_k^{(q)}\}_0^N * \{\{a_k^{*(p)}\}_0^N * \{h_k^{(q)}\}_0^M\} \\ & - \{b_k^{(q)}\}_0^N * \{\{a_k^{*(q)}\}_0^N * \{h_k^{(p)}\}_0^M\} \\ & - \{b_k^{(p)}\}_0^N * \{\{a_k^{*(q)}\}_0^N * \{h_k^{(q)}\}_0^M\} \end{aligned} \tag{31}$$

$$\begin{aligned} = {} & \{b_k^{(p)}\}_0^N * \{\{g_k^{(pp)}\}_0^{M+N} - \{g_k^{(qq)}\}_0^{M+N}\} \\ & - \{b_k^{(q)}\}_0^N * \{\{g_k^{(qp)}\}_0^{M+N} + \{g_k^{(pq)}\}_0^{M+N}\}, \end{aligned} \tag{32}$$

where $\{g_k^{(u)}\}_0^{M+N}$, $u = pp, pq, qp, qq$ are recognizable from (31).

We can now write the expression for $\rho(N, \tau, \mathbf{a}^*, \mathbf{b})$ as

$$\rho(N, \tau, \mathbf{a}^*, \mathbf{b}) = \frac{[(\mathbf{b}^{(p)}, \mathbf{g}_\tau^{(p)}) - (\mathbf{b}^{(q)}, \mathbf{g}_\tau^{(q)})]^2}{\sigma^2\|\mathbf{b}\|^2 + \sum_{k<\tau} [(\mathbf{b}^{(p)}, \mathbf{g}_k^{(p)}) - (\mathbf{b}^{(q)}, \mathbf{g}_k^{(q)})]^2 + \cdots + \sum_{k \leqq \tau} [(\mathbf{b}^{(q)}, \mathbf{g}_k^{(p)}) + (\mathbf{b}^{(p)}, \mathbf{g}_k^{(q)})]^2}, \tag{33}$$

[†] Also, we are assuming we will eliminate all postcursor ISI. However, in practice, only a few postcursors would be removed. Thus, some postcursor terms would appear in the denominator of (28) in that case.

where

$$\mathbf{g}_\tau^{(p)} = (g_\tau^{(pp)} - g_\tau^{(qq)}, g_{\tau-1}^{(pp)} - g_{\tau-1}^{(qq)}, \cdots, g_{\tau-N}^{(pp)} - g_{\tau-N}^{(qq)}), \tag{34}$$

and similarly for $\mathbf{g}_\tau^{(q)}$. To maximize $\rho(N, \tau, \mathbf{a}^*, \mathbf{b})$ subject to $\|\mathbf{b}\| = 1$, we first form the concatenated vectors of $2N + 2$ length:

$$\mathbf{b} = (\mathbf{b}^p, \mathbf{b}^q), \quad \mathbf{g}_k = (\mathbf{g}_k^{(p)}, -\mathbf{g}_k^{(q)}), \quad \mathbf{g}_k^c = (\mathbf{g}_k^{(q)}, \mathbf{g}_k^{(p)}). \tag{35}$$

Hence

$$\rho(N, \tau, \mathbf{a}^*, \mathbf{b}) = \frac{(\mathbf{b}, \mathbf{g}_\tau)^2}{\sigma^2\|\mathbf{b}\|^2 + \sum\limits_{k<\tau} (\mathbf{b}, \mathbf{g}_k)^2 + \sum\limits_{k \leqq \tau} (\mathbf{b}, \mathbf{g}_k^c)^2}, \tag{36}$$

and we proceed to find that

$$\max_{\|\mathbf{b}\|=1} \rho(N, \tau, \mathbf{a}^*, \mathbf{b}) = [\mathbf{g}_\tau, (\sigma^2 I + R + R_c)^{-1}\mathbf{g}_\tau] \tag{37}$$

achieved at

$$\mathbf{b}^* = \frac{(\sigma^2 I + R + R_c)^{-1}\mathbf{g}_\tau}{\|(\sigma^2 I + R + R_c)^{-1}\mathbf{g}_\tau\|}, \tag{38}$$

where R and R_c are channel response correlation matrices of the type encountered before.

3.2 Joint optimization

To jointly optimize the transmitter and receiver passband filters, we follow virtually the same procedure found successful for the baseband case. A comparable factorization problem arises here, for which only a combinatorial solution seems to exist.

The in-phase and quadrature responses through a passband transmitter, channel, and receiver are given by

$$\begin{aligned}\{r_k^{(p)}\}_0^{M+2N} = \{h_k^{(p)}\}_0^M * (\{a_k^{(p)}\}_0^N * \{b_k^{(p)}\}_0^N - \{a_k^{(q)}\}_0^N * \{b_k^{(q)}\}_0^N) \\ - \{h_k^{(q)}\}_0^M * (\{a_k^{(p)}\}_0^N * \{b_k^{(q)}\}_0^N + \{a_k^{(q)}\}_0^N * \{b_k^{(p)}\}_0^N)\end{aligned} \tag{39}$$

$$\begin{aligned}\{r_k^{(q)}\}_0^{M+2N} = \{h_k^{(p)}\}_0^M (\{a^{(p)}\}_0^N * \{b^{(q)}\}_0^N + \{a^{(q)}\}_0^N * \{b_0^{(p)}\}_0^N) \\ + \{h_k^{(q)}\}_0^M * (\{a_k^{(p)}\}_0^N * \{b_k^{(p)}\}_0^N - \{a_k^{(q)}\}_0^N * \{b_k^{(q)}\}_0^N).\end{aligned} \tag{40}$$

Rewriting (39) and (40) in terms of a combined passband filter with responses $\mathbf{c}^{(p)} = \{c_k^{(p)}\}_0^{2N}$ and $\mathbf{c}^{(q)} = \{c_k^{(q)}\}_0^{2N}$:

$$\{r_k^{(p)}\}_0^{M+2N} = \{h_k^{(p)}\}_0^M * \{c_k^{(p)}\}_0^{2N} - \{h_k^{(q)}\}_0^M * \{c_k^{(q)}\}_0^{2N} \tag{41}$$

$$\{r_k^{(q)}\}_0^{M+2N} = \{h_k^{(p)}\}_0^M * \{c_k^{(q)}\}_0^{2N} + \{h_k^{(q)}\}_0^M * \{c_k^{(p)}\}_0^{2N}, \tag{42}$$

we form the augmented vectors $\mathbf{c} = [\mathbf{c}^{(q)}, \mathbf{c}^{(p)}]$,

$$\mathbf{h}_k = (h_k^{(p)}, h_{k-1}^{(p)}, \cdots, h_{k-2N}^{(p)}, -h_k^{(q)}, \cdots, -h_{k-2N}^{(q)})$$

and

$$\mathbf{h}_k^R = (h_k^{(q)}, h_{k-1}^{(q)}, \cdots, h_{k-2N}^{(q)}, h_k^{(p)}, \cdots, h_{k-2N}^{(p)}), h_k^{(n)} = 0, k < 0, n = p,q.$$

Our signal-to-noise ratio becomes for the in-phase channel:

$$\rho(N, \tau, \mathbf{c}) = \frac{(\mathbf{c}, \mathbf{h}_\tau)^2}{\sigma^2\|\mathbf{b}\|^2 + (R\mathbf{c}, \mathbf{c}) + (R_c\mathbf{c}, \mathbf{c})}, \tag{43}$$

where R and R_c are the now-familiar channel correlation matrices and $\mathbf{b} = (\mathbf{b}^{(p)}, \mathbf{b}^{(q)})$. It is easy to show that the norm of the receiver filter is irrelevant in the maximization of $\rho(N, \tau, \mathbf{c})$. Hence, we choose $\|\mathbf{b}\| = 1$. We now specify the amount of signal power η^2 we will need at the receiver upon choosing the optimal filters. That is,

$$\sum_{k=0}^{M+2N} |(\mathbf{h}_k, \mathbf{c})|^2 + |(\mathbf{h}_k^R, \mathbf{c})|^2 = \eta^2. \tag{44}$$

But (44) can be rewritten

$$\frac{(Q\mathbf{c}, \mathbf{c})}{\eta^2} = 1, \tag{45}$$

where Q is a sum of two correlation matrices. Hence, (43) then yields the problem:

$$\max_{\substack{\|\mathbf{b}\|=1 \\ (Q\mathbf{c},\mathbf{c})=\eta^2}} \frac{(\mathbf{c}, \mathbf{h}_\tau)^2}{\dfrac{\sigma^2(Q\mathbf{c}, \mathbf{c})}{\eta^2} + (R\mathbf{c}, \mathbf{c}) + (R_c\mathbf{c}, \mathbf{c})} \tag{46}$$

to which the solution is

$$\mathbf{c}^* = k\left(\frac{\sigma^2 Q}{\eta^2} + R + R_c\right)^{-1}\mathbf{h}_\tau \tag{47}$$

and

$$\rho(N, \tau, \mathbf{c}^*) = \left[\mathbf{h}_\tau, \left(\frac{\sigma^2 Q}{\eta^2} + R + R_c\right)^{-1}\mathbf{h}_\tau\right].$$

The constant k is determined from the constraint that $(Q\mathbf{c}, \mathbf{c}) = \eta^2$. Since $\mathbf{c} = (\mathbf{c}^{(q)}, \mathbf{c}^{(p)})$ and the vectors $(\mathbf{a}^{(p)}, \mathbf{a}^{(q)})$ and $(\mathbf{b}^{(p)}, \mathbf{b}^{(q)})$ all make up $\mathbf{c}^{(p)}$ and $\mathbf{c}^{(q)}$, we encounter a factorization problem. We can choose $(\mathbf{b}^{(p)}, \mathbf{b}^{(q)})$, normalize the receiver filter, and then are left with the transmitter filter which has a given norm. This norm is then the transmitter power required to produce η^2/σ^2 generalized signal-to-noise power at the receiver.

IV. EXAMPLES

To illustrate the difference in performance between decision feedback and linear equalization, we have taken a telephone DDD toll connection as a linear channel model. Specifically, we would like to know the difference in performance on a telephone connection when both linear and decision feedback equalization schemes are constrained to use a finite number of taps. For comparison, we compute the performance asymptotes (infinite number of taps) for each equalization scheme (see Appendix C) realized when an infinite number of taps are available. We ask whether it is possible to approach these asymptotes with a reasonable (implementable) number of taps. Another point which is raised in every implementation of decision feedback equalization is that of postcursor size. If a mistake in symbol identification is made, then the subtraction, for example, of an erroneously signed postcursor may lead to a burst of errors if the postcursor size is large. We illustrate the postcursor sizes for a passband decision feedback equalization system operating on an average telephone connection.

Figure 3 illustrates the magnitude characteristic of the average DDD toll telephone connection as measured in the 1969–70 Toll Connection Survey of the Bell System. The corresponding delay characteristic follows a parabolic shape and has been numerically integrated to yield a phase curve. As discussed in Appendix B, the bandpass channel parameters have been calculated for various carriers and various flat Nyquist spectral widths assumed at the transmitter. The spectral width was controlled by superimposing a cosine rolloff (400-Hz width centered at the Nyquist frequency) on the in-phase and quadrature spectra. Figures 4 and 5 show typical passband spectra computed for this channel. When this decomposition of the bandpass channel into in-phase and quadrature responses is achieved, it is possible to compute the performance asymptotes for linear and decision feedback equalization given in Appendix C. The result of these computations is shown in Table I. It is seen that performance decreases

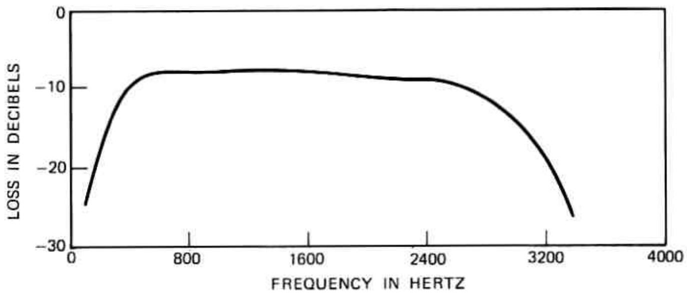


Fig. 3—Average amplitude characteristic for toll telephone connection (from 1969–70 Toll Connection Survey data).

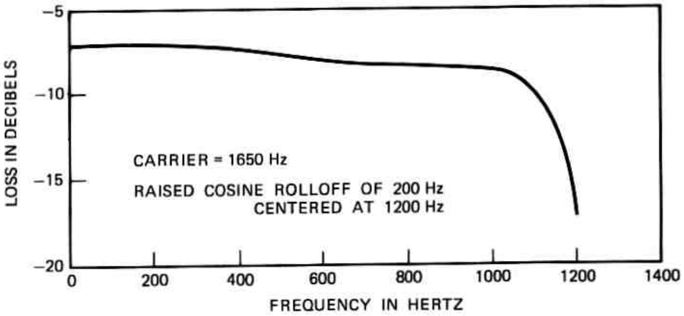


Fig. 4—Quadrature amplitude characteristic for average toll telephone connection.

with data rate and slightly with increased carrier frequency. The gap between decision feedback and linear equalization widens as speed is increased. For all computations, we have kept the total transmitted power through the channel fixed at −12 dBm, whereas the noise power spectral density was kept at that level corresponding to total noise power of −48.3 dBm over a 0–3000 Hz bandwidth. This noise level is 3 dB weaker[†] than the average noise power measured in the 1969–70 Toll Connection Survey.

A finite length receiving filter was increased in length until performance was reasonably close to the asymptote given in Table I for that speed and carrier. Figure 6 illustrates the difference in performance between decision feedback and linear equalization. It is seen that less than half the number of taps are required by the decision feedback

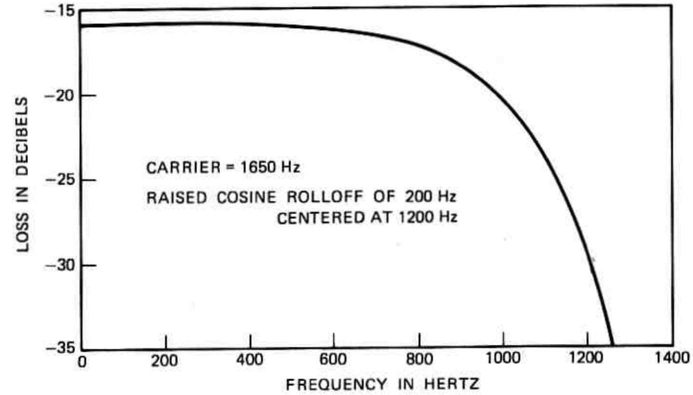


Fig. 5—In-phase amplitude characteristic for average toll telephone connection.

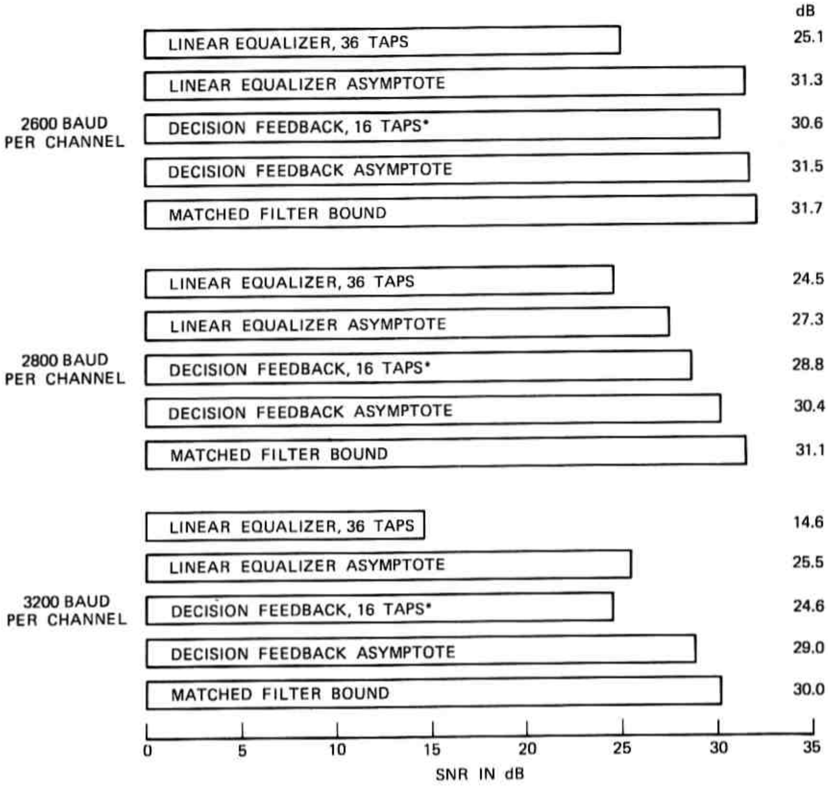
[†] Noise level was made weaker only for computational convenience.

Table I — QAM transmission asymptotic SNR in dB for average toll telephone connection
Noise at receiver — 48.3 dBm; transmitted power — 12 dBm

Rate in Baud/ch.	Carrier = 1650 Hz			Carrier = 1700 Hz			Carrier = 1800 Hz		
	L.E.	D.F.	M.F.	L.E.	D.F.	M.F.	L.E.	D.F.	M.F.
2400	31.4	31.9	32.1	30.7	31.6	31.9	29.2	30.9	31.6
2600	31.3	31.5	31.7	31.1	31.3	31.5	30.5	30.9	31.2
2800	27.3	30.4	31.1	28.5	30.2	30.9	28.8	29.9	30.5
3000	27.8	29.8	30.6	27.8	29.6	30.4	25.5	28.7	29.9
3200	25.5	29.0	30.1	25.1	28.6	29.9	19.4	27.4	29.4

L.E. = Linear equalization asymptote. D.F. = Decision feedback asymptote. M.F. = Matched filter bound.

equalizer to achieve a level of performance close to the asymptote. In addition, the linear equalizer even with its 36 tap length per channel could not keep an acceptable performance level when the data speed was increased to 3200 symbols/s/channel. On this basis, the premise that decision feedback equalization has significant advantages over linear equalization may be too readily accepted. For, if we examine postcursor sizes on one of these equalized bandpass channels, we can see that the high signal-to-noise ratio offered by decision feedback does not come without penalty. Figure 7 illustrates sample sizes of a toll telephone channel equalized with a 16-tap (8-feedback) decision feedback equalizer. The precursors, or samples before the main sample peak, are too small to be seen on this scale. However, it is clear that the postcursor adjacent to the signal sample, which is greater than half the latter's size, presents a problem. Should a decision error occur, the next signal sample could have its polarity reversed, since more than twice its strength could be subtracted out by the decision feedback processor. Thus, error propagation is possible with only a single mistake providing the ignition. Let us recall the hybrid equalization scheme discussed in Section II. We note in Fig. 8 that, for an alpha value of 0.01, we diminish the size of the large postcursor and more evenly distribute the heights of all the postcursors to be subtracted by the decision feedback processor. It is now apparent that no one postcursor is large enough to reverse the polarity of the signal should a decision error occur. It will take several consecutive decision errors, for example, before this can happen now. However, we lose 1 dB in signal-to-noise ratio for this example when we opt for this mitigation of the postcursor size problem. Of course, a trade-off exists between



NOISE AT RECEIVER = −48.3 dBm
TRANSMITTED POWER = −12 dBm
CARRIER FREQUENCY = 1650 Hz

* 8 TAPS ARE FEEDBACK

Fig. 6—Performance of finite equalizers for average toll telephone connection.

the loss in signal-to-noise ratio and reduction of postcursor size by means of this method.

V. SUMMARY

We have treated the design of finite length transmitting and receiving filters for a data system employing decision feedback equalization. Our purpose here was to examine the difference in performance between linear and decision feedback equalization on a given data channel. Sequential and joint optimization of transmitting and receiving filters were treated for an all-Nyquist equivalent data system. Although the solutions for the optimum tap settings

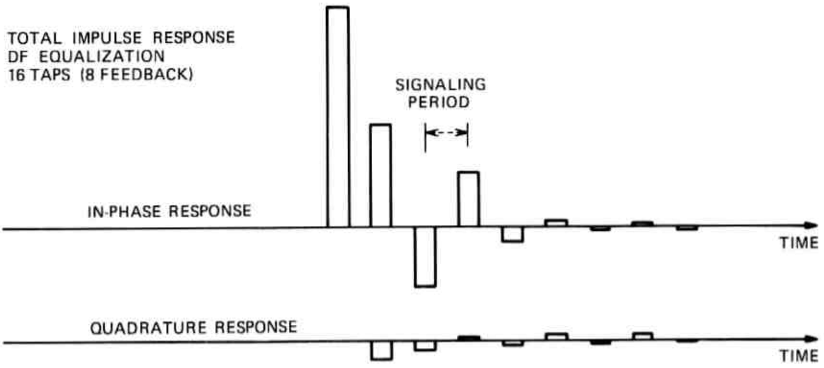


Fig. 7—Postcursor size problem and mitigation.

and signal-to-noise ratio were derived in general terms, applying the results to the spectrum of a toll telephone connection was of special interest. For this channel example, it was found that fewer filter taps were required for decision feedback equalization to achieve a reasonable performance level. The problem of postcursor size for an overall response of a passband decision feedback equalized system can be mitigated by a hybrid equalization scheme. The price for allowing the linear filter taps to diminish the postcursor sizes in this hybrid equalizer is a lower signal-to-noise ratio.

APPENDIX A

Details about the discrete channel model

The lowpass filters in the A/D or D/A conversion process shown in Figs. 1 and 2 delimit the channel frequency band which supports data

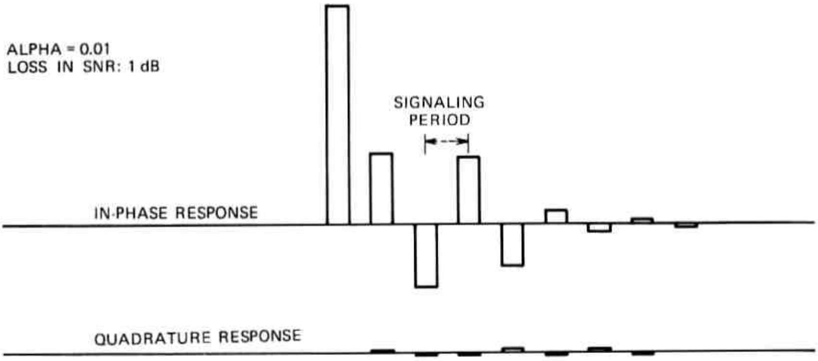


Fig. 8—Total impulse response hybrid equalization.

transmission. Hence, the channel can be seen as a bandlimited medium and can also be reduced to discrete form for M sufficiently large:

$$H(f) = \sum_{n=0}^{M} \equiv h_n \, e^{-jn2\pi fT} \quad |f| \leqq \frac{T}{2},$$

where T is the data symbol interval or $1/2T$ is the Nyquist frequency. The point here is that $\{h_n\}$ is dependent on the timing chosen for this reduction. Obviously, a timing exists which maximizes a signal-to-noise ratio, for example, of the unequalized response. We have found by experimentation that this timing was an excellent approximation to the timing which leads to a maximum signal-to-noise ratio after equalization.

For a bandpass channel, the decomposition into discrete form takes place in two steps. First, a carrier frequency is chosen, and in-phase and quadrature spectra are then computed. A constant carrier phase is then a variable parameter. However, it is easily shown that this carrier constant can be absorbed by either the demodulation process or the passband equalizer tap settings.

It is important to recall that the time samples of the spectrum

$$H(f) = \sum_{n=0}^{M} h_n \, e^{-jn2\pi fT}$$

are $\{\frac{1}{T} \cdot h_n\}_0^M$. Hence, in the formation of the signal-to-noise ratio:

$$\rho = \frac{h_\tau^2}{\sigma^2 T^2 + \text{ISI}}$$

we form the generalized variance parameter $\sigma^2 T^2$ where σ^2 is the noise sample variance. This accounts for this transformation from Fourier coefficients $\{h_n\}_0^M$ to time samples $\{1/T \cdot h_n\}_0^M$.

APPENDIX B

Channel data from 1969–70 toll connection survey

The average loss and delay measurements of over 600 toll voice-grade connections made in a 1969–70 survey are recorded in Ref. 4. For our channel model, interpolative curves were constructed from the average survey measurements made on 20 frequencies. A linear loss slope was appended at the lower frequency end to extrapolate loss down to zero frequency. The slope of the loss curve in decibels at the lowest measurement frequency (200 Hz) was used for this extrapolation. A constant was added to the integrated delay curve to achieve zero phase at zero frequency.

Passband responses at several carrier frequencies were then formulated from the interpolated baseband data. An impulse response was calculated for each in-phase and quadrature channel. Timing for the two channels was chosen to maximize the squared sampled signal to mean square ISI and CCISI before the receiver filter. One hundred eight Nyquist samples ($\{h_n^{(p)}\}_{h=0}^{179}$ and $\{h_n^{(q)}\}_0^{179}$) represented each passband channel.

APPENDIX C

Asymptotic MSE as derived by Falconer-Foschini [2]

We list here the formulas for the MSE as achieved by linear equalization and decision feedback for passband systems (here, independent binary ±1 transmission is assumed with $N_0/2$ input noise spectrum).

$$(\mathrm{MMSE})_{\mathrm{linear}} = \int_{-1/2T}^{1/2T} T\left(\frac{X_0(f)}{N_0} + 1\right)^{-1} \mathrm{df} \tag{48}$$

$$(\mathrm{MMSE})_{\mathrm{df}} = \exp\left\{T \int_{-1/2T}^{1/2T} \log\left(\frac{X_0(f)}{N_0} + 1\right)^{-1} \mathrm{df}\right\}, \tag{49}$$

where

$$X_0(f) = \frac{1}{T}\sum_n \left|G_1\left(f + \frac{n}{T}\right) + jG_2\left(f + \frac{n}{T}\right)\right|^2 \times \left|C_1\left(f + \frac{n}{T}\right) + jC_2\left(f + \frac{n}{T}\right)\right|^2.$$

The passband transmitter and channel characteristics are denoted by $G_1 + jG_2$ and $C_1 + jC_2$, respectively. For comparison purposes, it is simple to show that the matched filter bound is

$$(\mathrm{MMSE})_{\mathrm{mf}} = \left\{T \int_{-1/2T}^{1/2T} \left(\frac{X_0(f)}{N_0} + 1\right)\mathrm{df}\right\}^{-1}. \tag{50}$$

It is of interest to note that we can prove that expressions (48), (49), and (50) follow the sequence

$$(48) \leqq (49) \leqq (50)$$

by invoking Jensen's Inequality for the logarithm as the concave function. It is clear that, for the ideal channel and transmitter, i.e., $X_0(f) \equiv 1/T$, we have

$$(\mathrm{MMSE})_{\mathrm{linear}} = (\mathrm{MMSE})_{\mathrm{df}} = (\mathrm{MMSE})_{\mathrm{mf}} = \frac{1}{1 + (N_0T)^{-1}}.$$

APPENDIX D

A technique for separating transmitter and receiver filters

We wish to determine that factorization of

$$A(z^{-1})B(z^{-1}) = \prod_{n=0}^{N} Q_n(z^{-1}),$$

which minimizes $\|A(z^{-1})\|$ while $\|B(z^{-1})\| = 1$. Each

$$Q_n(z^{-1}) = 1 + a_1^{(n)}z^{-1} + a_2^{(n)}z^{-2}$$

is a quadratic factor. We assume no real roots occur, although the extension of the technique we present here to include real roots is obvious. Now

$$\|Q_n(z^{-1})\|^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} |Q_n(e^{-j2\pi fT})|^2 df = 1 + (a_1^{(n)})^2 + (a_2^{(n)})^2.$$

We notice that, upon choosing $B(z^{-1}) = \prod_{n_k \in N_B} Q_{n_k}(z^{-1})$ (where $N_B \cup N_A = \{0, 1, 2, \cdots, N\}$), then

$$\|A(z^{-1})\| = \| \prod_{n_k \in N_B} Q_{n_k}(z^{-1}) \| \, \| \prod_{n_m \in N_A} Q_{n_m}(z^{-1}) \|. \tag{51}$$

Thus, what we really want to do is select a partition of the Q_k factors so that the product of the norms of the partition factors is minimized. Much like the quadratic factor partitioning problem in digital filter implementation for minimizing roundoff noise, the only method for obtaining the global minimum of $\|A(z^{-1})\|$ seems to be the formation of all possible combinations of quadratic factors. When N is large, say, 20, this combinatorial method is time-consuming even when the filters are forced to be of the same order.

A technique for constructing the partition which sequentially minimizes $\|A(z^{-1})\|$ is first begun by reordering the quadratic factors by norm $\{Q_{nl}\}_{l=0}^{N}$. We think of the two norms of (51) as bins, and we sequentially fill those bins with quadratic factors. We insert one of two quadratic factors of largest norms into the first bin and the second factor into that same bin. We evaluate the norm of the first bin and now compare it to the product of the norms of the individual factors. Whichever placement results in smaller norm product, we choose as our partition initialization. Thus, at the end of the first step we have either

$$\begin{array}{cc} \text{bin 1} & \text{bin 2} \\ \|Q_{n_1}Q_{n_2}\| & \|1\| \end{array}$$

or

$$\overset{\text{bin 1}}{\|Q_{n_1}\|}\ \overset{\text{bin 2}}{\|Q_{n_2}\|},$$

depending on which product is smaller. The next factor, Q_{n_3}, is brought into the current partition and the products again are tested as to whether Q_{n_3} minimizes the product when placed in bin 1 or bin 2. The process continues until all quadratic factors are placed into either bin.

This procedure has been programmed and tested on actual filter quadratic factors. It has been our experience that the resulting factorization was close to the optimal one. To cite an example: Ten quadratic factors were randomly placed into two bins 500 times. The product of the norms of the two bins' contents ranged from 0.584 to 1183.33. The partition which our procedure yields for this set of quadratic factors had the product value of 0.646. Only 36 of the 500 partitions yielded smaller products. But little could be gained by using any of these 36 partitions. However, the worst partition was four orders of magnitude away from the outcome of our procedure. This is possibly what is most important, namely finding a partition very far away from the worst one.

REFERENCES

1. J. Salz, "Optimum Mean Square Decision Feedback Equalization," B.S.T.J., *52*, No. 8 (October 1973), pp. 1341–1373.
2. D. D. Falconer and G. J. Foschini, "Theory of Minimum Mean Square Error QAM Systems Employing Decision Feedback Equalization," B.S.T.J., *52*, No. 9 (December 1973), pp. 1821–1849.
3. P. R. Halmos, *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*, New York: Chelsea, 1957, p. 32.
4. T. W. Thatcher, Jr., and F. D. Duffy, "Analog Transmission Performance of Toll Connections on the DDD Telephone Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1311–1347.

Copyright © 1974 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 3, March 1974
Printed in U.S.A.

Some Properties of the Erlang Loss Function

By D. L. JAGERMAN

(Manuscript received March 5, 1973)

This paper develops the properties of the Erlang loss function, $B(N, a)$, used in telephone traffic engineering. The extension to a transcendental function of two complex variables is constructed, thus permitting the methods of complex analysis to be employed for the further study of its properties. Exact representations, Rodrigues formulas, and addition theorems are given both for the loss function and for the related Poisson-Charlier polynomials. Asymptotic formulas and approximations are developed for the loss function and also for its derivatives. A table of coefficients is included which, together with one of the asymptotic formulas, permits computation of $B(N, a)$ by simple means even when the number of trunks, N, is very large. This same table is used to obtain $\partial B(x, a)/\partial x$.

I. INTRODUCTION

The Erlang loss function

$$B(N, a) = \frac{a^N}{N!} \Big/ \sum_{j=0}^{N} \frac{a^j}{j!} \tag{1}$$

is fundamental to the study of telephone trunking problems. A. K. Erlang[1] used $B(N, a)$ to express the probability that a call, which is a member of a Poisson stream of parameter a, arriving at a group of N telephone trunks will be rejected. Later studies of trunking problems have shown the desirability of enlarging the scope of applications of the loss function. For example, the consideration of trunk groups with nonintegral number of trunks arises in determining the equivalent number of trunks in Wilkinson's "equivalent random method." [2] Methods for accomplishing the computation by interpolation are given by Rapp[3] while continued fraction procedures for accurate computation are given by Levy-Soussan[4] and Burke.[5] Derivatives with respect to N and a arise in optimal trunk group size apportionment problems. See, for example, Akimaru and Nishimura[6,7] who studied such models and prepared tables of derivatives. In some investigations, rapid and accurate approximate computations of $B(N, a)$ for very large trunk groups are needed. This occurred in the study of certain satellite telephonic communication problems.[8,9] The need thus arises of enlarging the definition of $B(N, a)$ as given in (1). Of course, that is done implicitly in the above investigations. It has been customary to extend the definition of $B(N, a)$ by use of an integral formula (Theorem 3) ascribed to Fortet. This integral formula is used in (23) to define a transcendent, $B(z, \alpha)$, for complex z and α. The extension to the complex plane in both z and α permits the powerful methods of complex analysis to be applied for obtaining exact, asymptotic, and approximate representations.

It is the purpose of this paper to provide an investigation into the properties of $B(z, \alpha)$ with the object of generalizing known results, obtaining new results, and presenting practical methods for application to the class of problems outlined above.

Part II derives exact relations satisfied by $B(z, \alpha)$. Similar relations for the related Poisson-Charlier polynomials, $G_j(z, \alpha)$, are derived in the appendix. These relations provide efficient means for exact computation; thus, Theorems 1 and 2 constitute a practical method of computing $B(N, a)$ to a prescribed accuracy for isolated computations. Similarly, the use of Theorem 5 enables one to compute $B(z, \alpha)$ even for nonintegral number of trunks. Theorem 6 may be similarly employed. The relationship of $B(z, \alpha)$ and $G_j(z, \alpha)$ to Whittaker functions as given in Theorems 7 and 24 is the key for linking up these functions with the more well-known functions of applied mathematics, i.e., hypergeometric functions and Laguerre polynomials. The Rodrigues Theorems 8 and 22 are useful for the evaluation of integrals of the form

$$\int f^{(r)}(a)a^{-1}e^{-a}B(N, a)^{-1}da, \qquad \int f^{(r)}(a)a^{z}e^{-a}G_j(z, a)da \qquad (2)$$

and, as in the case of Theorem 22, for obtaining an integral representation. The addition Theorems 9, 10, 26, and 27 yield group-theoretic structure information which is useful for simplifying formulas containing these functions, and for the evaluation of integrals. The evaluation of an integral, by means of generating functions and Theorem 10, was done in Part IV to obtain ultimately an approximate formula for $\partial B(x, a)/\partial x$. A general use of the exact relations is to serve as a springboard for asymptotic and approximate results and also for their error estimations. This is well illustrated in Part III of the paper.

The asymptotic expansions of Part III are also representations of $B(z, \alpha)$ but, unlike those of Part II, when used as approximate formulas for computation they cannot yield results of arbitrarily high accuracy, i.e., the accuracy depends on specific values of parameters. Theorem 11 is particularly useful for computation when $|z/\alpha|$ is small. It may be used for the computation of $B(z,\alpha)$ for fractional number of trunks by computing $B(z, \alpha)$ for $0 < z < 1$ and then using the recurrence formula of Theorem 4. Theorem 11 includes well-known asymptotic results, e.g.,

$$B\left(-\frac{1}{2}, a\right)^{-1} = \sqrt{\pi a}e^a(1 - \operatorname{erf} \sqrt{a}) \sim 1 - \frac{1}{2}a^{-1} + \frac{1.3}{2^2}a^{-2} - \frac{1.3.5}{2^3}a^{-3} + \cdots, \qquad a \to \infty, \tag{3}$$

$$B(-1, a)^{-1} = -ae^aE_i(-a) \sim 1 - a^{-1} + 2!\,a^{-2} - 3!\,a^{-3} + \cdots, \qquad a \to \infty.$$

An undesirable feature of many methods of computing $B(x, a)$ is the dependence of the computational effort, e.g., time of computation, on the value of x; thus, the larger the value of x the greater the computational effort. Theorem 14 overcomes this defect; the computational effort is independent of the size of x. Theorem 14 is easily usable even with a desk machine regardless of how large x is. The accuracy, however, depends on x and a parameter c. For fixed c the accuracy improves with increasing x. When x is fixed, the accuracy deteriorates when c is large and negative but greatly improves as c is increased. To facilitate the use of Theorem 14, Table I gives required coefficients, namely, $a_0(c)$, $a_1(c)$, $a_2(c)$. To use the table, one computes

$$c = \frac{a - x}{\sqrt{x}}, \tag{4}$$

then

$$B(x, a)^{-1} \cong a_0(c)\sqrt{x} + a_1(c) + \frac{a_2(c)}{\sqrt{x}}. \tag{5}$$

Possibly, one should comment that the range of values of x, c for which (5) is accurate is not as important as the fact that it is accurate over a wide range of values of $B(x, a)$, that is, values encompassing the ranges of most applications. For quantitative limitations, see Fig. 1. A method of obtaining $\partial B(x, a)/\partial x$ based on Theorem 14 is given in Part IV. This uses the formula

$$\frac{\partial B(x, a)}{\partial x} \cong -\frac{B(x, a)^2}{2\sqrt{x}}\left(a_0 - \frac{a_2}{x}\right) - \frac{x + a}{2x}B(x, a)\left\{\frac{x}{a} - 1 + B(x, a)\right\}. \tag{6}$$

Table I — Coefficients for evaluation of B(x, a) and ∂B(x, a)/∂x

c*	a_0	a_1	a_2
−3.0	225.3	2032	13726
−2.9	167.7	1367	8536
−2.8	126.0	925.4	5334
−2.7	95.63	630.5	3348
−2.6	73.28	432.2	2111
−2.5	56.70	298.0	1336
−2.4	44.29	206.7	848.1
−2.3	34.92	144.1	540.2
−2.2	27.80	100.9	345.0
−2.1	22.33	71.07	220.7
−2.0	18.10	50.27	141.4
−1.9	14.80	35.71	90.70
−1.8	12.21	25.49	58.17
−1.7	10.16	18.27	37.28
−1.6	8.521	13.15	23.86
−1.5	7.205	9.522	15.23
−1.4	6.139	6.936	9.692
−1.3	5.271	5.090	6.141
−1.2	4.557	3.772	3.872
−1.1	3.968	2.830	2.430
−1.0	3.477	2.159	1.519
−0.9	3.066	1.682	0.9486
−0.8	2.721	1.344	0.5960
−0.7	2.428	1.108	0.3816
−0.6	2.178	0.9435	0.2540
−0.5	1.964	0.8318	0.1804
−0.4	1.780	0.7580	0.1398
−0.3	1.620	0.7112	0.1187
−0.2	1.481	0.6840	0.1089
−0.1	1.360	0.6705	0.1052
0	1.253	0.6667	0.1044
0.1	1.159	0.6696	0.1048
0.2	1.076	0.6771	0.1052
0.3	1.002	0.6877	0.1052
0.4	0.9357	0.7000	0.1045
0.5	0.8764	0.7135	0.1031
0.6	0.8230	0.7274	0.1011
0.7	0.7749	0.7414	0.0985
0.8	0.7313	0.7552	0.0954
0.9	0.6917	0.7686	0.0920
1.0	0.6557	0.7814	0.0883
1.1	0.6227	0.7937	0.0845
1.2	0.5926	0.8053	0.0806
1.3	0.5649	0.8163	0.0767
1.4	0.5394	0.8267	0.0729
1.5	0.5158	0.8364	0.0691
1.6	0.4940	0.8455	0.0654
1.7	0.4739	0.8540	0.0619
1.8	0.4551	0.8619	0.0585
1.9	0.4376	0.8694	0.0552
2.0	0.4214	0.8763	0.0521
2.1	0.4062	0.8828	0.0492
2.2	0.3919	0.8889	0.0464
2.3	0.3786	0.8946	0.0438
2.4	0.3661	0.8999	0.0413
2.5	0.3543	0.9049	0.0390
2.6	0.3432	0.9095	0.0368
2.7	0.3327	0.9139	0.0347
2.8	0.3228	0.9179	0.0328
2.9	0.3134	0.9218	0.0309
3.0	0.3046	0.9254	0.0292
3.1	0.2962	0.9287	0.0276
3.2	0.2882	0.9319	0.0261
3.3	0.2806	0.9349	0.0247
3.4	0.2734	0.9377	0.0234
3.5	0.2666	0.9403	0.0222
3.6	0.2600	0.9428	0.0210
3.7	0.2538	0.9451	0.0199
3.8	0.2478	0.9473	0.0189
3.9	0.2421	0.9494	0.0179
4.0	0.2367	0.9514	0.0170

Calculate $c = \dfrac{a - x}{\sqrt{x}}$, then

$$B(x, a)^{-1} \simeq a_0\sqrt{x} + a_1 + a_2/\sqrt{x},$$

$$\frac{\partial B(x, a)}{\partial x} \simeq -\frac{B(x, a)^2}{2\sqrt{x}}\left(a_0 - \frac{a_2}{x}\right) - \frac{x + a}{2x} B(x, a)\left\{\frac{x}{a} - 1 + B(x, a)\right\}.$$

* Standardized offered load.

It is appropriate to mention, at this point, another method of approximating $B(x, a)$ by means of a formula whose computational effort is also independent of x and which, similarly, is applicable over a wide

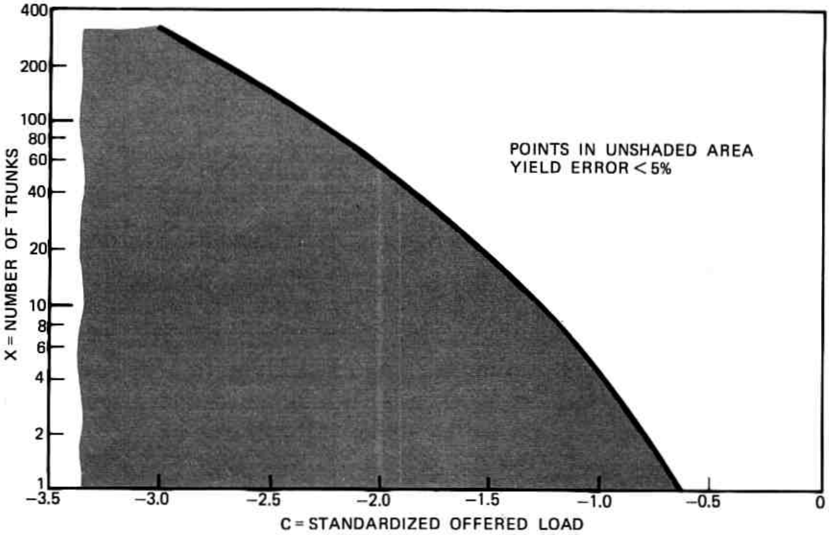


Fig. 1—Five-percent-error contour.

range of values of $B(x, a)$. This method is described in Refs. 10 and 11. A comparison of this method with that of Theorems 1 and 2 is given at length in a report by S. Miller.[9]

Derivatives and inequalities on derivatives are given in Part IV. Theorem 15 extends the well-known derivative formula for $B(x, a)$ with respect to the real variable a. Theorem 17 provides an accurate approximation for $\partial B(x, a)/\partial x$. Empirically, the accuracy seems to hold to four significant figures or better over a very wide range of values of x and a. Of significance is the corollary which shows that the approximate value obtained is always too small. If a quick appraisal of the derivative is desired, Theorem 18 may be used. The logarithmic convexity properties of $B(x, a)$ given in Theorem 19 provide the useful bounds of the corollary on the second derivatives. Also an application is given to the logarithmic interpolation of Theorem 20. This is very useful when, for example, one wishes to compute $B(x, a)$ for x between consecutive integers, say N, $N + 1$, and for which $B(N, a)$, $B(N + 1, a)$ are known. An extension of this idea is provided by Theorem 21 which permits accurate computation of $B(x, a)$.

It may be remarked that generally relations, representations, and asymptotics for $B(z, \alpha)^{-1}$ are simpler in structure than those for $B(z, \alpha)$ and may provide greater numerical accuracy in computations.

II. CONVERGENT REPRESENTATIONS

The study of telephone trunking problems, whether in equilibrium or transient condition, or even nonstationary,[12] engenders the Erlang loss function, $B(N, a)$, which initially arises in the form[13]

$$B(N, a) = \frac{a^N}{N!} \Big/ \sum_{j=0}^{N} \frac{a^j}{j!}, \qquad N \geqq 0 \text{ (integral)}, \quad a > 0. \tag{7}$$

For these reasons and for the purposes of studying certain forms arising in queuing theory related to $B(N, a)$ and also for the facilitation of numerical evaluation, it is useful to represent the loss function in diverse ways.

The numerical computation of $B(N, a)$ as given in (7) is awkward when a and N are large since then both numerator and denominator are large. A form well adapted to numerical work is

$$\begin{gathered} B(N, a)^{-1} = \sum_{j=0}^{N} N^{(j)} a^{-j}, \\ N^{(0)} = 1, \qquad N^{(j)} = N(N-1)\cdots(N-j+1) \quad (j > 0), \end{gathered} \tag{8}$$

which follows from

$$B(N, a)^{-1} = \sum_{j=0}^{N} \frac{N!}{j!} a^{j-N} = \sum_{j=0}^{N} \frac{N!}{(N-j)!} a^{-j} = \sum_{j=0}^{N} N^{(j)} a^{-j}. \tag{9}$$

A modified form of (8) is given in Theorem 1.

Theorem 1:

$$B(N, a)^{-1} = \sum_{j=0}^{\nu-1} N^{(j)} a^{-j} + N^{(\nu)} a^{-\nu} B(N-\nu, a)^{-1}, \qquad \nu \geqq 0.$$

Proof: Since

$$N^{(j+\nu)} = N^{(\nu)} (N-\nu)^{(j)}, \tag{10}$$

one has, from

$$B(N, a)^{-1} = \sum_{j=0}^{\nu-1} N^{(j)} a^{-j} + \sum_{j=\nu}^{N} N^{(j)} a^{-j}, \tag{11}$$

$$\begin{aligned} \sum_{j=\nu}^{N} N^{(j)} a^{-j} &= \sum_{j=0}^{N-\nu} N^{(j+\nu)} a^{-j-\nu} = N^{(\nu)} a^{-\nu} \sum_{j=0}^{N-\nu} (N-\nu)^{(j)} a^{-j} \\ &= N^{(\nu)} a^{-\nu} B(N-\nu, a)^{-1}. \end{aligned} \tag{12}$$

The formula of the theorem follows from (11) and (12).

Corollary: The case $\nu = 1$ implies the known[14] *difference equation*

$$B(N, a) = \frac{1}{1 + \dfrac{N}{aB(N - 1, a)}}.$$

R. Franks suggested using the value of $\tilde{B}_\nu(N, a)$ defined by

$$\tilde{B}_\nu(N, a) = 1 \Big/ \sum_{j=0}^{\nu-1} N^{(j)} a^{-j} \tag{13}$$

to approximate $B(N, a)$ in which, for any small number $\eta > 0$, the index ν is chosen so that

$$N^{(\nu)} a^{-\nu} \leqq \eta. \tag{14}$$

Theorem 2 bounds the error of the method.

Theorem 2:

$$\tilde{B}_\nu(N, a)(1 - \eta) \leqq B(N, a) \leqq \tilde{B}_\nu(N, a).$$

Proof: From Theorem 1 one has

$$\frac{1}{\tilde{B}_\nu(N, a)^{-1} + N^{(\nu)} a^{-\nu} B(N - \nu, a)^{-1}} = B(N, a) \leqq \tilde{B}_\nu(N, a). \tag{15}$$

Thus

$$\frac{B(N, a)}{\tilde{B}_\nu(N, a)} = \frac{1}{1 + N^{(\nu)} a^{-\nu} \tilde{B}_\nu(N, a) B(N - \nu, a)^{-1}}. \tag{16}$$

Since $N^{(\nu)} a^{-\nu}$ is strictly monotone increasing as a function of N, (8) shows that

$$B(N + 1, a) < B(N, a) \tag{17}$$

for all $N \geqq 0$; thus

$$\frac{B(N, a)}{\tilde{B}_\nu(N, a)} \geqq \frac{1}{1 + \eta \tilde{B}_\nu(N, a) B(N, a)^{-1}}, \tag{18}$$

and hence

$$\frac{B(N, a)}{\tilde{B}_\nu(N, a)} \geqq 1 - \eta. \tag{19}$$

The theorem follows from (15) and (19).

An integral representation, ascribed to Fortet,[15] may be obtained for $B(N, a)$.

Theorem 3:

$$B(N, a)^{-1} = a \int_0^\infty e^{-ay} (1 + y)^N dy.$$

Proof: From the Eulerian integral

$$\int_0^\infty e^{-ay}y^l dy = \Gamma(l+1)a^{-l-1}, \qquad l > -1, \tag{20}$$

one obtains

$$N^{(j)}a^{-j} = a\binom{N}{j}\int_0^\infty e^{-ay}y^j dy. \tag{21}$$

Use of (8) now yields

$$B(N, a)^{-1} = a\int_0^\infty e^{-ay}\sum_{j=0}^{N}\binom{N}{j}y^j dy = a\int_0^\infty e^{-ay}(1+y)^N dy. \tag{22}$$

The integral representation now permits extending $B(N, a)$ into the complex plane with respect to both N and a. One defines

$$B(z, \alpha)^{-1} = \alpha\int_0^\infty e^{-\alpha y}(1+y)^z dy \tag{23}$$

in which z, α may both be complex. Clearly, $B(z, \alpha)^{-1}$ is an entire function of z for $\text{Re}\,\alpha > 0$ (Re designates "real part"). The symbols N, a will be used for nonnegative integers and positive reals, respectively.

A generalization of Theorem 1 is given in Theorem 4.

Theorem 4:

$$B(z, \alpha)^{-1} = \sum_{j=0}^{\nu-1} z^{(j)}\alpha^{-j} + z^{(\nu)}\alpha^{-\nu}B(z-\nu, \alpha)^{-1}, \qquad \text{Re}\,\alpha > 0.$$

Proof: Integration by parts of (23).

It is of interest to investigate the relationship of $B(z, \alpha)$ to the function

$$\psi(z, \alpha) = e^{-\alpha}\frac{\alpha^z}{\Gamma(z+1)}, \tag{24}$$

which is an extension of the Poisson distribution function, $\psi(N, a)$, with parameter a. The function $\psi(N, a)$ is a good approximation to $B(N, a)$ when a is much less than N. Exact relations between $B(z, \alpha)$ and $\psi(z, \alpha)$ are given in Theorems 5 and 6. These relations provide convenient means of calculation of $B(z, \alpha)$ for general z, α; e.g., in trunk group blocking problems when a nonintegral number of trunks is considered.

Theorem 5:

$$B(z, \alpha)^{-1} = \psi(z, \alpha)^{-1} - \sum_{s=1}^{\infty}\frac{\alpha^S}{(z+1)\cdots(z+s)}.$$

The series converges uniformly everywhere in Re $z > -1$, Re $\alpha > 0$.

Proof: Let $u = 1 + y$ in (23), then

$$B(z, \alpha)^{-1} = \alpha e^{\alpha} \int_1^{\infty} e^{-\alpha u} u^z du; \tag{25}$$

hence,

$$B(z, \alpha)^{-1} = \alpha e^{\alpha} \int_0^{\infty} e^{-\alpha u} u^z du - \alpha \int_0^1 e^{\alpha(1-u)} u^z du, \tag{26}$$

and

$$B(z, \alpha)^{-1} = \psi(z, \alpha)^{-1} - \alpha \int_0^1 e^{\alpha(1-u)} u^z du. \tag{27}$$

To exhibit the integral in (27) as an inverse factorial series, consider the beta function integral

$$\int_0^1 u^{x-1}(1 - u)^{y-1} du = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} \tag{28}$$

which yields the special case ($s \geqq 0$ integral)

$$\int_0^1 u^z (1 - u)^s du = \frac{s\,!}{(z + 1) \cdots (z + 1 + s)}. \tag{29}$$

Use of the expansion

$$e^{\alpha(1-u)} = \sum_{s=0}^{\infty} \frac{\alpha^s}{s\,!} (1 - u)^s \tag{30}$$

in (27) and subsequent use of (29) yield the result of the theorem.

The Mittag-Leffler expansion for the integral of (27) leads to

Theorem 6:

$$B(z, \alpha)^{-1} = \psi(z, \alpha)^{-1} + e^{\alpha} \sum_{s=1}^{\infty} (-1)^s \frac{\alpha^s}{(s - 1)\,!\,(s + z)}.$$

Conditions of convergence are the same as in Theorem 5.

Proof: The expansion

$$e^{-\alpha u} = \sum_{s=0}^{\infty} (-1)^s \frac{\alpha^s u^s}{s\,!} \tag{31}$$

used in (27) leads immediately to the required result.

Whittaker functions,[16] $W_{k,m}(z)$, play a useful role in the discussion of $B(z, \alpha)$ and of Poisson-Charlier polynomials to be introduced later.

They may be introduced by

$$W_{k,m}(z) = \frac{e^{-\frac{1}{2}z}z^{k}}{\Gamma(\frac{1}{2} - k + m)} \int_0^\infty e^{-t}\left(1 + \frac{t}{z}\right)^{k-\frac{1}{2}+m} t^{-k-\frac{1}{2}+m}dt. \tag{32}$$

$$\operatorname{Re}\,(\tfrac{1}{2} - k + m) > 0, \qquad |\arg z| < \pi.$$

Theorem 7:

$$B(z, \alpha)^{-1} = e^{-z/2}e^{\frac{1}{2}\alpha}W_{z/2,(z+1)/2}(\alpha).$$

Proof: Let $t = \alpha y$ in (23), then

$$B(z, \alpha)^{-1} = \int_0^\infty e^{-t}\left(1 + \frac{t}{\alpha}\right)^{z} dt. \tag{33}$$

The required result follows on comparison with (32).

A Rodrigues type of relation for $B(N, a)^{-1}$ may be obtained from Theorem 3.

Theorem 8:

$$B(N + M, a)^{-1} = (-1)^{M}ae^{a}\frac{d^{M}}{da^{M}}[e^{-a}a^{-1}B(N, a)^{-1}].$$

Proof: From Theorem 3, one has

$$e^{-a}a^{-1}B(N, a)^{-1} = \int_0^\infty e^{-a(1+y)}(1 + y)^{N}dy; \tag{34}$$

hence,

$$(-1)^{M}\frac{d^{M}}{da^{M}}[e^{-a}a^{-1}B(N, a)^{-1}] = \int_0^\infty e^{-a(1+y)}(1 + y)^{N+M}dy. \tag{35}$$

The formula for $B(N + M, a)^{-1}$ now follows on multiplication by ae^{a}.

Corollary:

$$B(N, a)^{-1} = (-1)^{N}ae^{a}\frac{d^{N}}{da^{N}}[e^{-a}a^{-1}].$$

Additional formulas for $B(z, \alpha)^{-1}$ (as a function of α) provide convenient means of computation for values of α near some fixed point. Two such formulas are given in Theorems 9 and 10.

Theorem 9:

$$B(n, \alpha + t)^{-1} = \left(1 + \frac{t}{\alpha}\right)^{-n} \sum_{\nu=0}^{n} \binom{n}{\nu} B(n - \nu, \alpha)^{-1}\left(\frac{t}{\alpha}\right)^{\nu}.$$

Proof: The function $S_n(\alpha)$ given by

$$S_n(\alpha) = \sum_{\tau=0}^{n} \frac{\alpha^\tau}{\tau!} \tag{36}$$

is an Appell polynomial, that is,

$$\frac{dS_n(\alpha)}{d\alpha} = S_{n-1}(\alpha). \tag{37}$$

Thus the Taylor expansion for $S_n(\alpha + t)$ can be written in the form

$$S_n(\alpha + t) = \sum_{\nu=0}^{n} \frac{1}{\nu!} S_{n-\nu}(\alpha)t^\nu. \tag{38}$$

One obtains from (7)

$$B(n,\alpha)^{-1} = n!\,\alpha^{-n}S_n(\alpha) \tag{39}$$

and hence the theorem follows on substitution into (38).

Theorem 10:

$$B(z, \alpha + t)^{-1} = \left(1 + \frac{t}{\alpha}\right) e^t \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} B(z + \nu, \alpha)^{-1}t^\nu,$$
$$\operatorname{Re}\alpha > 0, \qquad |t| < \operatorname{Re}\alpha.$$

Proof: Let

$$l(z, \alpha) = e^{iz\pi-\alpha}\alpha^{-1}B(z, \alpha)^{-1}, \tag{40}$$

then, from (23),

$$\begin{aligned} \frac{d}{d\alpha} l(z, \alpha) &= \frac{d}{d\alpha} e^{iz\pi} \int_0^\infty e^{-\alpha(1+y)}(1 + y)^z dy, \\ &= -\, e^{iz\pi} \int_0^\infty e^{-\alpha(1+y)}(1 + y)^{z+1} dy, \\ &= l(z + 1, \alpha), \end{aligned} \tag{41}$$

and hence

$$\frac{d^\nu}{d\alpha^\nu} l(z, \alpha) = l(z + \nu, \alpha). \tag{42}$$

Thus, by Taylor's formula,

$$l(z, \alpha + t) = \sum_{\nu=0}^{\infty} \frac{t^\nu}{\nu!} l(z + \nu, \alpha). \tag{43}$$

Substitution of (40) into (43) yields the required result. One has

$$l(z + \nu, \alpha) = e^{i(z+\nu)\pi} \int_0^\infty e^{-\alpha(1+y)}(1 + y)^{z+\nu} dy, \tag{44}$$

hence the terms of (43) are $0[(t/\mathrm{Re}\,\alpha)^{\nu}\nu^{\mathrm{Re}\,z}]$. The stated convergence criterion now follows.

III. ASYMPTOTIC EXPANSIONS

Particularly simple and convenient forms for theoretical and numerical applications may be obtained by examining asymptotic expansions.

Theorem 11:

$$B(z, \alpha)^{-1} \sim \sum_{\nu=0}^{\infty} z^{(\nu)}\alpha^{-\nu}, \qquad \alpha \to \infty, \quad |\arg \alpha| < \pi.$$

Proof: The asymptotic expansion for $W_{k,m}(z)$ is[16]

$$W_{k,m}(z) \sim e^{-\frac{1}{2}z}z^{k} \times \left\{1 + \sum_{\nu=1}^{\infty} \frac{[m^2-(k-\frac{1}{2})^2][m^2-(k-\frac{3}{2})^2]\cdots[m^2-(k-\nu+\frac{1}{2})^2]}{\nu!\,z^{\nu}}\right\}, \qquad z \to \infty, \quad |\arg z| < \pi. \tag{45}$$

Substitution of the parameter values given by Theorem 7 establishes the result.

It should be remarked that the error, when using the partial sum $\sum_{\nu=0}^{k-1} z^{(\nu)}\alpha^{-\nu}$ to approximate $B(z, \alpha)^{-1}$, does not exceed $|\alpha|\,|z^{(k)}\alpha^{-k}|/\mathrm{Re}\,\alpha$ provided $\mathrm{Re}\,z \leqq k$, $\mathrm{Re}\,\alpha > 0$. This follows directly from Theorem 4 and (23).

For large z, one has

Theorem 12:

$$B(z, \alpha)^{-1} \sim \psi(z, \alpha)^{-1} - \sum_{s=1}^{\infty} \frac{\alpha^{s}}{(z+1)\cdots(z+s)},$$

$z \to \infty$, $|\arg z| < \pi/2$, *uniformly in any bounded region of the* α*-plane for which* $\mathrm{Re}\,\alpha > 0$.

Proof: The representation of Theorem 5 is used. One must show

$$B(z, \alpha)^{-1} - \psi(z, \alpha)^{-1} - \sum_{s=1}^{n} \frac{\alpha^{s}}{(z+1)\cdots(z+s)} = o\left(\frac{\alpha^{n}}{(z+1)\cdots(z+n)}\right); \tag{46}$$

that is,

$$\lim_{z\to\infty} \sum_{s>n} \frac{\alpha^{s-n}}{(z+n+1)\cdots(z+s)} = 0. \tag{47}$$

Let Re $z = x$, then one has

$$\left| \sum_{s>n} \frac{\alpha^{s-n}}{(z+n+1)\cdots(z+s)} \right| \\ \leqq \sum_{s>n} |\alpha|^{s-n} \frac{1}{(x+n+1)\cdots(x+s)}. \qquad (48)$$

Let $v = x + n$ and $l = s - n$, then the dexter of (48) is

$$\sum_{l=1}^{\infty} \frac{|\alpha|^l}{(v+1)\cdots(v+l)}. \qquad (49)$$

Use of (29) and (30) on (49) yields

$$\left| \sum_{s>n} \frac{\alpha^{s-n}}{(z+n+1)\cdots(z+s)} \right| \leqq |\alpha| \int_0^1 e^{|\alpha|(1-u)} u^v du; \qquad (50)$$

thus

$$\left| \sum_{s>n} \frac{\alpha^{s-n}}{(z+n+1)\cdots(z+s)} \right| \leqq \frac{|\alpha| e^{|\alpha|}}{v+1} \to 0, \qquad v \to \infty. \qquad (51)$$

The theorem is proved.

Useful asymptotic formulas are obtained when both α and z have infinite limits but approach infinity in a fixed ratio, that is, $\alpha = cz$, c fixed. The cases $c > 1$, $c = 1$ are discussed by A. Descloux[17] for large real z. Theorem 13 generalizes the result for $c > 1$ to complex z and provides the structure of the coefficients for the complete expansion. The case $c = 1$ is obtained as a corollary to Theorem 14 where the result is also generalized to complex z.

Theorem 13:

$$B(z, cz)^{-1} \sim \sum_{l=0}^{\infty} g_l z^{-l},$$

$$z \to \infty, \quad |\arg z| < \frac{\pi}{2}, \quad c > 1,$$

$$g_l = \left(\frac{c}{c-1} \frac{d}{dc} \right)^l \frac{c}{c-1}.$$

Proof:* One has, from (23)

$$B(z, cz)^{-1} = cz \int_0^{\infty} e^{-czy} (1+y)^z dy. \qquad (52)$$

* The author wishes to thank C. L. Mallows for this proof, which replaces a much longer proof originally supplied by the author.

Defining the function $h(y)$ by

$$h(y) = cy - \ln(1+y), \tag{53}$$

one may write, since $h(0) = 0$, $h(\infty) = \infty$, and $h(y)$ is monotonic increasing,

$$B(z, cz)^{-1} = z \int_0^\infty e^{-zh(y)} \frac{c(1+y)}{c(1+y)-1}\, dh. \tag{54}$$

The factor $c(1+y)/[c(1+y)-1]$ is now expanded in powers of h as follows:

$$\frac{c(1+y)}{c(1+y)-1} = \sum_{l=0}^{\infty} \frac{h^l}{l!} g_l. \tag{55}$$

A theorem on Abelian asymptotics for Laplace transforms[18] and (54), (55) yield the asymptotic behavior of $B(z, cz)^{-1}$ for $z \to \infty$, $|\arg z| < \pi/2$; thus,

$$B(z, cz)^{-1} \sim \sum_{l=0}^{\infty} g_l z^{-l}. \tag{56}$$

The coefficients g_l may be evaluated as follows. Let

$$w = c - \ln c, \tag{57}$$

and

$$k(w) = \frac{c}{c-1}, \tag{58}$$

then

$$k(w+h) = k[c(1+y) - \ln c(1+y)] = \frac{c(1+y)}{c(1+y)-1}. \tag{59}$$

Thus, Taylor expansion yields

$$\frac{c(1+y)}{c(1+y)-1} = \sum_{l=0}^{\infty} \frac{h^l}{l!} \left(\frac{d}{dw}\right)^l k(w). \tag{60}$$

One has

$$\frac{d}{dw} = \frac{c}{c-1} \frac{d}{dc}, \tag{61}$$

hence

$$\left(\frac{d}{dw}\right)^l k(w) = \left(\frac{c}{c-1}\frac{d}{dc}\right)^l \frac{c}{c-1} = g_l. \tag{62}$$

The following formula is obtained directly from Theorem 13.

$$B(z, cz)^{-1} \sim \frac{c}{c-1} - \frac{c}{(c-1)^3}\frac{1}{z} + \frac{2c^2 + c}{(c-1)^5}\frac{1}{z^2}. \tag{63}$$

The evaluation and behavior of $B(z, \alpha)$ for α in a neighborhood of z

is often of interest; accordingly, the function $B(z, z + c\sqrt{z})^{-1}$ will be considered for $z \rightarrow \infty$; c is a fixed real number.

Theorem 14: There exists a representation of the form

$$B(z, z + c\sqrt{z}) \sim \sum_{j=0}^{\infty} a_j(c) z^{-(j-1)/2},$$

$$z \rightarrow \infty, \quad |\arg z| < \frac{\pi}{2}, \quad c \text{ real},$$

in which

$$\begin{aligned}
a_0(c) &= e^{\frac{1}{2}c^2} \int_c^{\infty} e^{-\frac{1}{2}u^2} du, \\
a_1(c) &= \frac{2}{3} + \frac{1}{3} c^2 - \frac{1}{3} c^3 a_0(c), \\
a_2(c) &= -\frac{1}{18} c^5 - \frac{7}{36} c^3 + \frac{1}{12} c + \left(\frac{1}{18} c^6 + \frac{1}{4} c^4 + \frac{1}{12} \right) a_0(c).
\end{aligned}$$

Proof: From (23), one has

$$B(z, z + c\sqrt{z})^{-1} = (z + c\sqrt{z}) \int_0^{\infty} e^{-(z+c\sqrt{z})u} (1 + u)^z du, \tag{64}$$

$$|\arg z| < \frac{\pi}{2}.$$

Let $u = v/\sqrt{z}$, then

$$\begin{aligned}
B(z, z + c\sqrt{z})^{-1} &= \int_0^{\infty} e^{-(\frac{1}{2}v^2 + cv)} h(v, z) dv, \\
h(v, z) &= e^{\frac{1}{2}v^2 - \sqrt{z}v} \left(1 + \frac{v}{\sqrt{z}} \right)^z (\sqrt{z} + c).
\end{aligned} \tag{65}$$

Let K be a positive constant, then, for $|v| \leqq K$, $h(v, z)$ clearly possesses an asymptotic development in $\sqrt{z}$ uniformly in v; thus,

$$h(v, z) \sim \sum_{j=0}^{\infty} b_j(v, c) z^{-(j-1)/2}, \qquad z \rightarrow \infty, \tag{66}$$

in which the coefficients $b_j(v, c)$ are polynomials in v. In particular,

$$\begin{aligned}
b_0(v, c) &= 1, \\
b_1(v, c) &= \frac{1}{3} v^3 + c, \\
b_2(v, c) &= \frac{1}{3} cv^3 - \frac{1}{4} v^4 + \frac{1}{18} v^6.
\end{aligned} \tag{67}$$

Since

$$e^{-(\frac{1}{2}v^2+cv)}v^k\epsilon L(0,\ \infty) \tag{68}$$

for each $k > 0$ and any c, termwise integration of (66) leads to the required asymptotic expansion. Thus, letting

$$a_j(c) = \int_0^\infty e^{-(\frac{1}{2}v^2+cv)}b_j(v,\ c)dv, \tag{69}$$

one has

$$B(z,\ z + c\sqrt{z})^{-1} \sim \sum_{j=0}^{\infty} a_j(c)z^{-(j-1)/2}. \tag{70}$$

The formulas for $a_0(c)$, $a_1(c)$, $a_2(c)$ stated in the theorem are obtained by evaluation of (69) using $b_j(v,\ c)$ as given in (67).

Corollary:

$$B(z,\ z)^{-1} \sim \sqrt{\frac{\pi z}{2}} + \frac{2}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2z}},$$
$$z \to \infty, \qquad |\arg z| < \frac{\pi}{2}.$$

Proof: The result is obtained from Theorem 14 with $c = 0$.

This theorem helps explain the phenomenon of the efficiency of large trunk groups since even when $a > x$, $B(x,\ a)$ is small as long as a is in a small neighborhood of x; thus, Theorem 14 shows that $B(x,\ x + c\sqrt{x}) \sim 1/a_0\sqrt{x}$, $x \to \infty$.

Theorem 14 shows that the parameter c may be viewed as a standardized offered load measuring the deviation of a from x in units of $\sqrt{x}$. The value of this viewpoint is derived from the very simple approximating form for $B(x,\ a)$; thus,

$$B(x,\ a)^{-1} \simeq a_0\sqrt{x} + a_1 + \frac{a_2}{\sqrt{x}}. \tag{71}$$

An application of this is to the computation of $\partial B(x,\ a)/\partial x$ given in (92). Another advantage is the capability of computing $B(x,\ a)$ by means of a single-entry table against the standardized offered load c rather than the usual double-entry table against x and a.

Table I gives the values of $a_0(c)$, $a_1(c)$, $a_2(c)$ for $-3 \leqq c \leqq 4$ in steps of 0.1 with the intention of covering a practical range of values of $B(x,\ a)$. As an illustration, it is desired to compute $B(400,\ 378)$. Use of (71) with $c = -1.1$ gives the result 0.0122 correct to the last figure. If c does not appear in the table, then interpolation is used. For example, to compute $B(400,\ 377.6)$ for which $c = -1.12$ linear interpolation in the table of coefficients and use of (71) yields 0.0118 correct to the last figure. The method, of course, is valid even when the number of trunks is nonintegral. Consider, for example, $B(400.34, 420)$ for which $c = 0.98463$. The result obtained by linear interpolation in the table is 0.0713 correct to half a unit of the last figure.

The accuracy deteriorates when x is decreased or when c is large and negative. Thus, for $B(10, 8)$, one obtains 0.12144 as against the correct value 0.12166. In this case $c = -0.6325$ is not too disadvantageous. The case $B(10, 5)$ for which $c = -1.58114$ yields a much greater error, namely, 0.0256 as against the correct value 0.0184. This error occurs, however, for a small trunk group where exact calculation is quite feasible. To aid the delineation of suitable regions of (c, x) for which the table is accurate, a curve is given in Fig. 1 defining 5-percent error. When a computation is made from the table using any point (c, x) in the unbounded, unshaded region, the error incurred will be less than 5 percent of the true value of $B(x, a)$.

IV. DERIVATIVES AND INEQUALITIES

It is desired to obtain formulas for the derivatives of $B(z, \alpha)$, with respect to z and α.

Theorem 15:

$$\frac{\partial B(z, \alpha)}{\partial \alpha} = \left\{ \frac{z}{\alpha} - 1 + B(z, \alpha) \right\} B(z, \alpha), \qquad \operatorname{Re} \alpha > 0.$$

Proof: From (23), one has

$$\frac{\partial B(z, \alpha)^{-1}}{\partial \alpha} = \int_0^\infty e^{-\alpha u}(1 + u)^z du - \alpha \int_0^\infty e^{-\alpha u}(1 + u)^z u du; \tag{72}$$

hence,

$$\frac{\partial B(z, \alpha)^{-1}}{\partial \alpha} = \frac{1}{\alpha} B(z, \alpha)^{-1} - B(z + 1, \alpha)^{-1} + B(z, \alpha)^{-1}. \tag{73}$$

Use of Theorem 4 provides the relation

$$\frac{\partial B(z, \alpha)^{-1}}{\partial \alpha} = -\frac{z}{\alpha} B(z, \alpha)^{-1} - 1 + B(z, \alpha)^{-1}. \tag{74}$$

Since

$$\frac{\partial B(z, \alpha)^{-1}}{\partial \alpha} = -B(z, \alpha)^{-2} \frac{\partial B(z, \alpha)}{\partial \alpha}, \tag{75}$$

the result of the theorem follows from (74).

For the purpose of obtaining an approximate formula for the derivative with respect to z, consider

$$f(u) = aB(x, a)e^{-au}(1 + u)^x \tag{76}$$

in which $a > 0$, $x > 0$, and for which, by (23),

$$\int_0^\infty f(u)du = 1. \tag{77}$$

It is convenient to introduce the random variable ξ with density function $f(u)$. The power moments μ_r defined by

$$\mu_r = E\xi^r, \qquad r > 0 \text{ (integral)}, \tag{78}$$

are given in the following theorem.

Theorem 16:

$$\mu_r = B(x, a) \sum_{l=0}^{r} (-1)^{r-l} \binom{r}{l} B(x + l, a)^{-1}.$$

Proof: Define a generating function $\phi(t)$ by

$$\phi(t) = Ee^{t\xi} = aB(x, a) \int_0^\infty e^{-(a-t)u}(1 + u)^x du, \tag{79}$$

then, since

$$(a - t)B(x, a - t) \int_0^\infty e^{-(a-t)u}(1 + u)^x du = 1, \tag{80}$$

one has

$$\phi(t) = \frac{aB(x, a)}{(a - t)B(x, a - t)}. \tag{81}$$

Use of Theorem 10 in (81) provides the expansion

$$\phi(t) = B(x, a)e^{-t} \sum_{r=0}^{\infty} \frac{B(x + r, a)^{-1}}{r!} t^r. \tag{82}$$

Since

$$\phi(t) = \sum_{r=0}^{\infty} \frac{\mu_r}{r!} t^r, \tag{83}$$

the coefficient of t^r in the expansion of (82) in powers of t yields the required result. Thus

$$\mu_r = B(x, a)r! \sum_{l=0}^{r} \frac{(-1)^{r-l}}{(r - l)!} \frac{B(x + l, a)^{-1}}{l!} \tag{84}$$

and the formula of the theorem follows.

Corollary: *The central moments α_r are given by*

$$\alpha_r = E(\xi - \mu_1)^r = B(x, a) \sum_{l=0}^{r} (-1)^{r-l} \binom{r}{l} (\mu_1 + 1)^{r-l} B(x + l, a)^{-1}.$$

Proof: The same as for Theorem 16 but considering the function $e^{-\mu_1 t}\phi(t)$ instead of $\phi(t)$.

An approximation to $\partial B(x, a)/\partial x$ may now be obtained.

Theorem 17:

$$-B(x, a)^{-1} \frac{\partial B(x, a)}{\partial x} = \ln (1 + \mu_1) - \frac{1}{2} \frac{\alpha_2}{(1 + \mu_1)^2} + \frac{1}{3} \frac{\alpha_3}{(1 + \mu_1)^3} - \frac{1}{4} \alpha_4 \theta, \qquad 0 < \theta < 1.$$

Proof: From (23) and (76), one obtains

$$-B(x, a)^{-1} \frac{\partial B(x, a)}{\partial x} = E \ln (1 + \xi). \tag{85}$$

Since, by use of the mean value formula,

$$\ln (1 + \xi) = \ln (1 + \mu_1) + \frac{1}{1 + \mu_1} (\xi - \mu_1) - \frac{1}{2} \frac{1}{(1 + \mu_1)^2} (\xi - \mu_1)^2 + \frac{1}{3} \frac{1}{(1 + \mu_1)^3} (\xi - \mu_1)^3 - \frac{1}{4} \theta (\xi - \mu_1)^4, \qquad 0 < \theta < 1, \tag{86}$$

one has, from (85) and the corollary to Theorem 16, the required result.

Corollary:

$$-B(x, a)^{-1} \frac{\partial B(x, a)}{\partial x} < \ln (1 + \mu_1) - \frac{1}{2} \frac{\alpha_2}{(1 + \mu_1)^2} + \frac{1}{3} \frac{\alpha_3}{(1 + \mu_1)^3}.$$

Proof: The error term of Theorem 17 is omitted.

For ready reference the following formulas are given in which $B = B(x, a)$, $B_1 = B(x + 1, a)$, $B_2 = B(x + 2, a)$, $B_3 = B(x + 3, a)$, $B_4 = B(x + 4, a)$.

$$\begin{aligned}
\mu_1 &= -1 + BB_1^{-1}, \\
\alpha_2 &= (\mu_1 + 1)^2 - 2(\mu_1 + 1)BB_1^{-1} + BB_2^{-1}, \\
\alpha_3 &= -(\mu_1 + 1)^3 + 3(\mu_1 + 1)^2BB_1^{-1} - 3(\mu_1 + 1)BB_2^{-1} + BB_3^{-1}, \\
\alpha_4 &= (\mu_1 + 1)^4 - 4(\mu_1 + 1)^3BB_1^{-1} + 6(\mu_1 + 1)^2BB_2^{-1} \\
&\qquad - 4(\mu_1 + 1)BB_3^{-1} + BB_4^{-1}.
\end{aligned} \tag{87}$$

The evaluation of B_1^{-1}, B_2^{-1}, B_3^{-1}, B_4^{-1} is facilitated by successive use of Theorem 4.

An alternative method of obtaining $\partial B(x, a)/\partial x$ is based on Theorem 14. Let

$$f(x, c) = B(x, a), \qquad a = x + c\sqrt{x}, \tag{88}$$

then, from Theorem 14,

$$\frac{\partial f(x, c)^{-1}}{\partial x} \sim - \sum_{j=0}^{\infty} \frac{j-1}{2} a_j(c) x^{-(j+1)/2}, \qquad x \to \infty ; \tag{89}$$

hence,

$$\frac{\partial f(x, c)^{-1}}{\partial x} \simeq \frac{1}{2\sqrt{x}} \left\{ a_0(c) - \frac{a_2(c)}{x} \right\}. \tag{90}$$

Thus, the computation of $\partial f(x, c)/\partial x$ is easily accomplished with the help of Table I and the formula

$$\frac{\partial f(x, c)}{\partial x} = - f(x, c)^2 \frac{\partial f(x, c)^{-1}}{\partial x} \simeq - \frac{B(x, a)^2}{2\sqrt{x}} \left\{ a_0(c) - \frac{a_2(c)}{x} \right\}. \tag{91}$$

One now has

$$\frac{\partial B(x, a)}{\partial x} = \frac{\partial f(x, c)}{\partial x} - \frac{\partial B(x, a)}{\partial a} \left(1 + \frac{c}{2\sqrt{x}} \right). \tag{92}$$

A simple upper bound on $-B(x, a)^{-1}[\partial B(x, a)/\partial x]$ is given in the following theorem.

Theorem 18:

$$-B(x, a)^{-1} \frac{\partial B(x, a)}{\partial x} < \ln (1 + \mu_1).$$

Proof: Since the function $-\ln (1 + u)$ is convex for $u \geqq 0$, the required inequality follows from Jensen's inequality, namely,

$$g(E\xi) \leqq Eg(\xi) \tag{93}$$

valid for functions $g(x)$ convex over the range of the random variable ξ, and (85).

A function $g(x) > 0$ is said to be log-convex over a set if $\ln g(x)$ is convex over the set. It is known[19] that the sum of log-convex functions is log-convex and hence that the integral of a log-convex function with respect to a parameter is log-convex provided the function is log-convex for every value of the parameter. Since a necessary and sufficient condition that a twice-differentiable function be convex is the nonnegativity of its second derivative over the corresponding set, one derives the inequality

$$g''g - g'^2 \geqq 0 \tag{94}$$

as a necessary and sufficient condition that $g > 0$ be log-convex. One now has

Theorem 19: $B(x, a)^{-1}$, $[aB(x, a)]^{-1}$ *are log-convex functions of* x *and of* a, *respectively, for* $a > 0$ *and all* x.

Proof: The results are immediate from (23) and the observations that $(1 + u)^x$ is log-convex as a function of x for $u \geqq 0$, and e^{-au} is log-convex as a function of a for $u \geqq 0$.

Corollary:

$$B(x, a)\,\frac{\partial^2 B(x, a)}{\partial x^2} \leqq \left[\frac{\partial B(x, a)}{\partial x}\right]^2,$$

$$aB(x, a)\left[2\,\frac{\partial B(x, a)}{\partial a} + a\,\frac{\partial^2 B(x, a)}{\partial a^2}\right] \leqq \left[B(x, a) + a\,\frac{\partial B(x, a)}{\partial a}\right]^2.$$

Proof: Use of (94).

An immediate application of Theorem 19 is to the logarithmic interpolation of $B(x, a)$, that is, linear interpolation of $\ln B(x, a)$.

Theorem 20: *Let* $a, b, p, q > 0$, $p + q = 1$, *then*

$$B(x, a)^p B(y, a)^q \leqq B(px + qy, a),$$
$$[aB(x, a)]^p [bB(x, b)]^q \leqq (pa + qb) B(x, pa + qb).$$

Proof: Jensen's inequality applied to $-\ln B(x, a)$ and $-\ln [aB(x, a)]$, respectively.

An extension of the result of Theorem 20, for the purpose of obtaining an approximate formula for $B(x, a)$ when x is not an integer, may be derived from the corollary to Theorem 16. Let $N = [x]$, $\delta = x - N$, and α_r be the central moments computed for the density function

$$f(u) = aB(N, a)e^{-au}(1 + u)^N, \tag{95}$$

then one has

Theorem 21:

$$B(x, a)^{-1} = B(N, a)^{-1} \sum_{r=0}^{k-1} \binom{\delta}{r} \alpha_r (1 + \mu_1)^{\delta - r} + B(N, a)^{-1} \binom{\delta}{k} \alpha_k \theta,$$
$$k \text{ even}, \qquad |\theta| \leqq 1.$$

Proof: Let ξ be the random variable with density function $f(u)$, then

$$B(x, a)^{-1} = B(N, a)^{-1} E(1 + \xi)^\delta. \tag{96}$$

Since

$$(1+\xi)^{\delta} = \sum_{r=0}^{k-1} \binom{\delta}{r} (1+\mu_1)^{\delta-r}(\xi-\mu_1)^r + \binom{\delta}{k} (\xi-\mu_1)^k\theta, \quad (97)$$

the result follows from (96) and the corollary to Theorem 16.

A useful special case of Theorem 21 is

$$B(x, a) \simeq \frac{B^{1-\delta}B_1{}^{\delta}}{1-\dfrac{1}{2}\,\delta(1-\delta)\left(\dfrac{B_1^2}{BB_2}-1\right)} \quad (98)$$

in which

$$B = B(N, a), \qquad B_1 = B(N+1, a), \qquad B_2 = B(N+2, a). \quad (99)$$

V. CONCLUSION

Further investigations would be desirable; for example, one would like to know the contour function $g(z)$ for which $B[z, g(z)]$ is constant. Truncation error formulas for the asymptotic expansions of Theorems 13 and 14 would be useful; also, the general structure of the coefficients $a_j(c)$ of Theorem 14 should be determined. Asymptotic formulas of various types should be obtained for $G_j(z, \alpha)$ similar to those given for $B(z, \alpha)$. These formulas may then be used to obtain asymptotic results for its zeros which are needed in many transient and time-variable blocking analyses.

VI. ACKNOWLEDGMENTS

With pleasure I acknowledge the careful reading of the manuscript by R. Marzec and S. Horing and their valuable suggestions. I should also like to acknowledge the programming efforts of M. Zeitler who prepared Table I.

APPENDIX

The function $B(z, \alpha)$ is related to the Poisson-Charlier polynomials[20-22] much used in telephone traffic studies. Let

$$\begin{aligned} \psi_0(z, \alpha) &= \psi(z, \alpha), \\ \psi_j(z, \alpha) &= \frac{d^j}{d\alpha^j}\psi(z, \alpha), \end{aligned} \quad (100)$$

then the Poisson-Charlier polynomials, $G_j(z, \alpha)$, are defined by

$$\psi_j(z, \alpha) = \psi(z, \alpha)G_j(z, \alpha). \quad (101)$$

The Taylor expansion

$$\psi(z, \alpha + t) = \sum_{j=0}^{\infty} \frac{t^j}{j!} \psi_j(z, \alpha) \tag{102}$$

yields the generating function[23]

$$e^{-t}\left(1 + \frac{t}{\alpha}\right)^z = \sum_{j=0}^{\infty} G_j(z, \alpha) \frac{t^j}{j!}. \tag{103}$$

Thus, explicit formulas for $G_j(z, \alpha)$ are

$$\begin{aligned} G_j(z, \alpha) &= \frac{j!}{\alpha^j} \sum_{\nu=0}^{j} (-1)^\nu \binom{z}{j-\nu} \frac{\alpha^\nu}{\nu!} \\ &= \sum_{\nu=0}^{j} (-1)^{j-\nu} \binom{j}{\nu} \nu!\, \alpha^{-\nu} \binom{z}{\nu}. \end{aligned} \tag{104}$$

The first few polynomials are

$$\begin{aligned} G_0(z, \alpha) &= 1, \\ G_1(z, \alpha) &= \frac{1}{\alpha}(z - \alpha), \\ G_2(z, \alpha) &= \frac{1}{\alpha^2}[z^2 - (2\alpha + 1)z + \alpha^2], \\ G_3(z, \alpha) &= \frac{1}{\alpha^3}[z^3 - 3(\alpha + 1)z^2 + (3\alpha^2 + 3\alpha + 2)z - \alpha^3]. \end{aligned} \tag{105}$$

A recurrence relation derived from (103) is

$$G_{j+1}(z, \alpha) = \frac{z - j - \alpha}{\alpha} G_j(z, \alpha) - \frac{j}{\alpha} G_{j-1}(z, \alpha). \tag{106}$$

The polynomials, $G_j(z, \alpha)$, possess many properties analogous to those of $B(z, \alpha)^{-1}$. A Rodrigues formula is given in

Theorem 22:

$$G_{j+k}(z, \alpha) = \alpha^{-z} e^{\alpha} \frac{d^k}{d\alpha^k} [e^{-\alpha} \alpha^z G_j(z, \alpha)].$$

Proof: One has from (100)

$$\psi_{j+k}(z, \alpha) = \frac{d^k}{d\alpha^k} \psi_j(z, \alpha), \tag{107}$$

and hence

$$G_{j+k}(z, \alpha)\psi(z, \alpha) = \frac{d^k}{d\alpha^k} [\psi(z, \alpha) G_j(z, \alpha)]. \tag{108}$$

The result follows on use of (24).

Corollary:

$$G_j(z, \alpha) = \alpha^{-z} e^{\alpha} \frac{d^j}{d\alpha^j} [e^{-\alpha} \alpha^z].$$

Proof:

$$G_0(z, \alpha) = 1. \tag{109}$$

An integral representation for $G_j(z, \alpha)$ is given in

Theorem 23:

$$G_j(-z, \alpha) = (-1)^j \frac{\alpha^z}{\Gamma(z)} \int_0^\infty e^{-\alpha y}(1 + y)^j y^{z-1} dy,$$
$$\operatorname{Re} \alpha > 0, \qquad \operatorname{Re} z > 0.$$

Proof: From (20), one has

$$e^{-\alpha}\alpha^{-z} = \frac{1}{\Gamma(z)} \int_0^\infty e^{-\alpha(1+y)} y^{z-1} dy. \tag{110}$$

Substitution of (110) into the corollary of Theorem 22 yields the result.

Theorem 23 permits obtaining a Wittaker function representation.

Theorem 24:

$$G_j(z, \alpha) = (-1)^j \alpha^{-(z+j+1)/2} e^{\alpha/2} W_{(j+z+1)/2, (j-z)/2}(\alpha).$$

Proof: Comparison of Theorem 23 with (32) and replacement of $-z$ by z.

The representation of Theorem 24 remains valid, by analytic continuation, even outside the region of convergence of the integral of (32).

Corollary 1:

$$B(N, \alpha)^{-1} = (-1)^N G_N(-1, \alpha).$$

Proof: Comparison of Theorems 7 and 24.

Corollary 2:

$$G_j(z, \alpha) = \frac{z}{\alpha} G_{j-1}(z - 1, \alpha) - G_{j-1}(z, \alpha).$$

Proof: Substitution of the representation of Theorem 24 into the recurrence relation[24]

$$W_{k,m}(z) = \sqrt{z} W_{k-\frac{1}{2}, m-\frac{1}{2}}(z) + (\tfrac{1}{2} - k + m) W_{k-1,m}(z) \tag{111}$$

yields the result.

Corollary 3:

$$G_j(z, \alpha) = G_j(z - 1, \alpha) + \frac{j}{\alpha} G_{j-1}(z - 1, \alpha).$$

Proof: Same as for Corollary 2, except the following recurrence relation is used:

$$W_{k,m}(z) = \sqrt{z} W_{k-\frac{1}{2}, m+\frac{1}{2}}(z) + (\tfrac{1}{2} - k - m) W_{k-1,m}(z). \tag{112}$$

A representation of $G_j(z, \alpha)$ in terms of $B(z, \alpha)$ is given in

Theorem 25:

$$G_N(-j, \alpha) = (-1)^{N+j-1} \frac{\alpha^j}{(j-1)!} \frac{d^{j-1}}{d\alpha^{j-1}} \frac{1}{\alpha B(N, \alpha)}.$$

Proof: From (23),

$$\frac{1}{\alpha B(N, \alpha)} = \int_0^\infty e^{-\alpha y}(1 + y)^N dy, \tag{113}$$

and Theorem 23,

$$G_N(-j, \alpha) = (-1)^N \frac{\alpha^j}{(j-1)!} \int_0^\infty e^{-\alpha y}(1 + y)^N y^{j-1} dy, \tag{114}$$

one has the result on use of

$$\frac{d^{j-1}}{d\alpha^{j-1}} \int_0^\infty e^{-\alpha y}(1 + y)^N dy = (-1)^{j-1} \int_0^\infty e^{-\alpha y}(1 + y)^N y^{j-1} dy. \tag{115}$$

The Poisson-Charlier polynomials possess addition formulas similar to those of $B(z, \alpha)^{-1}$ as given in Theorems 9 and 10.

Theorem 26:

$$G_j(z, \alpha + t) = \left(1 + \frac{t}{\alpha}\right)^{-j} \sum_{\nu=0}^{j} \binom{j}{\nu} G_{j-\nu}(z, \alpha)(-1)^\nu \left(\frac{t}{\alpha}\right)^\nu.$$

Proof: Use of (103) shows that the system of functions

$$\left[\frac{(-1)^j}{j!}\right] \alpha^j G_j(z, \alpha)$$

has the generating function $e^{\alpha t}(1 - t)^z$, and hence[25] form an Appell system with respect to α, thus,

$$\frac{d}{d\alpha}\left[\frac{(-1)^j}{j!} \alpha^j G_j(z, \alpha)\right] = \frac{(-1)^{j-1}}{(j-1)!} \alpha^{j-1} G_{j-1}(z, \alpha). \tag{116}$$

The Taylor expansion of $[(-1)^j/j!](\alpha + t)^j G_j(z, \alpha, + t)$ in powers of t now yields the required result.

Theorem 27:

$$G_j(z, \alpha + t) = \left(1 + \frac{t}{\alpha}\right)^{-z} e^t \sum_{\nu=0}^{\infty} G_{j+\nu}(z, \alpha) \frac{t^\nu}{\nu!}.$$

The series is permanently convergent.

Proof: Use of (107) and Taylor's expansion yields

$$\psi_j(z, \alpha + t) = \sum_{\nu=0}^{\infty} \frac{t^\nu}{\nu!} \psi_{j+\nu}(z, \alpha). \tag{117}$$

The result is now obtained from (101) and (117). Since, from (104),

$$G_j(z, \alpha) \sim \left(\frac{z}{\alpha}\right)^j, \qquad j \to \infty, \tag{118}$$

the convergence is permanent.

An asymptotic expansion is given by

Theorem 28:

$$G_j(z, \alpha) \sim (-1)^j \left\{1 + \sum_{\nu=0}^{j} (-1)^\nu \frac{j^{(\nu)} z^{(\nu)}}{\nu ! \alpha^\nu}\right\}, \qquad \alpha \to \infty.$$

Proof: The result is obtained directly from (104). It may also be obtained from Theorem 24 and (45).

REFERENCES

1. A. K. Erlang, "Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges," P. O. Elect. Engrs. J., *10*, 189, 1917.
2. B. Wallström, "Congestion Studies in Telephone Systems with Overflow Facilities," Ericsson Technics, *22*, No. 3, 1966.
3. Y. Rapp, "Planning of Junction Network in a Multi-Exchange Area, Part 1," Ericsson Technics, *20*, 1964, pp. 77–130.
4. G. Levy-Soussan, "Numerical Evaluation of the Erlang Function Through a Continued-Fraction Algorithm," Elec. Commun., *42*, No. 2, 1968, pp. 163–168.
5. P. J. Burke, "An Interpolation Procedure of Arbitrary Accuracy for Blocking Probabilities Associated with a Nonintegral Number of Trunks," unpublished work.
6. H. Akimaru and T. Nishimura, "The Derivatives of Erlang's B Formula," Rev. Elec. Commun. Lab., *11*, No. 9–10, 1963.
7. H. Akimaru and T. Nishimura, "The Derivatives of Erlang's C Formula," Rev. Elec. Commun. Lab., *12*, No. 5–6, 1964.
8. S. Horing, H. Heffes, and J. M. Holtzman, "A Study of Demand Assignment of Satellite Capacity," unpublished work.
9. S. Miller, "Computational Methods for Calculating the Erlang B Function," unpublished work.
10. D. L. Jagerman, "A General Approximation Method with Applications to Erlangian Blocking," unpublished work.
11. D. L. Jagerman, "An Approximation Theorem of Central Limit Type," unpublished work.

12. D. L. Jagerman, "Non-Stationary Trunking Problem," unpublished work.
13. B. V. Gnedenko and I. N. Kovalenko, *Introduction to Queueing Theory*, Jerusalem: S. Monson, 1968, Chapter 1.
14. A. Y. Khintchine, *Mathematical Methods in the Theory of Queueing*, New York: Hafner, 1969.
15. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Edinburgh: Oliver and Boyd, 1959.
16. E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, New York: MacMillan, 1948, Chapter XVI.
17. A. Descloux, "Asymptotic Formulas for the Erlang Loss-Probability," unpublished work.
18. G. Doetsch, *Theorie Und Anwendung Der Laplace-Transformation*, New York: Dover Publications, 1943, Chapter 12.
19. E. Artin, *The Gamma Function*, New York: Holt, Rinehart and Winston, 1964, Chapter 1.
20. J. Riordan, *Stochastic Service Systems*, New York: Wiley, 1962.
21. V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.
22. C. Jordan, *Calculus of Finite Differences*, New York: Chelsea Publishing Co., 1947, Chapter VIII.
23. G. Doetsch, "Die in Der Statistik Seltener Ereignise Auftretenden Charlierschen Polynome Und Eine Damit Zusammenhängende Differentialgleichung," Math. Ann., *109*, 1933, pp. 257–266.
24. W. Magnus and F. Oberhettinger, *Formulas and Theorems for the Functions of Mathematical Physics*, New York: Chelsea, 1954.
25. E. B. McBride, *Obtaining Generating Functions*, New York: Springer-Verlag, 1971, Chapter V.

Contributors to This Issue

Dirk J. Bartelink, B. Sc., 1956, University of Western Ontario; M.S., 1959, and Ph.D., 1962, Stanford University; Bell Laboratories, 1961–1973. Mr. Bartelink has been engaged in fundamental studies of hot electrons in semiconductors and waves in solid-state plasmas. He has supervised a group concerned with various active solid-state devices, including characterization of the TRAPATT microwave diode. Before leaving Bell Laboratories, he investigated the physical processes of image recording on thin metallic films. Member, American Physical Society, IEEE.

William F. Bodtmann, Monmouth College, 1957–61; Bell Laboratories, 1941—. Mr. Bodtmann has been engaged in research on long- and short-haul microwave radio systems, frequency feedback receivers, and FM multiplex systems. He is presently engaged in work associated with communication systems operating at millimeter wavelengths.

Edgar N. Gilbert, B.S., 1943, Queens College; Ph.D., 1948, Massachusetts Institute of Technology; M.I.T. Radiation Laboratory, 1944–1946; Bell Laboratories, 1948—. Mr. Gilbert has done research in several branches of applied mathematics and is interested in communication theory. Member, American Mathematical Society, IEEE.

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.

D. L. Jagerman, B.E.E., 1949, Cooper Union; M.S., 1954, and Ph.D. (Mathematics), 1962, New York University; Bell Laboratories, 1964—. Mr. Jagerman has been engaged in mathematical research on numerical quadrature theory, interpolation, mathematical properties of pseudorandom number generators, dynamic programming, approximation theory, and widths and entropy with application to the storage and transmission of information. His recent work concerns the theory of queuing systems and its applications to telephone traffic problems. Member, Pi Mu Epsilon.

Jessie MacWilliams (Mrs. F. J.), B.A., 1939, M.A., 1941, Cambridge University (England); Ph.D., 1962, Harvard University; Bell Laboratories, 1956—. Mrs. MacWilliams has worked in transmission networks development and data communications engineering, and is now in the Mathematics and Statistics Research Center. Member, Mathematical Association of America, American Mathematical Society.

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.

G. Persky, B.S.E.E., 1959, Rensselaer Polytechnic Institute; M.S.E.E., 1961, and Ph.D. (Physics), 1968, Polytechnic Institute of Brooklyn; Bell Laboratories, 1967—. Mr. Persky has worked on problems of high-field transport in semiconductor devices. He is now engaged in software development for computer-aided design of integrated circuits. Member, IEEE, American Physical Society, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

Clyde L. Ruthroff, B.S.E.E., 1950, and M.A., 1952, University of Nebraska; Bell Laboratories, 1952—. Mr. Ruthroff has published contributions on the subjects of FM distortion theory, broadband transformers, FM limiters, threshold extension by feedback, microwave radio systems, rain attenuation, multiple-path propagation, linear phase modulators, and injection-locked FM receivers. He is interested in the extension of radio communication into the millimeter and optical wavelengths. Member, Sigma Xi, American Association for the Advancement of Science.

Andres C. Salazar, B.A. (Math), B.S.E.E., 1964, M.S., 1965, University of New Mexico; and Ph.D., 1967, Michigan State University; Bell Laboratories, 1967—. Mr. Salazar has been engaged in the statistical evaluation of data set performance on the switched telephone network. His current interests are in the areas of digital filter design and equalization techniques for voiceband data transmission systems. Member, IEEE, Phi Kappa Phi.

Neil J. A. Sloane, B.E.E., 1959, and B.A. (Hons.), 1960, University of Melbourne, Australia; Postmaster General's Department, Commonwealth of Australia, 1956–1961; M.S., 1964, and Ph.D., 1967, Cornell University; assistant professor of electrical engineering, Cornell University, 1967–1969; Bell Laboratories, 1969—. Mr. Sloane is engaged in research in coding theory, communication theory, and combinatorial mathematics. Member, IEEE, American Mathematical Society, Mathematical Association of America.

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.