THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 53 November 1974 Number 9



# Source Coding for a Simple Network

By R. M. GRAY and A. D. WYNER

(Manuscript received March 15, 1974)

*We consider the problem of source coding subject to a fidelity criterion for a simple network connecting a single source with two receivers via a common channel and two private channels. The region of attainable rates is formulated as an information-theoretic minimization. Several upper and lower bounds are developed and shown to actually yield a portion of the desired region in certain cases.*

## I. INTRODUCTION

### 1.1 *Informal statement of the problem*

To fix ideas, let us consider the following problem. Suppose that we are given a data source whose output is a sequence $U_1, U_2, \cdots$, that appears at the source output at the rate of 1 per second. The $\{U_k\}_1^\infty$ is a sequence of independent copies of the discrete random variable $U$, with probability distribution $\Pr\{U = u\} = Q(u)$, $u \in \mathfrak{U}$ a finite set. Our task is to transmit this data sequence over a communication channel having a capacity of $C$ bits per second so that it is represented at the output as $\hat{U}_1, \hat{U}_2, \cdots, \in \mathfrak{U}$. We assume that the data are trans-

The work of R. M. Gray was supported in part by NSF Grants GK-5452 and GK-31630 and by the Joint Services Program at Stanford Electronics Laboratory and U.S. Navy Contract N0014-67-A-112-004. Parts of this work were performed while A. D. Wyner was visiting Stanford University with the partial support of the ISL Stanford Affiliates.

mitted over the channel in blocks of length $n$, and allow processing at both the channel input and output (encoding and decoding). We define the "error rate" as

$$\Delta = E\frac{1}{n}\sum_{k=1}^{n} d_H(U_k, \hat{U}_k), \tag{1a}$$

where

$$d_H(u, \hat{u}) = \begin{cases} 0, & u = \hat{u}, \\ 1, & u \neq \hat{u}, \end{cases} \tag{1b}$$

is the Hamming metric. Thus, $\Delta$ is the average fraction of data digits delivered in error.

The question we pose is: What is the smallest capacity $C$ such that (for $n$ sufficiently large) we can transmit the data through the channel and achieve an arbitrarily small $\Delta$? The well-known answer to the question is that the minimum capacity $C$ is the entropy $H(U)$, defined by*

$$H(U) = -\sum_{u \in \mathcal{U}} Q(u) \log Q(u). \tag{2}$$

Now consider the case where the random variable $U$ is a pair $(X, Y)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We have

$$Q(u) = Q(x, y) = \Pr\{X = x, Y = y\},$$

and

$$H(U) = H(X, Y) = -\sum_{x,y} Q(x, y) \log Q(x, y).$$

Setting $\hat{U} = (\hat{X}, \hat{Y})$, $\Delta$ [as defined in (1)] is the fraction of pairs delivered in error. Thus, we conclude that $H(X, Y)$ is the minimum channel capacity required to transmit the source output $\{(X_k, Y_k)\}$ with the error rate $\Delta$ arbitrarily small.

Next, let us assume that, as above, $U = (X, Y)$, but that it is only required to transmit the sequence $\{X_k\}$ through a channel having a capacity $C_1$, and to deliver it at the channel output as $\{\hat{X}_k\}$. Let

$$\Delta_X = E\frac{1}{n}\sum_{k=1}^{n} d_H(X_k, \hat{X}_k)$$

be the error rate for a system with block coding of block length $n$. The special assumption here is that the random sequence $\{Y_k\}_{k=1}^{n}$ is available to the encoder and the decoder. See Fig. 1.

* All logarithms in this paper are assumed to be taken to the base 2.

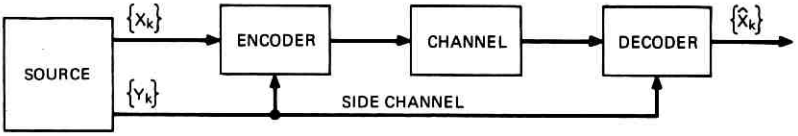



Fig. 1—Source coding with side information.

Again we ask: What is the minimum capacity $C_1$ required to transmit $\{X_k\}$ with $\Delta_X$ arbitrarily small (with $n$ sufficiently large)? The answer[1,2] is that the minimum $C_1$ is the "conditional entropy," $H(X|Y)$, defined by

$$H(X|Y) = -\sum_{x,y} Q(x, y) \log \frac{Q(x, y)}{Q_Y(y)}$$

$$= -\sum_{y} Q_Y(y)[\sum_{x} Q_{X|Y}(x|y) \log Q_{X|Y}(x|y)], \quad (3a)$$

where

$$Q_Y(y) = \Pr\{Y = y\} = \sum_{x \in \mathfrak{X}} Q(x, y) \quad (3b)$$

and

$$Q_{X|Y}(x|y) = \frac{Q(x, y)}{Q_Y(y)} = \Pr\{X = x | Y = y\}. \quad (3c)$$

Note that $H(X|Y) + H(Y) = H(X, Y)$.

Let us remark that the above still holds if, instead of delivering $\{Y_k\}$ to the decoder, we delivered a sequence $\{\hat{Y}_k\}$, where $\Delta_Y = E(1/n)\sum_{k=1}^{n} d_H(Y_k, \hat{Y}_k)$ can be made arbitrarily small. Thus, the capacity of the "side channel" must be at least $H(Y)$.

Finally, we turn our attention to the problem to which this paper is devoted. Let the source output be $\{(X_k, Y_k)\}_{k=1}^{\infty}$, as above. We assume here, however, that there are *two* receivers. Receiver 1 is interested in obtaining a reproduction $\{\hat{X}_k\}$ of the sequence $\{X_k\}$, and receiver 2 is interested in obtaining a reproduction $\{\hat{Y}_k\}$ of the sequence $\{Y_k\}$. Assume further that a network consisting of three channels is available, as in Fig. 2. The first of these channels is a "common" channel (with capacity $C_0$) that connects the transmitter to both receivers, and the other two are "private" channels that connect the transmitter to each of the two receivers (with capacities $C_1$ and $C_2$). Assuming that we use block coding with block length $n$, the error rates are

$$\Delta_X = E\frac{1}{n}\sum_{k=1}^{n} d_H(X_k, \hat{X}_k) \quad (4a)$$

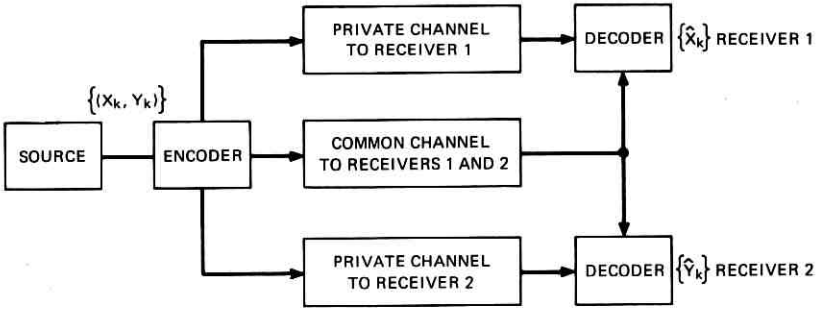



Fig. 2—Source coding for a network.

and

$$\Delta_Y = E\frac{1}{n}\sum_{k=1}^{n} d_H(Y_k, \hat{Y}_k). \tag{4b}$$

We say that a "rate-triple" $(R_0, R_1, R_2)$ is *achievable* if, for any triple of channel capacities $(C_0, C_1, C_2)$ for which $C_i > R_i$ $(i = 0, 1, 2)$ and any $\epsilon > 0$, transmission over the network of Fig. 2 (with these capacities) is possible (with $n$ sufficiently large) with $\Delta_X$, $\Delta_Y \leqq \epsilon$. Our problem is the determination of the set $\mathfrak{R}$ of achievable rate-triples.

Before stating our results, we digress to give a formal and precise statement of the problem as well as some other specialized information. This digression can be omitted by the casual reader.

### 1.2 Digression—formal statement of the problem

Let $\{(X_k, Y_k)\}_{k=1}^{\infty}$ be a sequence of independent drawings of a pair of random variables $(X, Y)$, $X \in \mathfrak{X}$, $Y \in \mathfrak{Y}$. $\mathfrak{X}$ and $\mathfrak{Y}$ are finite sets and $\Pr\{X = x, Y = y\} = Q(x, y)$, $x \in \mathfrak{X}$, $y \in \mathfrak{Y}$. The marginal distributions are

$$Q_X(x) = \sum_{y\in\mathfrak{Y}} Q(x, y) \qquad \text{and} \qquad Q_Y(y) = \sum_{x\in\mathfrak{X}} Q(x, y).$$

Often, when the random variables are clear from the context, we write $Q_X(x)$ as $Q(x)$, etc. Define, for $m = 1, 2, \cdots$, the set

$$I_m = \{0, 1, 2, \cdots, m - 1\}. \tag{5}$$

An *encoder* with parameters $(n, M_0, M_1, M_2)$ is a mapping

$$f_E: \mathfrak{X}^n \times \mathfrak{Y}^n \to I_{M_0} \times I_{M_1} \times I_{M_2}. \tag{6}$$

Given an encoder, a *decoder* is a pair of mappings

$$f_D^{(X)}: I_{M_0} \times I_{M_1} \to \mathfrak{X}^n \tag{7a}$$

$$f_D^{(Y)}: I_{M_0} \times I_{M_2} \to \mathfrak{Y}^n. \tag{7b}$$

An encoder-decoder with parameters $(n, M_0, M_1, M_2)$ is applied as follows. Let

$$f_E(\mathbf{X}, \mathbf{Y}) = (S_0, S_1, S_2), \tag{8a}$$

where

$$\mathbf{X} = (X_1, \cdots, X_n) \qquad \text{and} \qquad \mathbf{Y} = (Y_1, \cdots, Y_n).$$

Then let

$$\hat{\mathbf{X}} = f_D^{(X)}(S_0, S_1), \tag{8b}$$

$$\hat{\mathbf{Y}} = f_D^{(Y)}(S_0, S_1). \tag{8c}$$

The resulting error rate is

$$\Delta = \max(\Delta_X, \Delta_Y), \tag{9a}$$

where

$$\Delta_X = E\frac{1}{n}\sum_{k=1}^{n} d_H(X_k, \hat{X}_k), \tag{9b}$$

$$\Delta_Y = E\frac{1}{n}\sum_{k=1}^{n} d_H(Y_k, \hat{Y}_k), \tag{9c}$$

$d_H(\cdot, \cdot)$ is defined by (1b), and $\hat{X}_k$, $\hat{Y}_k$ are the $k$th coordinate of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, respectively. The Hamming distance $D_H(\mathbf{u}, \mathbf{v})$ between the $n$-vectors $\mathbf{u}$ and $\mathbf{v}$ is the number of positions in which $\mathbf{u}$ and $\mathbf{v}$ differ. Thus, $\Delta_X = E(1/n)D_H(\mathbf{X}, \mathbf{Y})$ and $\Delta_Y = E(1/n)D_H(\mathbf{Y}, \hat{\mathbf{Y}})$.

The correspondence between the encoder-decoder pair (or "code") as defined here and the communication system of Fig. 2 should be clear. Note that the capacities of the channels in that diagram must be at least $C_i = (1/n)\log_2 M_i$ $(i = 0, 1, 2)$.

A triple $(R_0, R_1, R_2)$ is said to be *achievable* if, for arbitrary $\epsilon > 0$, there exists (for $n$ sufficiently large) a code with parameters $(n, M_0, M_1, M_2)$ with $M_i \leqq 2^{n(R_i+\epsilon)}$, $i = 0, 1, 2$, and error rate $\Delta \leqq \epsilon$. We define $\mathfrak{R}$ as the set of achievable rates. Our main problem is to ascertain the region $\mathfrak{R}$.

It follows from the definition that $\mathfrak{R}$ is a closed subset of Euclidean three-space and the $\mathfrak{R}$ has the property that

$$(R_0, R_1, R_2) \in \mathfrak{R} \to (R_0 + \epsilon_0, R_1 + \epsilon_1, R_2 + \epsilon_2) \in \mathfrak{R}, \tag{10}$$

$\epsilon_i \geqq 0$, $i = 0, 1, 2$. The region $\mathfrak{R}$ is therefore completely defined by giving its lower boundary $\overline{\mathfrak{R}}$, where

$$\overline{\mathfrak{R}} \triangleq \{R_0, R_1, R_2) \in \mathfrak{R}: (\hat{R}_0, \hat{R}_1, \hat{R}_2) \in \mathfrak{R}, \\ \hat{R}_i \leqq R_i (i = 0, 1, 2) \rightarrow \hat{R}_i = R_i (i = 0, 1, 2)\}. \tag{11}$$

It follows immediately that $\overline{\mathfrak{R}}$ too is closed.

It can also be verified by a simple "time-sharing" argument that $\mathfrak{R}$ is convex (see appendix). This leads us to the following equivalent formulation of the problem. Let $\alpha_i \geqq 0$, $i = 0, 1, 2$ be arbitrary. Then define

$$T_1(\alpha_0, \alpha_1, \alpha_2) = \min_{(R_0, R_1, R_2) \in \mathfrak{R}} (\alpha_0 R_0 + \alpha_1 R_1 + \alpha_2 R_2).$$

Then it follows from the convexity of $\mathfrak{R}$ that the lower boundary $\overline{\mathfrak{R}}$ is the upper envelope of the family of planes $\sum_0^2 \alpha_i R_i = T_1(\alpha_0, \alpha_1, \alpha_2)$.

We can think of $T_1(\alpha_0, \alpha_1, \alpha_2)$ as the minimum cost of transmitting, using a code with rate-triple $(R_0, R_1, R_2)$ over the network of Fig. 2, when the cost of transmitting a bit per second over the common channel is $\alpha_0$ and the costs of transmitting a bit per second over the private channels to receivers 1 and 2 are $\alpha_1$ and $\alpha_2$, respectively. Now, since information sent over the common channel (in Fig. 2) can alternatively be sent over *both* private channels, it is never necessary to consider the case where the sum of the costs of a bit per second on the private channels $\alpha_1 + \alpha_2 < \alpha_0$, the cost of a bit per second on the common channel. Similarly, we need never consider the cases where $\alpha_1 > \alpha_0$, or $\alpha_2 > \alpha_0$, since information transmitted over a private channel can alternatively be sent over the common channel. Since we can normalize $\alpha_0$ as unity, the following theorem should be plausible. A complete proof is given in the appendix.

For $\mathbf{R} = (R_0, R_1, R_2)$ satisfying $R_i \geqq 0$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ arbitrary, let the "cost" be defined by

$$C(\boldsymbol{\alpha}, \mathbf{R}) = R_0 + \alpha_1 R_1 + \alpha_2 R_2. \tag{12}$$

With $\boldsymbol{\alpha}$ held fixed, let

$$T(\boldsymbol{\alpha}) = \min_{\mathbf{R} \in \mathfrak{R}} C(\boldsymbol{\alpha}, \mathbf{R}). \tag{13}$$

The indicated minimum exists because $\mathfrak{R}$ is closed. For $\boldsymbol{\alpha}$ arbitrary, let $S(\boldsymbol{\alpha})$ be the set of $\mathbf{R} \in \mathfrak{R}$ that achieve $T(\boldsymbol{\alpha}) = C(\boldsymbol{\alpha}, \boldsymbol{R})$.

*Theorem 1*:

$$(i)\ \overline{\mathfrak{R}} \subseteq \bigcup_{\alpha \in \mathfrak{A}} \mathfrak{S}(\alpha),$$

$$(ii)\ \bigcup_{\alpha \in \mathfrak{A}'} \mathfrak{S}(\alpha) \subseteq \overline{\mathfrak{R}},$$

*where the boundary* $\overline{\mathfrak{R}}$ *is defined in (11),* $\mathfrak{A}$ *is the set of* $\alpha = (\alpha_1, \alpha_2)$ *that satisfy*

$$0 \leqq \alpha_1, \alpha_2 \leqq 1, \qquad \alpha_1 + \alpha_2 \geqq 1,$$

*and* $\mathfrak{A}'$ *is* $\mathfrak{A}$ *with the elements* $(0, 1)$ *and* $(1, 0)$ *deleted.*

*Remarks*:

(1) $(0, 1)$ and $(1, 0)$ are the only pairs in $\mathfrak{A}$ with zero elements. Thus, $\mathfrak{A}$ and $\mathfrak{A}'$ are nearly identical.

(2) The theorem implies that $\overline{\mathfrak{R}}$ is upper envelope in $(R_0, R_1, R_2)$-space of the family of planes defined by

$$R_0 + \alpha_1 R_1 + \alpha_2 R_2 = T(\alpha),$$

$\alpha \in \mathfrak{A}$.

### 1.3 Upper and lower bounds on $\mathfrak{R}$

#### 1.3.1 Lower bounds

We can immediately give some lower bounds to the region $\mathfrak{R}$. We state them as

*Theorem 2*: *If* $(R_0, R_1, R_2) \in \mathfrak{R}$, *then*

(a) $R_0 + R_1 + R_2 \geqq H(X, Y)$,

(b) $R_0 + R_1 \geqq H(X)$,

(c) $R_0 + R_2 \geqq H(Y)$.

*Proof*: Suppose that $(R_0, R_1, R_2) \in \mathfrak{R}$. Then, for arbitrary $\epsilon > 0$, we can (for sufficiently large block length $n$) reproduce $\{X_k\}$, and $\{Y_k\}$ with arbitrarily small $\Delta_X$, $\Delta_Y$, with capacity triple (in Fig. 2)

$$(C_0, C_1, C_2) = (R_0 + \epsilon, R_1 + \epsilon, R_2 + \epsilon).$$

That is, with a code with $M_i = 2^{nC_i}$, $i = 0, 1, 2$.

Since the total capacity of the three channels is $C_0 + C_1 + C_2$, we must have

$$C_0 + C_1 + C_2 = R_0 + R_1 + R_2 + 3\epsilon \geqq H(X, Y).$$

Letting $\epsilon \to 0$, we have established $(a)$. Inequality $(b)$ follows in an identical way on observing that the common channel (with capacity $C_0$) and the private channel to receiver 1 (with capacity $C_1$) together transmit $\{X_k\}$. Inequality $(c)$ follows, similarly.

Let us remark that inequality $(a)$ is an expression of the fact that a communication system with the constraints imposed in Fig. 2 cannot perform better than in the "best of all possible worlds" situation in which the receivers can collaborate. It is therefore called the "Pangloss bound." The set of triples $(R_0, R_1, R_2)$ that satisfy $\sum_0^2 R_i = H(X, Y)$ are called the "Pangloss plane." Corresponding to rate-triples that lie on the intersection of $\mathcal{R}$ and the Pangloss plane, the approximately $H(X, Y)$ bits per second that characterize $\{X_k, Y_k\}$ can be split up into three parts (corresponding to the information transmitted over the three channels in our network) such that $\{X_k, Y_k\}$ can be essentially perfectly reconstructed by the three receivers in the network. In this situation, the information transmitted over the common channel represents a kind of "core" process. Furthermore, the smallest $R_0$, such that $(R_0, R_1, R_2) \in \mathcal{R}$ and lies on the Pangloss plane (for some $R_1, R_2$), can be thought of as a measure of the "common information" of $\{X_k\}$ and $\{Y_k\}$. This point is explored thoroughly in Ref. 3.

#### *1.3.2 Some easily achievable rate-triples*

We now assert that certain rate-triples are achievable.

*Theorem 3*: *The following triples belong to* $\mathcal{R}$:

$(A)$ $R_0 = H(X, Y), \quad R_1 = R_2 = 0$

$(B)$ $R_0 = 0, \quad R_1 = H(X), \quad R_2 = H(Y)$

$(C)$ $R_0 = H(Y), \quad R_1 = H(X \mid Y), \quad R_2 = 0$

$(D)$ $R_0 = H(X), \quad R_1 = 0, \quad R_2 = H(Y \mid X).$

*Proof*: To achieve $(A)$, simply transmit $\{(X_k, Y_k)\}$ over the common channel (and do not use the private channels). To achieve $(B)$, transmit $\{X_k\}$ and $\{Y_k\}$ over the private channels to receivers 1 and 2, respectively (and do not use the common channel). To achieve $(C)$, transmit $\{Y_k\}$ over the common channel (requiring a capacity of about $H(Y)$), and deliver $\{\hat{Y}_k\}$ to receiver 1 to use as side information for transmitting $\{X_k\}$ over the private channel to receiver 1. This will require a capacity of about $H(X \mid Y)$. We do not use the private channel

to receiver 2. Triple $(D)$ can be achieved as in $(C)$ by reversing to roles of $X$ and $Y$.

Let us remark that points $(C)$ and $(D)$ lie on the Pangloss plane (i.e., they satisfy relation $(a)$ of Theorem 2 with equality), since $H(X) + H(Y|X) = H(Y) + H(X|Y) = H(X, Y)$. Furthermore, because of the convexity of $\mathfrak{R}$, all triples that are linear combinations of triples $(A)$ to $(D)$ are also members of $\mathfrak{R}$. The situation is summarized in Fig. 3. The plane labeled "$(a)$" in the figure is the Pangloss plane defined by $R_0 + R_1 + R_2 = H(X, Y)$. Theorem $2(a)$ states that the region $\mathfrak{R}$ (and therefore its lower boundary $\overline{\mathfrak{R}}$) lies above this plane. Similarly, Theorem $2(b, c)$ states that $\mathfrak{R}$ and $\overline{\mathfrak{R}}$ lie above the planes labeled "$(b)$" and "$(c)$" in Fig. 3.

Now the points labeled "$A$," "$B$," "$C$," and "$D$" in the figure are points $(A)$ to $(D)$ respectively in Theorem 3. As we mentioned previously, points $C$ and $D$ (as well as $A$) lie on plane $a$. Thus (from the convexity of $\mathfrak{R}$), the triangle $ADC$ lies in $\mathfrak{R}$ and must therefore be part of the lower boundary $\overline{\mathfrak{R}}$. Further, since points $D$ and $B$ lie on plane $b$,

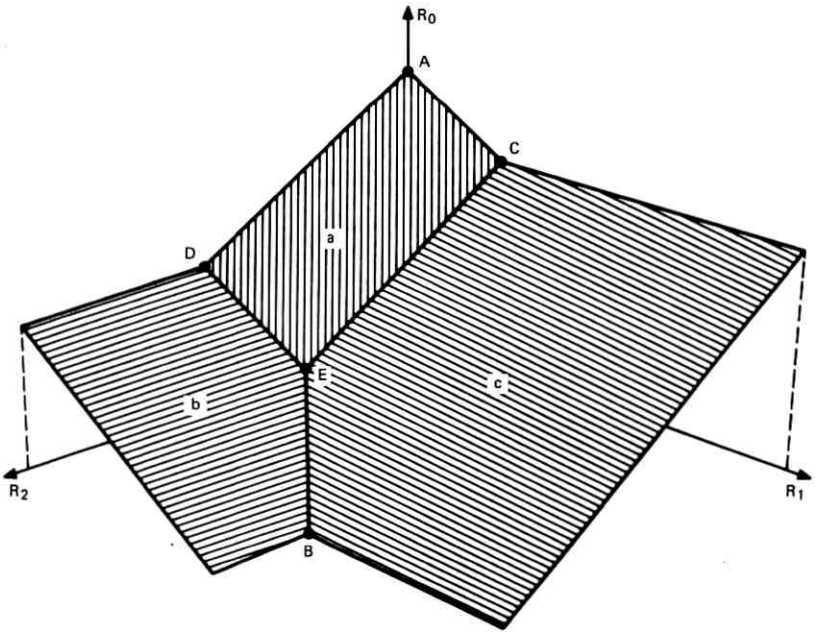



Fig. 3—Estimates of rate-region $\mathfrak{R}$.

the line $DB$ is part of $\overline{\mathfrak{R}}$. Similarly, line $BC$ is part $\overline{\mathfrak{R}}$. Finally, since points $B$, $C$, and $D$ are achievable, so are the points on the triangle $BCD$. Thus, the only unknown part of the lower boundary $\overline{\mathfrak{R}}$ lies in the (upside-down) triangular pyramid with base $BCD$ and apex at point $E$ (the intersection of planes $a$, $b$, $c$). The coordinates of point $E$ are easily seen to be $(R_0, R_1, R_2) = [I(X;Y), H(X|Y), H(Y|X)]$.

Let us remark here that there is one special source distribution $Q(x, y)$ for which point $E$ is achievable.† In this case, the entire boundary region $\overline{\mathfrak{R}}$ lies on planes $a$, $b$, $c$. This special case is when $X$, $Y$ can be written $X = (X', V)$, $Y = (Y', V)$, where $X'$ and $Y'$ are conditionally independent given $V$. Then $I(X, Y) = H(V)$, $H(X|Y) = H(X'|V)$, $H(Y|X) = H(Y'|V)$, so that point $E$ is $R_0 = H(V)$, $R_1 = H(X'|V)$, $R_2 = H(Y'|V)$. Clearly, if, in the system of Fig. 2, we transmit $V$ over the common channel and $X'$ and $Y'$ over the two private channels, we can reconstruct $X = (X', V)$ at receiver 1 and $Y = (Y', V)$ at receiver 2. This requires a capacity triple $C_0 = H(V) + \epsilon$, $C_1 = H(X'|V) + \epsilon$, $C_2 = H(Y'|V) + \epsilon$ ($\epsilon > 0$ arbitrary), so that point is in fact achievable.

We now give a characterization of the region $\mathfrak{R}$ (and therefore of $\overline{\mathfrak{R}}$) in terms of information theoretic quantities. This characterization is, in fact, the main result.

### 1.4 Characterization of $\mathfrak{R}$—the main result

Suppose we are given $Q(x, y)$, $x \in \mathfrak{X}$, $y \in \mathfrak{Y}$, an arbitrary probability function, where $\mathfrak{X}$, $\mathfrak{Y}$ are finite. Let $\mathcal{P}$ be the family of probability functions $p(x, y, w)$, where $x \in \mathfrak{X}$, $y \in \mathfrak{Y}$, $w \in \mathfrak{W}$, and $\mathfrak{W}$ is another finite set, for which

$$\sum_{w \in \mathfrak{W}} p(x, y, w) = Q(x, y), x \in \mathfrak{X}, y \in \mathfrak{Y}. \tag{14}$$

Each $p \in \mathcal{P}$ defines discrete random variables $X$, $Y$, $W$ in an obvious way. For each $p \in \mathcal{P}$, define the subset of Euclidean three-space

$$\mathfrak{R}^{(p)} = \{(R_0, R_1, R_2) : R_0 \geqq I(X, Y; W), \quad R_1 \geqq H(X|W), \quad R_2 \geqq H(Y|W)\}, \tag{15a}$$

and then let

$$\mathfrak{R}^* = \Big(\bigcup_{p \in \mathcal{P}} \mathfrak{R}^{(p)}\Big)^c, \tag{15b}$$

† M. Kaplan has shown that, in fact, this special case is the only one for which point $E$ is achievable.

where $(\ )^c$ denotes set closure. Then our main result (the proof of which is given in Section III) is

*Theorem 4*: $\mathfrak{R} = \mathfrak{R}^*$.

*Remarks*:

(1) Let us define $\mathcal{P}_T$ as the family of "test channel" transition probabilities. That is, $\mathcal{P}_T$ is the family of all $p_t(w|x,y)\,(x \in \mathfrak{X}, y \in \mathfrak{Y}, w \in \mathfrak{W})$, where $\mathfrak{W}$ is a finite set, and for each $(x, y)$, $p_t(w|x, y)$ is a probability function on $\mathfrak{W}$. Corresponding to each $p_t \in \mathcal{P}_T$, we have $p(x, y, w) = Q(x, y)p_t(w|x, y) \in \mathcal{P}$. Further, for each $p \in \mathcal{P}$, we have $p_t(w|x, y) = [p(x, y, w)/Q(x, y)] \in \mathcal{P}_T$. Thus $\mathcal{P}$ is in 1-1 correspondence with $\mathcal{P}_T$.

(2) Since $\mathfrak{R}$ is convex, Theorem 4 implies that $\mathfrak{R}^*$ is convex also.

(3) Theorem 4 can be invoked to show that $T(\boldsymbol{\alpha})$ defined in (13) is also given by

$$T(\boldsymbol{\alpha}) = \inf_{p \in \mathcal{P}} [I(X, Y; W) + \alpha_1 H(X|W) + \alpha_2 H(Y|W)]. \quad (16)$$

Thus, from Theorem 1, the lower boundary $\overline{\mathfrak{R}}$, and therefore $\mathfrak{R}$, is essentially determined by $T(\boldsymbol{\alpha})$ given by (16).

(4) Theorems 2 and 3 can be verified easily by using Theorem 4. Thus, if $(R_0, R_1, R_2) \in \mathfrak{R}$, from Theorem 4 for arbitrary $\epsilon > 0$ we can find a triple of random variables $X$, $Y$, $W$ such that

$$\begin{aligned} R_0 + R_1 + R_2 &\geqq I(X, Y; W) + H(X|W) + H(Y|W) - \epsilon \\ &= H(X, Y) + [H(X|W) + H(Y|W) - H(X, Y|W)] - \epsilon \\ &\geqq H(X, Y) - \epsilon \to H(X, Y), \qquad \text{as } \epsilon \to 0. \end{aligned} \quad (17)$$

This is Theorem 2(a). The second inequality in (17) follows from the fact that the entropy of a pair of random variables is less than the sum of the respective entropies. Part $(b)$ of Theorem 2 follows from

$$\begin{aligned} I(X, Y; W) + H(X|W) &= I(X; W) + I(Y; W|X) \\ &\quad + H(X|W) \geqq I(X; W) + H(X|W) = H(X). \end{aligned} \quad (18)$$

The first equality in (18) follows from a standard identity [Ref. 4, Eq. (2.2.29)].

Theorem 3 follows from Theorem 4 on taking $W$ as follows: $(A)$ $W = (X, Y)$, $(B)$ $W = 0$, $(C)$ $W = Y$, $(D)$ $W = X$.

(5) Although Theorem 4 characterizes $\mathfrak{R}$ and $\overline{\mathfrak{R}}$ by an information theoretic minimization, it must be emphasized that the minimization is not, in general, easy. In fact, there is no nontrivial case for which we have succeeded in calculating the entire boundary $\overline{\mathfrak{R}}$ analytically.

Its major utility at this point has been in finding upper bounds on $\overline{\mathfrak{R}}$ by guessing at a $p$ or $p_t$ and calculating the corresponding triple $[I(X, Y; W), H(X|W), H(Y|W)]$, which must lie above $\overline{\mathfrak{R}}$. See the example below. The problem of computation of $\overline{\mathfrak{R}}$ both analytically and numerically is still open.†

(6) For $p \in \mathcal{P}$, we can define the quantities

$$\beta_{xy}(w) = \Pr\{X = x, Y = y | W = w\}, \qquad x \in \mathfrak{X}, y \in \mathfrak{Y}, w \in \mathfrak{W},$$

which can be thought of as the transition probabilities of the "backward test channel." For a given $(x, y)$, we can think of $\beta_{xy} = \beta_{x,y}(W)$ as a random variable. Of course, $\beta_{xy}$ must satisfy

$$\beta_{xy} \geqq 0, \tag{19a}$$

$$\sum_{x,y} \beta_{xy} = 1, \tag{19b}$$

and

$$E\beta_{xy} = Q(x, y), \tag{19c}$$

where the expectation is taken over the distribution for $W$. Further,

$$I(X, Y; W) = H(X, Y) - H(X, Y|W) = H(X, Y) - E\sum_{x,y} \beta_{xy} \log \frac{1}{\beta_{xy}}, \tag{20a}$$

$$H(X|W) = E\sum_{x} \beta_x^{(1)} \log \frac{1}{\beta_x^{(1)}}, \qquad H(Y|W) = E\sum_{y} \beta_y^{(2)} \log \frac{1}{\beta_y^{(2)}}, \tag{20b}$$

where

$$\beta_x^{(1)} = \sum_y \beta_{xy} = \Pr\{X = x|W\}, \quad \text{and} \quad \beta_y^{(2)} = \sum_x \beta_{xy} = \Pr\{Y = y|W\}, \tag{20c}$$

and the expectation is taken over the distribution for $W$. Using this idea, it is possible to characterize, for example, $T(\boldsymbol{\alpha})$ as follows (see Ref. 3, for a precise proof of this characterization). Given $Q(x, y)$, $x \in \mathfrak{X}$, $y \in \mathfrak{Y}$, define $\mathfrak{B}$ as the family of collections of random variables, $\{\beta_{xy}\}$, $x \in \mathfrak{X}$, $y \in \mathfrak{Y}$, which satisfy (19). Then

$$T(\boldsymbol{\alpha}) = \min[I(X, Y; W) + \alpha_1 H(X|W) + \alpha_2 H(Y|W)],$$

† One reason for the difficulty is that $I(X, Y; W) + \alpha_1 H(X|W) + \alpha_2 H(Y|W)$ is apparently neither convex nor concave in $p_t$.

where $I(X, Y; W)$, $H(X|W)$, $H(Y|W)$ are given by (20) and the minimum (which can be shown to exist) is over all sets $\{\beta_{xy}\}$ in $\mathfrak{B}$.

This characterization may have value in the computation problem, since the quantities in (20) are *linear* functions of the joint distribution function for the $\{\beta_{xy}\}$ and the constraints of (19) are also linear inequalities in this distribution function. Thus, calculation of $T(\alpha)$ is a linear programming problem.

(7) If $p \in \mathcal{P}$ is such that $X$ and $Y$ are conditionally independent given $W$, then $H(X, Y|W) = H(X|W) + H(Y|W)$. Thus, with $R_0 = I(X, Y; W)$, $R_1 = H(X|W)$, $R_2 = H(Y|W)$,

$$\begin{aligned} R_0 + R_1 + R_2 &= H(X, Y) - H(X, Y|W) + H(X|W) + H(Y|W) \\ &= H(X, Y), \end{aligned}$$

and $(R_0, R_1, R_2) \in \mathfrak{R}$ and lies on the Pangloss plane. Reference 3 shows that this class of triples (corresponding to a $p \in \mathcal{P}$, with $X$, $Y$ conditionally independent given $W$) completely characterizes the intersection of $\mathfrak{R}$ and the Pangloss plane.

### 1.5 An example

As an example of the preceding, let us consider the special case where the source is the "doubly symmetric binary source" (DSBS), where $\mathfrak{X} = \mathfrak{Y} = \{0, 1\}$, and

$$Q(x, y) = \tfrac{1}{2}(1 - p_0)\delta_{x,y} + \tfrac{1}{2}p_0(1 - \delta_{x,y}), \quad x, y = 0, 1, \quad (21)$$

and the parameter $p_0$ satisfies $0 \leqq p_0 \leqq \frac{1}{2}$. We can think of $X$ as being an unbiased binary input into a binary symmetric channel (BSC) with crossover probability $p_0$, and $Y$ as being the corresponding output, or vice versa. To get a clearer picture of the set of achievable rates $\mathfrak{R}$, let us restrict ourselves to the plane in $(R_0, R_1, R_2)$-space, where $R_1 = R_2$. The intersection of $\mathfrak{R}$ and this plane can be plotted in a two-dimensional picture.

Let us first take a look at the implications of Theorems 2 and 3. In this source,

$$H(X) = H(Y) = 1, \qquad H(X|Y) = H(Y|X) = h(p_0)$$

and

$$H(X, Y) = H(X) + H(Y|X) = 1 + h(p_0),$$

where

$$h(\lambda) = -\lambda \log \lambda - (1 - \lambda) \log (1 - \lambda), \qquad 0 \leqq \lambda \leqq 1 \quad (22)$$

is the entropy function. [We take $h(0) = h(1) = 1$.] With $R_1 = R_2$,

Theorem 2 yields

$$R_0 + 2R_1 \geqq 1 + h(p_0), \tag{23a}$$

$$R_0 + R_1 \geqq 1. \tag{23b}$$

Thus, $\mathcal{R}$ and therefore the lower boundary $\overline{\mathcal{R}}$ must lie above the lines labeled $a$ and $b$ in Fig. 4.

Now Theorem 3 implies that points $A[R_0 = 1 + h(p_0),\ R_1 = 0]$, and $B(R_0 = 0,\ R_1 = 1)$ are achievable, so that any point on the line connecting them is also achievable. But we can do better. Let us drop for a moment the requirement that $R_1 = R_2$. From Theorem 3, $C$ and $D$, the points $[R_0 = 1, R_1 = h(p_0), R_2 = 0]$ and $[R_0 = 1, R_1 = 0, R_2 = h(p_0)]$

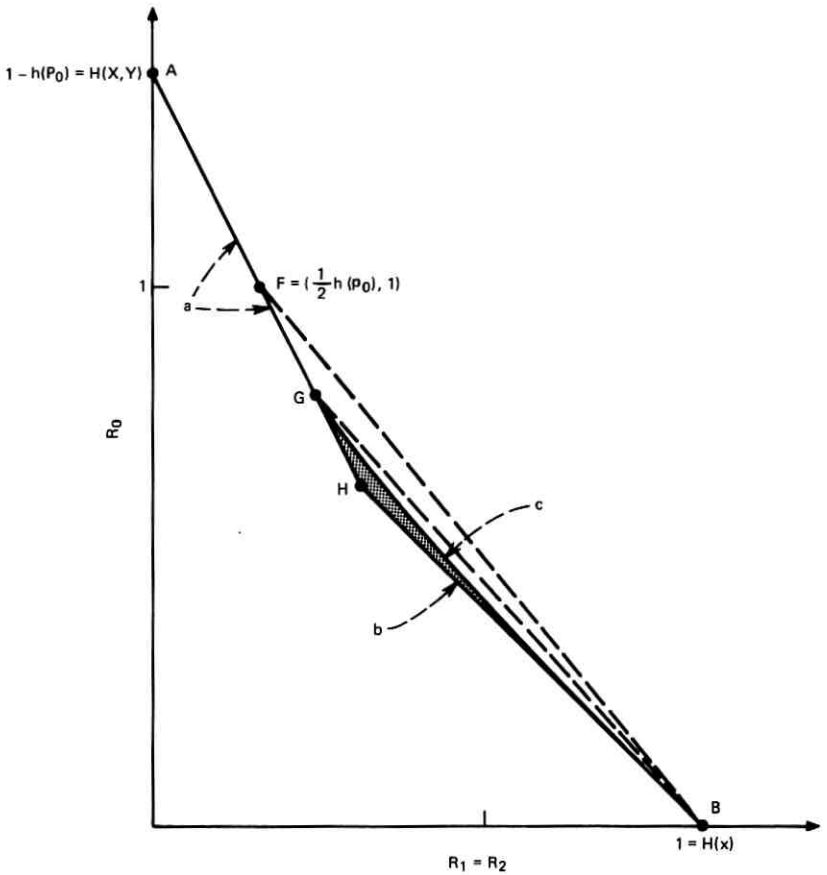



Fig. 4—Estimates of rate-region $\mathcal{R}$ for the DSBS.

are achievable. Thus, the point in $(R_0, R_1, R_2)$-space halfway between them is also achievable. But this point,

$$[R_0 = 1, R_1 = \tfrac{1}{2}h(p_0), R_2 = \tfrac{1}{2}h(p_0)],$$

satisfies $R_1 = R_2$, and is therefore of interest to us now. Point F in Fig. 4 is therefore achievable, and therefore so are line segments $AF$ and $FB$. But line segment $AF$ coincides with line $a$, so that it must be on the boundary $\overline{\mathfrak{R}}$. So far, the unknown part of the boundary curve $\overline{\mathfrak{R}}$ lies in triangle $FHB$. We can do better, however, by using Theorem 4.

Theorem 4 asserts that any triple in $\mathfrak{R}^{(p)}$, $p \in \mathcal{P}$, is achievable. We therefore guess at a $p \in \mathcal{P}$ that defines random variables $X$, $Y$, $W$, and then assert that the triple $R_0 = I(X, Y; W)$, $R_1 = H(X|W)$, $R_2 = H(Y|W)$ is achievable. Since we choose a $p \in \mathcal{P}$ such that $R_1 = R_2$, this triple is of interest in our present discussion. The $p \in \mathcal{P}$ we have chosen is (with $\mathcal{W} = \{0, 1\}$) given by Table I. The quantity $p_1 = \frac{1}{2}(1 - \sqrt{1 - 2p_0})$. One way of characterizing $p$ is to think of $W$ as an unbiased binary input and $X$, $Y$ the respective outputs of two independent BSC's, each with crossover probability $p_1$. Note that these two BSC's in cascade are equivalent to a single BSC with crossover probability, $2p_1(1 - p_1) = p_0$.

With $X$, $Y$, $W$ so defined, $X$, $Y$ are conditionally independent given $W$, so that $(R_0, R_1, R_2)$ lies on the Pangloss plane. [See remark (6) following Theorem 4.] We have

$$\begin{aligned} R_0 &= I(X, Y; W) = H(X, Y) - H(X, Y|W) \\ &= 1 + h(p_0) - 2h(p_1), \\ R_1 &= R_2 = H(X|W) = h(p_1). \end{aligned} \tag{24}$$

This is point $G$ in Fig. 4. Line segment $AG$ is therefore on the boundary $\overline{\mathfrak{R}}$. From these simple arguments, we see that the unknown part of the boundary $\overline{\mathfrak{R}}$ lies in the triangle $GHB$.

To obtain a still tighter bound on $\overline{\mathfrak{R}}$, we employ the same technique as above—i.e., "guessing" at a $p \in \mathcal{P}$ and then deducing that $\mathfrak{R}^{(p)}$

**Table I**

| $W$ \ $XY$ | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| 0 | $\frac{1}{2}(1 - p_1)^2$ | $\frac{1}{2}p_1(1 - p_1)$ | $\frac{1}{2}p_1(1 - p_1)$ | $\frac{1}{2}p_1^2$ |
| 1 | $\frac{1}{2}p_1^2$ | $\frac{1}{2}p_1(1 - p_1)$ | $\frac{1}{2}p_1(1 - p_1)$ | $\frac{1}{2}(1 - p_1)^2$ |

**Table II**

| $W$ \ $XY$ | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| 0 | $\frac{1}{2}\left(1-\beta-\frac{p_0}{2}\right)$ | $p_{0/4}$ | $p_{0/4}$ | $\frac{1}{2}\left(\beta-\frac{p_0}{2}\right)$ |
| 1 | $\frac{1}{2}\left(\beta-\frac{p_0}{2}\right)$ | $p_{0/4}$ | $p_{0/4}$ | $\frac{1}{2}\left(1-\beta-\frac{p_0}{2}\right)$ |

$\subseteq \mathcal{R}$. Let $\beta$ be a parameter for which

$$p_1 = \tfrac{1}{2}(1 - \sqrt{1 - 2p_0}) \leqq \beta \leqq \tfrac{1}{2}. \tag{25}$$

Then let $\mathcal{W} = \{0, 1\}$, and $p(x, y, w)$ be given by Table II. Then the triples $(R_0, R_1, R_2) \in \mathcal{R}$, where

$$\begin{aligned} R_0 &= I(X, Y; W) = H(X, Y) - H(X, Y|W) \\ &= 1 + h(p_0) + \frac{1}{2}\left(1-\beta-\frac{p_0}{2}\right)\log\frac{1}{2}\left(1-\beta-\frac{p_0}{2}\right) \\ &\qquad + \frac{p_0}{2}\log\frac{p_0}{4} + \frac{1}{2}\left(\beta-\frac{p_0}{2}\right)\log\frac{1}{2}\left(\beta-\frac{p_0}{2}\right) \end{aligned} \tag{26a}$$

and

$$R_1 = R_2 = H(X|W) = H(Y|W) = h(\beta). \tag{26b}$$

For $\beta = p_1$, the triple of (26) coincides with that of (25), i.e., point $G$ in Fig. 4. For $\beta = \frac{1}{2}$, the triple of (26) is $R_0 = 0$, $R_1 = R_2 = 1$, i.e., point $B$ of Fig. 4. As $\beta$ increases from $p_1$ to $\frac{1}{2}$, the family of rate-triples of (26) generate a curve $c$, which lies below the line $GB$ and therefore constitutes a tighter upper bound on $\overline{\mathcal{R}}$. We conclude that the unknown portion of $\overline{\mathcal{R}}$ lies in the shaded region in Fig. 4.

In Section 2.5 we give some insight into how we "guessed" at these distributions $p \in \mathcal{P}$.

## II. GENERALIZATION TO A FIDELITY CRITERION

In this section we formulate a generalization of the problem of Section I in which we require that the source sequences $\{X_k\}$ and $\{Y_k\}$ be reproduced to within a specified fidelity criterion and not, as in Section I, essentially perfectly. The proofs of the main theorems appear in Section III.

### *2.1 Definitions and formulation of the problem*

Let $\{(X_k, Y_k)\}_{k=1}^{\infty}$ be a sequence of independent drawings of a pair of random variables $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, where the "source alphabets"

$\mathfrak{X}$ and $\mathfrak{Y}$ are either discrete sets, the reals, or arbitrary measurable spaces. We assume that we are given a probability law that defines $(X, Y)$. If $\mathfrak{X}$ and $\mathfrak{Y}$ are discrete, then we write

$$Q(x, y) = \Pr\{X = x, Y = y\}, \qquad x \in \mathfrak{X}, y \in \mathfrak{Y}.$$

If $\mathfrak{X}$, $\mathfrak{Y}$ are the reals, then $(X, Y)$ may be defined by a probability density $Q(x, y)$, $-\infty < x, y < \infty$. For arbitrary measurable $\mathfrak{X}$, $\mathfrak{Y}$, the pair $(X, Y)$ is defined by a probability measure $Q$ on $\mathfrak{X} \times \mathfrak{Y}$. The marginal distribution for $X$, $Y$ will be defined similarly by $Q_X$, $Q_Y$ respectively.

As in (5), define the set $I_m = \{0, 1, \cdots, m-1\}$ for $m = 1, 2, \cdots$. An *encoder* with parameters $(n, M_0, M_1, M_2)$ is [as in (6)] a mapping

$$f_E: \mathfrak{X}^n \times \mathfrak{Y}^n \to I_{M_0} \times I_{M_1} \times I_{M_2}. \tag{27}$$

We assume that the sequences $\{X_k\}$ and $\{Y_k\}$ are to be reproduced as sequences of elements of sets $\hat{\mathfrak{X}}$ and $\hat{\mathfrak{Y}}$, respectively, called "reproducing alphabets." Thus [as in (7)], corresponding to a given encoder, a decoder is a pair of mappings

$$f_D^{(X)}: I_{M_0} \times I_{M_1} \to \hat{\mathfrak{X}}^n, \tag{28a}$$

$$f_D^{(Y)}: I_{M_0} \times I_{M_1} \to \hat{\mathfrak{Y}}^n. \tag{28b}$$

Let us adopt the convention of denoting $n$-vectors with bold-face type (either upper or lower case) and the components as the same subscripted letter in ordinary type. For example, $\mathbf{u} = (u_1, \cdots, u_n)$.

An encoder-decoder with parameters $(n, M_0, M_1, M_2)$ is applied as follows. Say

$$f_E(\mathbf{X}, \mathbf{Y}) = (S_0, S_1, S_2), \tag{29a}$$

where $\mathbf{X} \in \mathfrak{X}^n$, $\mathbf{Y} \in \mathfrak{Y}^n$, and $(S_0, S_1, S_2)$ is a triplet of indices. Then set

$$\hat{\mathbf{X}} = f_D^{(X)}(S_0, S_1), \qquad \hat{\mathbf{Y}} = f_D^{(Y)}(S_0, S_2), \tag{29b}$$

where $\hat{\mathbf{X}} \in \hat{\mathfrak{X}}^n$, $\hat{\mathbf{Y}} \in \hat{\mathfrak{Y}}^n$. The encoder-decoder is said to have *average distortion* $(\Delta_X, \Delta_Y)$, where

$$\Delta_X = ED_1(\mathbf{X}, \hat{\mathbf{X}}), \qquad \Delta_Y = ED_2(\mathbf{Y}, \hat{\mathbf{Y}}), \tag{30a}$$

and the single-letter distortion functions are defined by

$$D_1(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{k=1}^{n} d_1(x_k, \hat{x}_k), \tag{30b}$$

$$D_2(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{k=1}^{n} d_2(y_k, \hat{y}_k), \tag{30c}$$

$\mathbf{x} \in \mathcal{X}^n$, $\hat{\mathbf{x}} \in \hat{\mathcal{X}}^n$, $\mathbf{y} \in \mathcal{Y}^n$, $\hat{\mathbf{y}} \in \mathcal{Y}^n$, and $d_1(\cdot, \cdot)$ is a given nonnegative per-letter distortion function for the X-receiver and $d_2(\cdot, \cdot)$ is a given nonnegative per-letter distortion function for the Y-receiver. An encoder-decoder with parameters $(n, M_0, M_1, M_2)$ with average distortion $(\Delta_X, \Delta_Y)$ is said to be a *code* $(n, M_0, M_1, M_2, \Delta_X, \Delta_Y)$.

A rate-triple $(R_0, R_1, R_2)$ is said to be $(\Delta_1, \Delta_2)$-*achievable* if, for arbitrary $\epsilon > 0$ and $n$ sufficiently large, there exists a code $(n, M_0, M_1, M_2, \Delta_X, \Delta_Y)$ with

$$M_i \leqq 2^{n(R_i+\epsilon)}, \qquad i = 0, 1, 2,$$

and

$$\Delta_X \leqq \Delta_1 + \epsilon, \qquad \Delta_Y \leqq \Delta_2 + \epsilon.$$

The set of all $(\Delta_1, \Delta_2)$-achievable rate-triples is called $\mathfrak{R}(\Delta_1, \Delta_2)$. Our main problem is to ascertain $\mathfrak{R}(\Delta_1, \Delta_2)$, $\Delta_1, \Delta_2 \geqq 0$. Clearly, this generalized problem reduces to the problem of Section I, if $\mathcal{X} = \hat{\mathcal{X}}$, $\mathcal{Y} = \hat{\mathcal{Y}}$, $d_1 = d_2 = d_H$, and $\Delta_1 = \Delta_2 = 0$. As in Section I, the region $\mathfrak{R}(\Delta_1, \Delta_2)$ is completely defined by the boundary $\overline{\mathfrak{R}}(\Delta_1, \Delta_2)$, where $\overline{\mathfrak{R}} = \mathfrak{R}(\Delta_1, \Delta_2)$ is defined in (11). Further, we show in the appendix that $\mathfrak{R}(\Delta_1, \Delta_2)$ is convex and that Theorem 1 holds with $\mathfrak{R} = \mathfrak{R}(\Delta_1, \Delta_2)$.

### 2.2 Rate-distortion functions and conditional rate-distortion functions

A major tool in this study is rate-distortion theory. Specifically, joint, marginal, and conditional rate-distortion functions (or simply "rates") are used both in evaluations and bounds. These functions and their properties are dealt with in Refs. 1, 4, and 5. Here we only summarize some pertinent definitions and properties.

The marginal, joint, and conditional rates are defined as follows. Consider first the case where the alphabets $\mathcal{X}$, $\hat{\mathcal{X}}$, $\mathcal{Y}$, $\hat{\mathcal{Y}}$, are finite and $Q(x, y)$, $Q_X(x)$, $Q_Y(y)$ are probability functions. Then the (joint) rate-distortion function is defined by

$$R_{XY}(\Delta_1, \Delta_2) = \min I(XY; \hat{X}\hat{Y}), \tag{31}$$

where the random variables $\hat{X}\hat{Y}$ are defined by a "test-channel" $q_t(\hat{x}; \hat{y} | x, y)$—i.e., a probability function on $\hat{\mathcal{X}} \times \hat{\mathcal{Y}}$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The information in (31) is calculated for the joint distribution

$$\Pr\{X = x, Y = y, \hat{X} = \hat{x}, \hat{Y} = \hat{y}\} = Q(x, y) q_t(\hat{x}, \hat{y} | x, y). \tag{32}$$

The minimum in (31) is taken with respect to all test channels $q_t$ such that $Ed_1(X, \hat{X}) \leqq \Delta_1$, $Ed_2(Y, \hat{Y}) \leqq \Delta_2$, where the expectations are taken with respect to the distribution (32). The minimum always

exists. Similarly, the marginal rates are defined by

$$R_X(\Delta_1) = \min_{q_t(\hat{x}|x):Ed_1(x,\hat{x}) \leqq \Delta_1} I(X; \hat{X}), \tag{33a}$$

$$R_Y(\Delta_2) = \min_{q_t(\hat{y}|y):Ed(Y,\hat{Y}) \leqq \Delta_2} I(Y; \hat{Y}), \tag{33b}$$

where the expressions in (33) are interpreted analogously to that of (31). Detailed discussions of these quantities and their significance can be found in Refs. 1 and 4.

Another quantity that plays a crucial role in our study is the "conditional rate distortion function." Let $\mathfrak{X}$, $\mathfrak{Y}$ be finite, and let $Q(x, y)$ be given. Let $p(x, y, w)$ be a probability function on $\mathfrak{X} \times \mathfrak{Y} \times \mathfrak{W}$, where $\mathfrak{W}$ is a finite set such that $\sum_w p(x, y, w) = Q(x, y)$. Then $p(x, y, w)$ defines a triple of random variables $X$, $Y$, $W$, where the marginal distribution for $X$, $Y$ is $Q$. The conditional rate-distortion functions are defined as

$$R_{XY|W}(\Delta_1, \Delta_2) = \min I(X, Y; \hat{X}\hat{Y}|W), \tag{34}$$

where the minimum (which always exists) is taken with respect to all test channels $q_t(\hat{x}, \hat{y}|x, y, w)$ such that $Ed_1(X, \hat{X}) \leqq \Delta_1$, $Ed_2(Y, \hat{Y}) \leqq \Delta_2$. The conditional information in (34) is defined in Ref. 4, p. 21. The conditional rates $R_{X|W}(\Delta_1)$, $R_{Y|W}(\Delta_2)$ are defined analogously. A detailed discussion of conditional rates is given in Ref. 5. Of course, these definitions are meaningful if $X \equiv W$ or $Y \equiv W$. Roughly speaking, $R_{X,Y|W}(\Delta_1, \Delta_2)$ is the channel capacity required to transmit $X$, $Y$ and to reproduce it as $\hat{X}$, $\hat{Y}$ to within an average distortion $(\Delta_1, \Delta_2)$ when both the transmitter and receiver know $W$.

We shall need several properties of the conditional rate-distortion function in the sequel. The first is given in Ref. 5. For $\Delta \geqq 0$,

$$R_{X|W}(\Delta) = \min \sum_w \Pr\{W = w\} R_{X|W=w}(\Delta_w), \tag{35}$$

where $R_{X|W=w}(\cdot)$ is the rate-distortion function calculated for a source with outputs $x \in \mathfrak{X}$ with probability distribution $P_{X|W}(x|w)$ (the conditional probability function for $X$ given $W = w$). The minimum is taken over all sets $\{\Delta_w\}_{w \in \mathfrak{W}}$ such that $\sum_w \Pr\{W = w\} \Delta_w \leqq \Delta$.

A second fact of importance is that, say, $R_{X|W}(\Delta)$ is a continuous, convex, nonincreasing function of $\Delta$ for $\Delta \geqq 0$. That $R_{X|W}(\Delta)$ is nonincreasing follows from the definition. The proof that it is convex parallels the proof of the convexity of the ordinary rate-distortion function. The continuity of $R_{X|W}(\Delta)$, $\Delta > 0$ follows from its convexity.

Finally, the continuity of $R_{X|W}(\Delta)$ at $\Delta = 0$ follows from (35), and the continuity of $R_{X|W=w}(\Delta)$ at $\Delta = 0$.

A third fact we shall need is that, for any $X$, $W_1$, $W_2$, $\Delta \geqq 0$,

$$R_{X|W_1W_2}(\Delta) \leqq R_{X|W_1}(\Delta). \tag{36}$$

This follows from $R_{X|W_1W_2}(\Delta) = \inf I(X; \hat{X}|W_1W_2)$, where the infimum is with respect to test channels $q_t(\hat{x}|x, w_1, w_2)$ such that $Ed_1(X, \hat{X}) \leqq \Delta$. Included in this class of test channels are those that are independent of $w_2$, i.e., $q_t(\hat{x}|x, w_1, w_2) = q_t(\hat{x}|x_1, w_1)$. This subclass is exactly the class of test channels in the minimization for computing $R_{X|W_1}(\Delta)$.

The final property of conditional rates is stated as a lemma below. The proof is given in the appendix.

Let $X \in \mathfrak{X}$ be a random variable with probability distribution $Q_X(x) = \Pr(X = x)$, where $\mathfrak{X}$ is a finite source alphabet. Let $\hat{\mathfrak{X}}$ be a finite reproducing alphabet and let $d(x, \hat{x}) \geqq 0$, $x \in \mathfrak{X}$, $\hat{x} \in \hat{\mathfrak{X}}$ be a distortion function.

Now let $\{\mathfrak{W}_k\}_{k=1}^n$ be a family of disjoint finite sets and let $\{p_k(x, w)\}_{k=1}^n$ be a family of probability distributions on $\mathfrak{X} \times \mathfrak{W}_k$ such that

$$\sum_{w \in \mathfrak{W}_k} p_k(x, w) = Q_X(x).$$

The random pairs $(X, W_k)$ are defined by

$$\Pr\{X = x, W_k = w\} = p_k(x, w), \qquad x \in \mathfrak{X}, w \in \mathfrak{W}_k.$$

Let $R_{X|W_k}(\Delta)$, $\Delta \geqq 0$ be the corresponding conditional rate-distortion function.

Next, set $\mathfrak{W} = \sum_{k=1}^n \mathfrak{W}_k$, where $\sum$ indicates union of disjoint sets. Define the probability distribution on $\mathfrak{X} \times \mathfrak{W}$:

$$p^*(x, w) = \frac{1}{n} p_k(x, w), \qquad \text{for } w \in \mathfrak{W}_k, 1 \leqq k \leqq n,$$

and let $(X, W)$ be the corresponding random pair with conditional rate-distortion function $R_{X|W}(\Delta)$, $\Delta \geqq 0$. Clearly, $p^*(\cdot)$ is a mixture of the $n$ disjoint probability distributions $\{p_k\}$, with prior probability $1/n$. We now state the lemma.

*Lemma 5*: *For arbitrary* $\{\Delta_k\}_{k=1}^n$, $\Delta_k \geqq 0$,

$$R_{X|W}\left(\frac{1}{n}\sum_1^n \Delta_k\right) \leqq \frac{1}{n}\sum_{k=1}^n R_{X|W_k}(\Delta_k).$$

We note here that all the above is meaningful for the case where $Q(x, y)$ is a probability density function or $Q$ is an abstract probability measure. We need only make the obvious correspondences between discrete distributions and more general probability measures and replace "minimum" in (31), (33), (34), and (35) by "infimum."

We conclude this section by taking a look at the specialization of the above to the case where $d_1 = d_2 = d_H$, the Hamming distortion defined in (1b), and $\Delta_1 = \Delta_2 = 0$. Then

$$\begin{aligned} R_{XY}(0, 0) &= H(X, Y), \quad R_X(0) = H(X), \quad R_Y(0) = H(Y), \\ R_{XY|W}(0, 0) &= H(X, Y|W), \quad R_{X|W}(0) = H(X|W), \\ R_{Y|W}(0) &= H(Y|W), \end{aligned}$$

where the entropy $H(\cdot)$ and the conditional entropy $H(\cdot|\cdot)$ are defined in (2) and (3), respectively. Analogous to the relation

$$H(X|Y) + H(Y) = H(X, Y) \leqq H(X) + H(Y), \tag{37a}$$

which holds for this special case, the following is established in Ref. 5 for the general case:

$$R_{X|Y}(\Delta_1) + R_Y(\Delta_2) \leqq R_{XY}(\Delta_1, \Delta_2) \leqq R_X(\Delta_1) + R_Y(\Delta_2). \tag{37b}$$

Further, it is shown in Ref. 5, Corollary 3.2, that the left inequality in (37b) holds with equality in some neighborhood of the origin $\{(\Delta_1, \Delta_2): 0 \leqq \Delta_1, \Delta_2 \leqq \gamma\}$, provided that

$$Q(x, y) > 0, \quad \text{all } x \in \mathfrak{X}, y \in \mathfrak{Y}, \tag{38a}$$

and $d_1$, $d_2$ satisfy

$$\begin{aligned} d_1(x, \hat{x}) &> d_1(x, x) = 0, \quad x \neq \hat{x}, \\ d_2(y, \hat{y}) &> d_2(y, y) = 0, \quad y \neq \hat{y}. \end{aligned} \tag{38b}$$

### 2.3 *Characterization of $\mathcal{R}(\Delta_1, \Delta_2)$—the main result*

We first state two simple theorems that are generalizations of Theorems 2 and 3. The proofs are analogous to the proofs of Section I, and are therefore omitted. Theorem 6(a) is also called the Pangloss bound.

*Theorem 6*: *If* $(R_0, R_1, R_2) \in R(\Delta_1, \Delta_2)$, *then*

(*a*) $R_0 + R_1 + R_2 \geqq R_{XY}(\Delta_1, \Delta_2)$.

(*b*) $R_0 + R_1 \geqq R_X(\Delta_1)$.

(*c*) $R_0 + R_2 \geqq R_Y(\Delta_2)$.

*Theorem 7*: *The following triples belong to* $\mathcal{R}(\Delta_1, \Delta_2)$:

(*A*) $R_0 = R_{XY}(\Delta_1, \Delta_2), \quad R_1 = R_2 = 0.$

(*B*) $R_0 = 0, \quad R_1 = R_X(\Delta_1), \quad R_2 = R_Y(\Delta_2).$

It is also possible to generalize Theorem 3(*C*) and (*D*), but this must await presentation of the main result, which we now give.

Consider first the case where $\mathcal{X}$, $\mathcal{Y}$ are finite. Let $Q(x, y)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ be given. Now let $\mathcal{P}$ be the family of probability functions $p(x, y, w)$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $w \in \mathcal{W}$, $\mathcal{W}$ is another finite set, and

$$\sum_{w \in \mathcal{W}} p(x, y, w) = Q(x, y), \quad x \in \mathcal{X}, y \in \mathcal{Y}. \tag{39}$$

Thus, $\mathcal{P}$ is exactly as in Section 1.4. Now each $p \in \mathcal{P}$ defines three discrete random variables $X$, $Y$, $W$ in the obvious way. For $p \in \mathcal{P}$ and $\Delta_1, \Delta_2 \geqq 0$, define the subset of Euclidean three-space

$$\begin{aligned} \mathcal{R}_1^{(p)}(\Delta_1, \Delta_2) = \{(R_0, R_1, R_2) : R_0 &\geqq I(X, Y; W), \\ R_2 &\geqq R_{X|W}(\Delta_1), \qquad R_2 \geqq R_{Y|W}(\Delta_2)\}. \end{aligned} \tag{40a}$$

Then let

$$\mathcal{R}^*(\Delta_1, \Delta_2) = \Big[\bigcup_{p \in \mathcal{P}} \mathcal{R}^{(p)}(\Delta_1, \Delta_2)\Big]^c, \tag{40b}$$

where $(\ )^c$ denotes set closure. Since $R_{X|W}(\Delta_1)$ and $R_{Y|W}(\Delta_2)$ are continuous for $\Delta_1, \Delta_2 \geqq 0$, we conclude that $\mathcal{R}^*(\Delta_1, \Delta_2)$ is continuous in $(\Delta_1, \Delta_2)$ according to the Hausdorff set metric. This metric $\rho(S_1, S_2)$ between two subsets $S_1$, $S_2$ of a Euclidean space is defined by

$$\rho(S_1, S_2) = \sup_{r_1 \in S_1} \inf_{r_2 \in S_2} \|r_1 - r_2\| + \sup_{r_2 \in S_2} \inf_{r_1 \in S_1} \|r_1 - r_2\|,$$

where $\|\cdot\|$ denotes Euclidean norm.

If $Q$ is either a density or a probability measure, then $\mathcal{R}^*(\Delta_1, \Delta_2)$ can be defined in an analogous way. In this more general case, we must require that the source has the property that there exists an $\hat{x} \in \hat{\mathcal{X}}$, $\hat{y} \in \hat{\mathcal{Y}}$ such that

$$Ed_1(X, \hat{x}) < \infty, \quad Ed_2(Y, \hat{y}) < \infty. \tag{41}$$

If $\mathcal{X}$, $\mathcal{Y}$ are finite, then (41) is always satisfied. We can now state our main result.

*Theorem 8*: $\mathcal{R}(\Delta_1, \Delta_2) = \mathcal{R}^*(\Delta_1, \Delta_2)$.

*Remarks*:

(1) Theorem 8 reduces to Theorem 4 when $\mathcal{X}$, $\mathcal{Y}$ are finite, $d_1 = d_2 = d_H$, and $\Delta_1 = \Delta_2 = 0$.

(2) If we define $\mathcal{P}_T$ as in remark (1) following Theorem 4 as the set of test channels $p_t(w|x, y)$, then $\mathcal{P}_T$ is in 1-1 correspondence with $\mathcal{P}$.

(3) Since $\mathcal{R}(\Delta_1, \Delta_2)$ is convex, Theorem 8 implies that $\mathcal{R}^*(\Delta_1, \Delta_2)$ is convex also.

(4) Since Theorem 1 is valid for $\mathcal{R}(\Delta_1, \Delta_2)$, the present theorem implies that $T(\alpha)$, defined in (13) is also given by

$$T(\alpha) = \inf_{p \in \mathcal{P}} [I(X, Y; W) + \alpha_1 R_{X|W}(\Delta_1) + \alpha_2 R_{Y|W}(\Delta_2)]. \quad (42)$$

Thus, from Theorem 1, the lower boundary $\overline{\mathcal{R}}(\Delta_1, \Delta_2)$, and therefore $\mathcal{R}(\Delta_1, \Delta_2)$, is determined by $T(\alpha)$ given in (42).

(5) As in remark (4) after Theorem 4, Theorems 6 and 7 can be obtained directly from Theorem 8. The steps parallel those in remark (4) and will be omitted. We will, however, give the generalization of Theorem 3(*C*) and (*D*). We state this as follows. The following triples $(R_0, R_1, R_2) \in \mathcal{R}(\Delta_1, \Delta_2)$:

(*C*) $R_0 = R_Y(\Delta_2), \quad R_1 = R_{X|\hat{Y}}(\Delta_1), \quad R_2 = 0,$

(*D*) $R_0 = R_X(\Delta_1), \quad R_1 = 0, \quad R_2 = R_{Y|\hat{X}}(\Delta_2),$

where the random variable $\hat{Y}$ is defined by the test channel that achieves the infimum in $R_Y(\Delta_2)$ (assuming that the infimum can be achieved; if not, a simple modification is possible), and $\hat{X}$ is defined by the test channel that achieves $R_X(\Delta_1)$. In the discrete case, we can achieve point (*C*) as follows. Let $p_t^*(\mathcal{Y}|y)$ be the test channel that achieves $I(Y; \hat{Y}) = R_Y(\Delta_2)$. Let $\mathcal{W} = \hat{\mathcal{Y}}$ and let

$$p(x, y, \mathcal{Y}) = Q(x, y) p_t^*(\mathcal{Y}|y) \in \mathcal{P}.$$

The random variables $X$, $Y$, $\hat{Y}$ are defined in an obvious way by $p(x, y, \mathcal{Y})$. Further, since $X$, $\hat{Y}$ are conditionally independent given $Y$,

$$\begin{aligned} I(X, Y; \hat{Y}) &= I(Y; \hat{Y}) + I(X; \hat{Y}|Y) \\ &= I(Y; \hat{Y}) = R_Y(\Delta_2). \end{aligned}$$

Also, the conditional rate $R_{Y|\hat{Y}}(\Delta_2) = 0$. Thus, from Theorem 8, with $W = \hat{Y}$, we have $(R_0, R_1, R_2) \in \mathcal{R}(\Delta_1, \Delta_2)$ where

$$\begin{aligned} R_0 &= I(X, Y; W) = R_Y(\Delta_2) \\ R_1 &= R_{X|W}(\Delta_1) = R_{X|\hat{Y}}(\Delta_1), \\ R_2 &= R_{Y|W}(\Delta_2) = 0. \end{aligned}$$

This is point (*C*). Point (*D*) is obtained on reversing the roles of $X$ and $Y$.

Since $\mathfrak{R}(\Delta_1, \Delta_2)$ is convex, any linear combination of points $(A)$ and $(B)$ of Theorem 7 and $(C)$ and $(D)$ above also belongs to $\mathfrak{R}(\Delta_1, \Delta_2)$. But there is no guarantee in this case that points $(C)$ and $(D)$ will lie on the Pangloss plane. There are cases for which a portion of the Pangloss plane is known to be realizable, as is shown in the example below.

### 2.4 A technique for overbounding $\overline{\mathfrak{R}}(\Delta_1, \Delta_2)$

In this section we present an intuitively sensible ad hoc scheme for choosing probability distributions $p \in \mathcal{P}$ that yield triples $[I(X, Y; W), R_{X|W}(\Delta_1), R_{Y|W}(\Delta_2)]$ which are often close to or actually on the boundary curve $\overline{\mathfrak{R}}(\Delta_1, \Delta_2)$. In fact, in many cases this triple will lie on the Pangloss plane.

A natural coding scheme to apply to our network would be to send a "coarse" version of the source output $(\mathbf{X}, \mathbf{Y})$ over the common channel, and then send to each receiver over its private channel only the necessary "fine tuning" it needs to meet its fidelity requirement. This reasoning leads us to the following family of rate triples that belong to $\mathfrak{R}(\Delta_1, \Delta_2)$. Assume for simplicity that $\mathfrak{X}$, $\mathfrak{Y}$, $\hat{\mathfrak{X}}$, $\hat{\mathfrak{Y}}$ are finite.

Let $\Delta_1, \Delta_2 \geqq 0$ be given. Let $\beta_1, \beta_2$ satisfy

$$\beta_1 \geqq \Delta_1, \quad \beta_2 \geqq \Delta_2.$$

Now let $q_t(\tilde{x}, \tilde{y} | x, y)$ be the test channel that achieves $I(X, Y; \tilde{X}, \tilde{Y}) = R_{XY}(\beta_1, \beta_2)$. Then with $W = (\tilde{X}, \tilde{Y})$ we have that the triple $(R_0, R_1, R_2) \in \mathfrak{R}(\Delta_1, \Delta_2)$, where

$$R_0 = I(X, Y; W) = R_{XY}(\beta_1, \beta_2),$$

and

$$R_1 = R_{X|\tilde{X}\tilde{Y}}(\Delta_1), \quad R_2 = R_{Y|\tilde{X}\tilde{Y}}(\Delta_2). \tag{43}$$

Note that the rates corresponding to Theorem 7 $(A)$ and $(B)$ and to points $(C)$ and $(D)$ in remark (5) following Theorem 8 can be generated as special cases of the rate in (43). We do this as follows:

$A$: Let $(\beta_1, \beta_2) = (\Delta_1, \Delta_2)$.

$B$: Let $\beta_1, \beta_2$ be large enough so that $R_{XY}(\beta_1, \beta_2) = 0$. Then $\tilde{X}$, $\tilde{Y}$ are degenerate.

$C$: Let $\beta_1$ be large enough so that $R_X(\beta_1) = 0$, and let $\beta_2 = \Delta_2$. Then $\tilde{X}$ is degenerate.

$D$: Let $\beta_2$ be large enough so that $R_Y(\beta_2) = 0$, and let $\beta_1 = \Delta_1$.

The power of this technique is illustrated by the following theorem, which asserts that under weak assumption the family of rates given

by (43) includes a substantial subfamily that lies on the Pangloss bound and therefore on the boundary.

*Theorem 9: Given a source that satisfies*

(*i*) $\mathfrak{X} = \hat{\mathfrak{X}}$, $\mathcal{Y} = \hat{\mathcal{Y}}$, $\mathfrak{X}$, $\mathcal{Y}$ *finite,*

(*ii*) $Q(x, y) > 0$, *all* $x \in \mathfrak{X}$, $y \in \mathcal{Y}$,

(*iii*) $d_1(x, \hat{x}) > d_1(x, x) = 0$, *all distinct* $x$, $\hat{x} \in \mathfrak{X}$, *and* $d_2(y, \hat{y}) > d_2(y, y) = 0$, *all distinct* $y$, $\hat{y} \in \mathcal{Y}$.

*Then there exists two neighborhoods of the origin*

$$\eta_1 = \{(\Delta_1, \Delta_2) : 0 \leqq \Delta_1, \Delta_2 \leqq a\}$$
$$\eta_2 = \{(\beta_1, \beta_2) : 0 \leqq \beta_1, \beta_2 \leqq b\},$$

*where* $0 < a \leqq b$, *such that, if* $(\Delta_1, \Delta_2) \in \eta_1$ *and* $(\beta_1, \beta_2) \in \eta_2$, *then*

$$R_0 + R_1 + R_2 = R_{XY}(\Delta_1, \Delta_2),$$

*where* $(R_0, R_1, R_2)$ *is given by* (43).

The theorem can be proved using Shannon lower-bound techniques[1,5] and, in particular, the proof is similar to that of Theorem 32 in Ref. 5. Since the proof requires the generation of special machinery that is only tangential to the main ideas in this paper, we have elected to omit it.

### 2.5 Examples

(*A*) Our first example will be the DSBS considered in the example of Section 1.5. Here $\mathfrak{X} = \mathcal{Y} = \{0, 1\}$, and

$$Q(x, y) = \tfrac{1}{2}(1 - p_0)\delta_{x,y} + \tfrac{1}{2}p_0(1 - \delta_{x,y}), \quad x, y = 0, 1, \tag{44}$$

where the parameter $p_0 \in [0, \frac{1}{2}]$. The distortion function will be the Hamming metric, i.e., $d_1 = d_2 = d_H$, where $d_H$ is defined in (1b). Again, as in Section 1.4, we consider only the plane in $(R_0, R_1, R_2)$-space where $R_1 = R_2$ and $\Delta_1 = \Delta_2 = \Delta$. We employ the technique of Section 2.4 to obtain an upper bound for $\overline{\mathfrak{R}}(\Delta, \Delta)$.

Making use of Ref. 1, pp. 46–50 (Ex. 2.7.2), we have

$$R_{XY}(\beta, \beta) = \begin{cases} 1 + h(p_0) - 2h(\beta), & 0 \leqq \beta \leqq p_1 \\ L(1 - p_0) - \tfrac{1}{2}\{L(2\beta - p_0) + L[2(1 - \beta) - p_0]\}, & p_1 \leqq \beta \leqq \tfrac{1}{2} \end{cases} \tag{45a}$$

where

$$p_1 = \tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - 2p_0}, \tag{45b}$$

$$h(\lambda) = \lambda \log \lambda - (1 - \lambda) \log (1 - \lambda), \quad 0 \leqq \lambda \leqq 1, \tag{45c}$$

$$L(\lambda) = -\lambda \log \lambda, \quad 0 \leqq \lambda \leqq 1. \tag{45d}$$

Now, from Ref. 1, the random variables $\tilde{X}$ and $\tilde{Y}$, which satisfy $I(X, Y; \tilde{X}\tilde{Y}) = R_{XY}(\beta, \beta)$ are such that

$$\begin{aligned}\Pr\{X = x | \tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}\} &= \Pr\{X = x | \tilde{X} = \tilde{x}\} \\ &= (1 - \beta)\delta_{x,\tilde{x}} + \beta(1 - \delta_{x,\tilde{x}}), \quad x, \tilde{x}, \tilde{y} = 0, 1\end{aligned} \tag{46a}$$

and

$$\begin{aligned}\Pr\{Y = y | \tilde{X} = \tilde{x}; \tilde{Y} = \tilde{y}\} &= \Pr\{Y = y | Y = \tilde{y}\} \\ &= (1 - \beta)\delta_{y,\tilde{y}} + \beta(1 - \delta_{y,\tilde{y}}), \quad y, \tilde{y}, \tilde{x} = 0, 1.\end{aligned} \tag{46b}$$

Thus, again from Ref. 1 (p. 46, Ex. 2.7.1), for $0 \leqq \Delta \leqq \beta$,

$$\begin{aligned}R_{X|\tilde{X}\tilde{Y}}(\Delta) &= R_{X|\tilde{X}}(\Delta) = R_{Y|\tilde{X}\tilde{Y}}(\Delta) = R_{Y|\tilde{Y}}(\Delta) \\ &= h(\beta) - h(\Delta).\end{aligned} \tag{47}$$

Thus, we conclude that, for arbitrary $0 \leqq \Delta \leqq \beta \leqq \frac{1}{2}$, the triple $(R_0, R_1, R_2) \in \mathcal{R}(\Delta, \Delta)$, where $R_0 = R_{XY}(\beta, \beta)$ [as in (45)], and $R_1 = R_2 = h(\beta) - h(\Delta)$. Let us note that, for $0 \leqq \Delta \leqq \beta \leqq p_1$, these rate-triples $(R_0, R_1, R_2)$ satisfy

$$R_0 + R_1 + R_2 = 1 + h(p_0) - 2h(\Delta) = R_{XY}(\Delta, \Delta), \tag{48}$$

and therefore lie on the Pangloss plane and $\overline{\mathcal{R}}(\Delta, \Delta)$. One special case occurs when $\Delta = 0$, $\beta = p_1$. This yields the rate-triple of (24)—i.e., point $G$ in Fig. 4. In fact, the distribution $p(x, y, w) \in \mathcal{P}$, which we guessed at in Section 1.5, was obtained by setting $W = (\tilde{X}, \tilde{Y})$, where $\tilde{X}$, $\tilde{Y}$ are as above for $\beta \geqq p_1$.

(*B*) Our second example is a source where $Q(x, y)$ is a density function and $\mathfrak{X}$, $\hat{\mathfrak{X}}$, $\mathfrak{Y}$, $\hat{\mathfrak{Y}}$ are the reals. The ad hoc technique used in the previous example (*A*) will work here with obvious modifications. The random variables $X$, $Y$ in this case will be jointly gaussian with $EX = EY = 0$, $EX^2 = EY^2 = 1$, and $EXY = r$, $0 \leqq r \leqq 1$. Thus, the density

$$Q(x, y) = \frac{1}{2\pi(1 - r^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x^2 + y^2 - 2rxy)}{2(1 - r^2)}\right\}. \tag{49}$$

We take the distortion to be $d_1(\cdot, \cdot) = d_2(\cdot, \cdot)$, where

$$d_1(x, \hat{x}) = (x - \hat{x})^2, \quad -\infty < x, \hat{x} < \infty.$$

For $0 < \beta < \infty$, it can be shown[1,4] that

$$R_{XY}(\beta, \beta) = \begin{cases} \frac{1}{2}\log\left(\frac{1-r^2}{\beta^2}\right), & 0 \leqq \beta \leqq 1-r, \\ \frac{1}{2}\log\left(\frac{1+r}{2\beta-(1-r)}\right), & 1-r \leqq \beta \leqq 1, \\ 0, & \beta \geqq 1. \end{cases} \tag{50}$$

Further, the random variables $\tilde{X}$, $\tilde{Y}$ which satisfy $I(X, Y; \tilde{X}, \tilde{Y}) = R_{XY}(\beta, \beta)$, $0 < \beta \leqq 1$ are such that, given $\tilde{X} = \tilde{x}$, $\tilde{Y} = \tilde{y}$, $X$ and $Y$ are gaussian with

$$\begin{aligned} E(X|\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) &= \tilde{x}, \\ E(Y|\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) &= \tilde{y}, \\ \text{var}\,(X|\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) &= \text{var}\,(Y|\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) = \beta. \end{aligned}$$

Thus,[1,4] for $0 < \Delta \leqq \beta < 1$,

$$R_{X|\tilde{X}\tilde{Y}}(\Delta) = R_{Y|\tilde{X}\tilde{Y}}(\Delta) = \frac{1}{2}\log\frac{\beta}{\Delta}.$$

Thus, we conclude that, for arbitrary $0 < \Delta \leqq \beta \leqq 1$, the triple $(R_0, R_1, R_2) \in \mathfrak{R}(\Delta, \Delta)$, where $R_0 = R_{XY}(\beta, \beta)$ [as in (50)] and $R_1 = R_2 = \frac{1}{2}\log \beta/\Delta$. Again, observe that for $0 \leqq \Delta \leqq \beta \leqq 1 - r$,

$$R_0 + R_1 + R_2 = \frac{1}{2}\log\left(\frac{1-r^2}{\Delta}\right) = R_{XY}(\Delta, \Delta), \tag{51}$$

and therefore $(R_0, R_1, R_2)$ lies on the Pangloss plane and therefore on $\overline{\mathfrak{R}}(\Delta, \Delta)$.

## III. PROOF OF THE MAIN RESULT—THEOREM 8

The proof of Theorem 8 consists of two parts: (*i*) the "converse" part, which asserts that any point in $\mathfrak{R}(\Delta_1, \Delta_2)$ belongs to $\mathfrak{R}^*(\Delta_1, \Delta_2)$ and (*ii*) the "direct" or "positive" part, which asserts that any point in $\mathfrak{R}^*(\Delta_1, \Delta_2)$ belongs to $\mathfrak{R}(\Delta_1, \Delta_2)$. We give the proof for the case where $\mathfrak{X}$, $\mathfrak{Y}$ are finite sets. The proof for arbitrary $\mathfrak{X}$, $\mathfrak{Y}$ follows in a parallel way with integrals replacing sums, etc., in the standard way. We will begin with the converse.

### 3.1 The converse

Let $(f_E, f_D^{(X)}, f_D^{(Y)})$ define a code $(n, M_0, M_1, M_2, \Delta_X, \Delta_Y)$. We find a $p^*(x; y, w) \in \mathcal{P}$ for an appropriate set $\mathcal{W}$ such that

$$\left(\frac{1}{n}\log M_0, \frac{1}{n}\log M_1, \frac{1}{n}\log M_2\right) \in \mathfrak{R}^{(p^*)}(\Delta_X, \Delta_Y) \subseteq \mathfrak{R}^*(\Delta_X, \Delta_Y). \tag{52}$$

The converse follows on applying the definition of $(\Delta_1, \Delta_2)$-achievable rates and applying the continuity of $\mathfrak{R}^*$ as discussed in Section II.

First, let $f_E(\mathbf{X}, \mathbf{Y}) = (S_0, S_1, S_2)$ where $S_i \in I_{M_i}$ is a random variable $(i = 0, 1, 2)$. Then we have

$$\frac{1}{n}\log M_0 \overset{(1)}{\geqq} \frac{1}{n} H(S_0) \overset{(2)}{\geqq} \frac{1}{n} I(\mathbf{X}, \mathbf{Y}; S_0) \overset{(3)}{=} \frac{1}{n}[H(\mathbf{X}, \mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}|S_0)]$$
$$\overset{(4)}{=} \frac{1}{n}\sum_{k=1}^{n}[H(X_k, Y_k) - H(X_k, Y_k|S_0, X_1, \cdots, X_{k-1}, Y_1, \cdots, Y_{k-1})]. \tag{53}$$

These steps are justified as follows:

(1) From $S_0 \in I_{M_0}$.
(2) Standard inequality.
(3) Definition of $I(\mathbf{X}, \mathbf{Y}; S_0)$.
(4) $H(\mathbf{X}, \mathbf{Y}) = \sum_k H(X_k, Y_k)$ follows from the independence of the pairs $(X_k, Y_k)$, $k = 1, 2, \cdots, n$. The rest is also a standard identity.

Now, for $1 \leqq k \leqq n$, let $W_k = (S_0, X_1, \cdots, X_{k-1}, Y_1, \cdots, Y_{k-1})$, a random variable belonging to, say, $\mathcal{W}_k$, a finite set.† Relation (53) is then

$$\frac{1}{n}\log M_0 \geqq \frac{1}{n}\sum_{k=1}^{n} I(X_k, Y_k; W_k). \tag{54}$$

Next, let $\hat{\mathbf{X}} = f_D^{(X)} \circ f_E(\mathbf{X}, \mathbf{Y})$. Let $\Delta_{1k} = Ed_1(X_k, \hat{X}_k)$, $1 \leqq k \leqq n$. Then

$$\Delta_X = ED_1(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{n}\sum_{k=1} \Delta_{1k}. \tag{55a}$$

We now write

$$\frac{1}{n}\log M_1 \overset{(1)}{\geqq} \frac{1}{n} H(\hat{\mathbf{X}}|S_0) \overset{(2)}{\geqq} \frac{1}{n} I(\mathbf{X}; \hat{\mathbf{X}}|S_0)$$
$$\overset{(3)}{=} \frac{1}{n}\sum_{k=1}^{n} I(X_k; \hat{\mathbf{X}}|S_0, X_1, \cdots, X_{k-1})$$
$$\overset{(4)}{\geqq} \frac{1}{n}\sum_{k=1}^{n} I(X_k; \hat{X}_k|S_0, X_1, \cdots, X_{k-1})$$
$$\overset{(5)}{\geqq} \frac{1}{n}\sum_{k=1}^{n} R_{X_k|V_k}(\Delta_{1k}) \overset{(6)}{\geqq} \frac{1}{n}\sum_{k=1}^{n} R_{X_k|W_k}(\Delta_{1k}), \tag{55b}$$

where $V_k = (S_0, X_1, \cdots, X_{k-1})$, and $W_k = (V_k, Y_1, \cdots, Y_{k-1})$ as above. These steps are justified as follows:

† We can, of course, take $\mathcal{W}_k = I_{M_0} \times \mathcal{X}^{k-1} \times \mathcal{Y}^{k-1}$.

(1) Since $\hat{X}$ is a function of $S_0$ and $S_1$, we have that, conditioned on $S_0 = s_0$, $\hat{X}$ can take no more than $M_1$ values (since $S_1 \in I_{M_1}$). Thus, $H(\hat{\mathbf{X}} | S_0 = s_0) \leqq \log M_1$, all $s_0$.

(2–4) Standard inequalities and identities.

(5) From the definition of $\mathcal{R}_{X_k|Y_k}(\Delta_{1k})$, since $Ed_1(X_k, \hat{X}_k) = \Delta_{1k}$.

(6) Follows from (36).

A similar derivation yields

$$\frac{1}{n} \log M_2 \geqq \frac{1}{n} \sum_{k=1}^{n} R_{Y_k|W_k}(\Delta_{2k}), \tag{56a}$$

where

$$\Delta_Y = \frac{1}{n} \sum_{k=1}^{n} \Delta_{2k}. \tag{56b}$$

We are now in a position to define $p^*(x, y, w)$. With $W_k$ defined as above, let

$$p_k(x, y, w) = \Pr\{X_k = x, Y_k = y, W_k = w\},$$
$$x \in \mathfrak{X}, y \in \mathfrak{Y}, w \in \mathfrak{W}_k.$$

Let

$$p_{1k}(x, w) = \sum_y p_k(x, y, w),$$
$$p_{2k}(y, w) = \sum_x p_k(x, y, w)$$

be the marginal distributions for $(X_k, W_k)$ and $(Y_k, W_k)$, respectively. The $\{\mathfrak{W}_k\}$ can be considered a class of disjoint sets. Let $\mathfrak{W} = \Sigma \mathfrak{W}_k$, and define the probability function on $\mathfrak{X} \times \mathfrak{Y} \times \mathfrak{W}$

$$p^*(x, y, w) = \frac{1}{n} p_k(x, y, w), \quad w \in \mathfrak{W}_k, \quad 1 \leqq k \leqq n.$$

Since

$$\sum_{w \in \mathfrak{W}} p^*(x, y, w) = \sum_{k=1}^{n} \sum_{w \in \mathfrak{W}_k} \frac{1}{n} p_k(x, y, w) = Q(x, y),$$

we have $p^* \in \mathcal{P}$. The random variables $X$, $Y$, $W$ are defined by $p^*$ in the obvious way. We can think of $W$ as being generated by choosing an integer $K \in [1, n]$ without bias, and setting $W = W_k$ when $K = k$, $1 \leqq k \leqq n$. A straightforward calculation yields

$$I(X,Y; W) = \frac{1}{n} \sum_{k=1}^{n} I(X, Y; W_k). \tag{57a}$$

Furthermore, Lemma 5 can be applied to $p_{1k}$, $p_{2k}$ to yield

$$R_{X|W}(\Delta_X) \leqq \frac{1}{n}\sum_{k=1}^{n} R_{X_k|W_k}(\Delta_k), \tag{57b}$$

$$R_{Y|W}(\Delta_Y) \leqq \frac{1}{n}\sum_{k=1}^{n} R_{Y_k|W_k}(\Delta_k). \tag{57c}$$

Inequalities (57a, b, c) can be substituted into (54), (55), (56), respectively, to obtain

$$\frac{1}{n}\log M_0 \geqq I(X, Y; W)$$

$$\frac{1}{n}\log M_1 \geqq R_{X|W}(\Delta_X)$$

$$\frac{1}{n}\log M_2 \geqq R_{Y|W}(\Delta_Y),$$

which is (52). This completes the proof of the converse.

### 3.2 The direct half

We begin the proof by stating a lemma concerning conventional source coding for a single memoryless source. The source is defined by a random variable $X \in \mathfrak{X}$, with probability distributions $Q_X(x)$, and a reproducing alphabet $\hat{\mathfrak{X}}$ with distortion function $d_1(x, \hat{x})$. As above, $\mathbf{X} = (X_1, \cdots, X_n)$ are $n$ independent copies of $X$. Let $Q_X^{(n)}(\mathbf{x}) = \Pi_{k=1}^{n} Q_X(x_k)$ be the probability distribution for $\mathbf{X}$. Let $R(\Delta)$ be the rate-distortion function.

A source *code* with parameters $(n, M)$ may be thought of as a mapping $f: \mathfrak{X}^n \to \mathcal{C} \subseteq \hat{\mathfrak{X}}^n$, where card $\mathcal{C} \leqq M$. Let $\hat{\mathbf{X}} = (\hat{X}_1, \cdots, \hat{X}_n) = f(\mathbf{X})$. Then $D_1(\mathbf{X}, \hat{\mathbf{X}}) = 1/n\sum_{k=1}^{n} d_1(X_k, \hat{X}_k)$ is a random variable. We are interested in the quantity

$$\Gamma(\Delta_1 + \delta) = \Pr\{D_1(\mathbf{X}, \hat{\mathbf{X}}) \geqq \Delta + \delta\} = \sum_{\mathbf{x} \in \mathfrak{X}^n} Q_X^{(n)}(\mathbf{x})\Phi(\mathbf{x}), \tag{58}$$

where $\Phi(\mathbf{x}) = 1$, if $D[\mathbf{x}, f(\mathbf{x})] \geqq \Delta + \delta$, and $\Phi(\mathbf{x}) = 0$, otherwise. We now state a lemma, which follows immediately from Lemma 9.3.1 and inequality (9.3.31) of Gallager.[4]

*Lemma 10*: *Let* $\Delta \geqq 0$, *and* $\epsilon$, $\delta > 0$ *be arbitrary. Then there exist* $A$, $B > 0$ *such that for all* $n = 1, 2, \cdots$ *there exists a code with parameters* $(n, M)$ *satisfying*

$$M \leqq 2^{n[R(\Delta)+\epsilon]},$$

*and*

$$\Gamma(\Delta + \delta) = \Pr\{D_1(\mathbf{X}, \hat{\mathbf{X}}) \geqq \Delta + \delta\} \leqq Ae^{-Bn}.$$

With the aid of this lemma the standard source coding theorem follows readily (Ref. 4, Theorem 9.3.1).

Next, let us consider a compound source for which the source output in $n$ time units is an $n$-vector $\mathbf{X} = (X_1, \cdots, X_n) \in \mathfrak{X}^n$. The $\{X_k\}$ are still independent, but the $X_k$ are not identically distributed.

Let $n_1, n_2, \cdots, n_J$ be such that $\sum_{j=1}^{J} n_j = n$, and let $Q_1(\cdot)$, $Q_2(\cdot), \cdots, Q_J(\cdot)$ be $J$ probability distribution functions on $X$. The source is characterized by the fact that a known subset $n_j$ of the $n$ coordinates of $\mathbf{X}$ are distributed according to $Q_j(\cdot)$, $j = 1, \cdots, J$. Let $R_j(\Delta)$ be the rate-distortion function corresponding to $Q_j(\cdot)$ relative to the distortion function $d_1$. A code is defined exactly as above, and $\hat{\mathbf{X}} = f(\mathbf{X})$. We now have

*Corollary 11*: *Let* $\Delta_j \geqq 0$, $j = 1, 2, \cdots, J$, *and* $\epsilon$, $\delta > 0$ *be arbitrary. Then there exist* $A_j, B_j > 0, j = 1, \cdots, J$, *such that, for all* $n = 1, 2, \cdots$ *and any set* $\{n_j\}_1^J$ *such that* $\Sigma n_j = n$, *there exists a code with parameter* $M$ *satisfying*

$$M \leqq \prod_{j=0}^{J-1} \exp_2 \{n_j[R_j(\Delta_j) + \epsilon]\} = \exp_2 \{\sum n_j[R_j(\Delta_j) + \epsilon]\} \quad (59a)$$

*and*

$$\Gamma(\Delta + \delta) = \Pr \{D_1(\mathbf{X}, \hat{\mathbf{X}}) \geqq \Delta + \delta\} \leqq \sum_{j=0}^{J-1} A_j 2^{-B_j n_j}, \quad (59b)$$

*where* $\Delta = n^{-1}\sum n_j\Delta_j$. *The* $(A_j, B_j)$*'s are the* $(A, B)$ *of Lemma 10 corresponding to* $Q_j(\cdot)$.

The corollary follows immediately from Lemma 10 on noting that, for any random variables $\{U_j\}$ and any set of constants $\{c_j\}$,

$$\Pr \{\sum_j U_j \geqq \sum_j c_j\} \leqq \sum_j \Pr \{U_j \geqq c_j\}.$$

Let us also remark that the $Q^{(n)}(\mathbf{x})$ used to compute $\Gamma(\Delta + \delta)$ in the corollary is of the form

$$Q_X^{(n)}(\mathbf{x}) = \prod_{k=1}^{n_1} Q_1(x_{i_{1k}}) \prod_{k=1}^{n_2} Q_2(x_{i_{2k}}) \cdots \prod_{k=1}^{n_J} Q_J(x_{i_{Jk}}), \quad (60)$$

where the $i_{jk}$th coordinate of $\mathbf{x}$ has distribution $Q_j(\cdot)$, $1 \leqq k \leqq n_j$, $0 \leqq j \leqq J - 1$.

Let us now turn to our network coding problem. An alternative (though equivalent) way of defining a code for our network with

parameters $(n, M_0, M_1, M_2)$ is

(1) A mapping

$$g: \mathfrak{X}^n \times \mathfrak{Y}^n \to \mathfrak{C}, \tag{61a}$$

where $\mathfrak{C}$ is an arbitrary set with cardinality $\leqq M_0$. The mapping $g$ is called a "core code."

(2) For each $\mathbf{w} \in \mathfrak{C}$, a mapping

$$g_{\mathbf{w}}^{(X)}: \mathfrak{X}^n \times \mathfrak{Y}^n \to \mathfrak{C}_{\mathbf{w}}^{(X)} \subseteq \hat{\mathfrak{X}}^n, \tag{61b}$$

where card $\mathfrak{C}_{\mathbf{w}}^{(X)} \leqq M_1$.

(3) For each $\mathbf{w} \in \mathfrak{C}$, a mapping

$$g_{\mathbf{w}}^{(Y)}: \mathfrak{X}^n \times \mathfrak{Y}^n \to \mathfrak{C}_{\mathbf{w}}^{(Y)} \subseteq \hat{\mathfrak{Y}}^n, \tag{61c}$$

where card $\mathfrak{C}_{\mathbf{w}}^{(Y)} \leqq M_2$.

The code defined in this way can be used on our network (Fig. 2) as follows. Let $\mathfrak{C} = \{\mathbf{w}_i\}_1^{M_0}$. Then, if $g(\mathbf{x}, \mathbf{y}) = \mathbf{w}_i$, the index $i$ is transmitted over the common channel. Let $\mathfrak{C}_{\mathbf{w}_i}^{(X)} = \{\hat{\mathbf{x}}_{il}\}_{l=1}^{M_1}$, $1 \leqq i \leqq M_0$. Then, if $g(\mathbf{x}, \mathbf{y}) = \mathbf{w}_i$, and $g_{\mathbf{w}_i}^{(X)}(\mathbf{x}, \mathbf{y}) = \hat{\mathbf{x}}_{il}$, we transmit the index $l$ over the private channel to receiver 1. The decoder at receiver 1, knowing the indices $i$ and $l$, emits $\hat{\mathbf{x}}_{il}$, and the resulting distortion is $D_1(\mathbf{x}, \hat{\mathbf{x}}_{il})$. Receiver 2 works analogously.

Let us fix our attention on receiver 1, and assume that $g_{\mathbf{w}_i}^{(X)}(\mathbf{x}, \mathbf{y}) = g_{\mathbf{w}_i}^{(X)}(\mathbf{x})$. Then we define the quantity $q(\mathbf{x}, \mathbf{w}_i)(\mathbf{x} \in \mathfrak{X}^n, \mathbf{w}_i \in \mathfrak{C})$ as the probability that $\mathbf{X} = \mathbf{x} \in \mathfrak{X}^n$ and $\mathbf{Y} = \mathbf{y}$ such that $g(\mathbf{x}, \mathbf{y}) = \mathbf{w}_i$. Thus,

$$q(\mathbf{x}, \mathbf{w}_i) = \sum_{\mathbf{y}: g(\mathbf{x}, \mathbf{y}) = \mathbf{w}_i} Q^{(n)}(\mathbf{x}, \mathbf{y}). \tag{62}$$

$Q^{(n)}(\mathbf{x}, \mathbf{y}) \triangleq \prod_{k=1}^n Q(x_k, y_k)$ is the probability distribution for $\mathbf{X}$ and $\mathbf{Y}$.

Then, as in (58) with $\hat{X} = g_{\mathbf{W}}^{(X)}(\mathbf{X})$, $\mathbf{W} = g(\mathbf{X}, \mathbf{Y})$,

$$\Gamma(\Delta_1 + \delta) \triangleq \Pr\{D_1(\mathbf{X}, \hat{\mathbf{X}}) > \Delta_1 + \delta\} = \sum_{i=1}^{M_0} \sum_{\mathbf{x}} q(\mathbf{x}, \mathbf{w}_i)\Phi_i(\mathbf{x}), \tag{63a}$$

where

$$\Phi_i(\mathbf{x}) = \begin{cases} 1, \text{ if } D_1[\mathbf{x}, g_{\mathbf{w}_i}^{(X)}(\mathbf{x})] > \Delta_1 + \delta, \\ 0, \text{ otherwise.} \end{cases} \tag{63b}$$

Substituting (62) into (63), we obtain

$$\Gamma(\Delta_1 + \delta) = \sum_{i=1}^{M_0} \{ \sum_{(x,y) \in G_i} Q^{(n)}(\mathbf{x}, \mathbf{y})\Phi_i(\mathbf{x}) \}, \tag{64a}$$

where

$$G_i = \{(\mathbf{x}, \mathbf{y}) : g(\mathbf{x}, \mathbf{y}) = \mathbf{w}_i\}, \quad 1 \leqq i \leqq M_0. \tag{64b}$$

Now our goal is to show that there exists a code for $n$ sufficiently large, with $M_0$, $M_1$, $M_2$ appropriately chosen, and with $\Gamma(\Delta_1 + \delta)$ arbitrarily small.

Let us assume that we are given a probability distribution $p(x, y, w) \in \mathcal{P}$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $w \in \mathcal{W}$. Let $p_W(w) = \sum_{x,y} p(x, y, w)$, $w \in \mathcal{W}$ be the marginal distribution of $W$. Assume with no loss of generality that $p_W(w) > 0$. Let

$$p_b(x, y \mid w) = \frac{p(x, y, w)}{p_W(w)}$$

be the "backward test channel." For $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{y} \in \mathcal{Y}^n$, $\mathbf{w} \in \mathcal{W}^n$, let

$$p^{(n)}(\mathbf{x}, \mathbf{y}, \mathbf{w}) = \Pi_{k=1}^n p(x_k, y_k, w_k)$$

be the probability distribution for $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{W}$ ($n$ independent drawings of $X$, $Y$, $W$). Let $p_W^{(n)}(\mathbf{w}) = \prod_{k=1}^n p_W(w_k)$, and $p_b^{(n)}(\mathbf{x}, \mathbf{y} \mid \mathbf{w}) = \prod_{k=1}^n p_b(x_k, y_k \mid w_k)$. For $(\mathbf{x}, \mathbf{y}, \mathbf{w}) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{W}^n$, let

$$i^{(n)}(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \log \frac{p_b^{(n)}(\mathbf{x}, \mathbf{y} \mid \mathbf{w})}{Q^{(n)}(\mathbf{x}, \mathbf{y})} = \sum_{k=1}^n \log \frac{p_b(x_k, y_k \mid w_k)}{Q(x_k, y_k)}, \tag{65}$$

be the information "density." Of course,

$$Ei^{(n)}(\mathbf{X}, \mathbf{Y}; \mathbf{W}) = I\{\mathbf{X}, \mathbf{Y}; \mathbf{W}\} = nI\{X, Y; W\}.$$

Finally, let $\Delta_1 \geqq 0$ be given and let $\{\Delta_w\}_{w \in \mathcal{W}}$ satisfy

$$R_{X \mid W}(\Delta_1) = \sum_{w \in \mathcal{W}} p_W(w) R_{X \mid W=w}(\Delta_w), \tag{66a}$$

and

$$\Delta_1 = \sum_{w \in \mathcal{W}} p_W(w) \Delta_w. \tag{66b}$$

See (35). A similar expression can be written for $R_{Y \mid W}(\Delta_2)$.

We now return to our network coding problem. With $p \in \mathcal{P}$ given, we set out to construct a core code $g$ with certain desirable properties. For any core code $g$: $\mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{C} = \{\mathbf{w}_i\}_1^{M_0} \subseteq \mathcal{W}^n$, let $N_{iw}$ = the number of occurrences of symbol $w$ in code vector $\mathbf{w}_i$, $1 \leqq i \leqq M_0$, $w \in \mathcal{W}$. The existence of a desirable core code is assured by

*Lemma 12*: *Let $p \in \mathcal{P}$ and $\epsilon > 0$ be arbitrary. Let $I^* = I(X, Y; W)$ correspond to $p \in \mathcal{P}$. For $n$ sufficiently large, there exists a code $g$ as in*

(61a) *such that*

$$(i)\ M_0 \leqq 2^{n(I^*+\epsilon)}$$

$$(ii)\ \left|\frac{N_{iw}}{n} - p_W(w)\right| \leqq \epsilon[\min_w p_W(w)],\quad \textit{for all } w \in \mathcal{W},$$

$$(iii)\ \Pr\,(S_\epsilon^c) = \sum_{(\mathbf{x},\mathbf{y}) \notin S_\epsilon} Q^{(n)}(\mathbf{x},\mathbf{y}) \leqq \epsilon,$$

*where*

$$S_\epsilon = \{(\mathbf{x},\mathbf{y}) : \frac{1}{n} i^{(n)}[\mathbf{x},\mathbf{y}; g(\mathbf{x},\mathbf{y})] \geqq I^* - \epsilon\},$$

*and* $i(\mathbf{x},\mathbf{y};\mathbf{w})$ *is defined in* (68).

We defer discussion on the proof of Lemma 12 to the end of this section.

Let us suppose that $g$ is a code that satisfies conditions $(i)$, $(ii)$, and $(iii)$ of Lemma 12. Let $\{g_{\mathbf{w}_i}^{(X)}\}_1^{M_0}$ be a family of encoders as in (61b). Consider expression (64a). The term in braces is

$$\sum_{(\mathbf{x},\mathbf{y}) \in G_i} Q^{(n)}(\mathbf{x},\mathbf{y})\Phi_i(\mathbf{x})$$

$$= \sum_{(\mathbf{x},\mathbf{y}) \in G_i \cap S_\epsilon} Q^{(n)}(\mathbf{x},\mathbf{y})\Phi_i(\mathbf{x}) + \sum_{(\mathbf{x},\mathbf{y}) \in G_i \cap S_\epsilon^c} Q^{(n)}(\mathbf{x},\mathbf{y}). \quad (67)$$

But if $(\mathbf{x},\mathbf{y}) \in G_i$ [i.e., $g(\mathbf{x},\mathbf{y}) = \mathbf{w}_i$] and $(\mathbf{x},\mathbf{y}) \in S_\epsilon$, then

$$Q^{(n)}(\mathbf{x},\mathbf{y}) \leqq 2^{-(I^*-\epsilon)n} p_b^{(n)}(\mathbf{x},\mathbf{y}|\mathbf{w}_i),$$

so that the first summation in the right member of (67) can be overbounded:

$$\begin{aligned}
&\leqq 2^{-(I^*-\epsilon)n} \sum_{\substack{(\mathbf{x},\mathbf{y}):\, g(\mathbf{x},\mathbf{y})=w_i \\ (\mathbf{x},\mathbf{y}) \in S_\epsilon}} p_b^{(n)}(\mathbf{x},\mathbf{y}|\mathbf{w}_i)\Phi_i(\mathbf{x}) \\
&\leqq 2^{-(I^*-\epsilon)n} \sum_{\mathbf{x},\mathbf{y}} p_b^{(n)}(\mathbf{x},\mathbf{y})\,|\,w_i)\Phi_i(\mathbf{x}) \qquad (68) \\
&\leqq 2^{-(I^*-\epsilon)n} \sum_{\mathbf{x}} p_b^{(n)}(\mathbf{x}|\mathbf{w}_i)\Phi_i(\mathbf{x}).
\end{aligned}$$

Combining (68), (67), and (64), we have

$$\begin{aligned}
\Gamma(\Delta_1 + \delta) &= \Pr\,\{D_1(\mathbf{x},\hat{\mathbf{x}}) > \Delta_1 + \delta\} \\
&\leqq 2^{-(I^*-\epsilon)n} \sum_{i=1}^{M_0} \sum_{\mathbf{x}} p_b^{(n)}(\mathbf{x}|\mathbf{w}_i)\Phi_i(\mathbf{x}) + \sum_{i=1}^{M_0} \sum_{(\mathbf{x},\mathbf{y}) \in G_i \cap S_\epsilon^c} Q^{(n)}(\mathbf{x},\mathbf{y}) \\
&\leqq 2^{-(I^*-\epsilon)n} \sum_{i=1}^{M_0} \sum_{\mathbf{x}} p_b^{(n)}(\mathbf{x}|\mathbf{w}_i)\Phi_i(\mathbf{x}) + \Pr\,(S_\epsilon^c). \qquad (69)
\end{aligned}$$

Now consider

$$p_b^{(n)}(\mathbf{x}|\mathbf{w}_i) = \prod_{k=1}^{n} p_b(x_k|w_{ik}),$$

where $\mathbf{w}_i = (w_{i1}, w_{i2}, \cdots, w_{in})$, $1 \leqq i \leqq M_0$. With $N_{iw}$ as defined just above Lemma 12, we see that (for a given $i$) $p_b^{(n)}(\mathbf{x}|\mathbf{w}_i)$ is the same form as $Q^{(n)}(\mathbf{x})$ in (60) with $n_j = N_{iw}$. It then follows from Corollary 11 that, with $\mathbf{w}_i$ held fixed, we can find a source code for $X$—i.e., a mapping $g_{\mathbf{w}_i}^{(X)}$—with parameter $M = M_1$ such that, for arbitrary $\epsilon, \delta > 0$,

$$M_1 \leqq \exp_2 \{ \sum_{w \in \mathcal{W}} N_{iw}[R_{X|W=w}(\Delta_w) + \epsilon]\}, \tag{70a}$$

and

$$\sum_{\mathbf{x}} p_b^{(n)}(\mathbf{x}|\mathbf{w}_i)\Phi_i(\mathbf{x}) \leqq \sum_{w} A_w 2^{-B_w N_{iw}}, \tag{70b}$$

where $\Phi_i(\mathbf{x})$ is defined in (63b) with $\Delta_1 = n^{-1}\sum_w N_{iw}\Delta_w$, and $\{\Delta_w\}$ satisfying (66). The $\{(A_w, B_w)\}$ correspond to $p_b(x|w)$. Further, since the $\{N_{iw}\}$ satisfy condition (*ii*) of Lemma 12, (70) becomes [using (66)]

$$M_1 \leqq \exp_2 \{n \sum_{w \in \mathcal{W}} (p_W[w] + \epsilon)(R_{X|W=w}[\Delta_w] + \epsilon)\}$$

$$\leqq \exp_2 \{n(R_{X|W}[\Delta_1] + \epsilon H[X] + \epsilon)\} \tag{71a}$$

and

$$\sum_{\mathbf{x}} p_b^{(n)}(\mathbf{x}|\mathbf{w}_i)\Phi_i(\mathbf{x}) \leqq \sum_{w \in \mathcal{W}} A_w 2^{-B_w n p_W(w)(1-\epsilon)} \leqq C 2^{-nB(1-\epsilon)}, \tag{71b}$$

where $C = (\text{card } \mathcal{W}) \cdot \max_w A_w$ and $B = \min_w B_w p_w(w)$. Substituting (71b) into (69) and using conditions (*i*) and (*iii*) of Lemma 12, we have

$$\Gamma(\Delta_1 + \delta) \leqq 2^{-n(I^*-\epsilon)} M_0 \cdot C 2^{-nB(1-\epsilon)} + \Pr(S_\epsilon^c)$$

$$\leqq 2^{-n(B-B\epsilon-2\epsilon)} + \epsilon \to 0, \quad \text{as } n \to \infty \text{ and then } \epsilon \to 0.$$

Since we can do an identical construction for $Y$, we have proved

*Lemma 13*: *Let* $p \in \mathcal{P}$, *and let the corresponding information be* $I(X, Y; W) = I^*$. *Let* $\Delta_1, \Delta_2 \geqq 0$ *and* $\epsilon, \delta > 0$ *be arbitrary. Then, for* $n$ *sufficiently large, there exists a coding scheme as in* (61) *with parameters* $(n, M_0, M_1, M_2)$ *such that*

(*i*) $M_0 \leqq 2^{n(I^*+\epsilon)}$,
(*ii*) $M_1 \leqq 2^{n(R_{X|W}(\Delta_1)+\epsilon)}$,
(*iii*) $M_2 \leqq 2^{n(R_{Y|W}(\Delta_2)+\epsilon)}$,

*and*

(*iv*) $\Pr\{D_1(\mathbf{X}, \hat{\mathbf{X}}) > \Delta_1 + \delta\} \leqq \epsilon$,
(*v*) $\Pr\{D_2(\mathbf{Y}, \mathbf{Y}) > \Delta_2 + \delta\} \leqq \epsilon$.

The following corollary follows from Lemma 13 in the usual way (exactly as does Theorem 9.3.1 in Gallager[4]).

*Corollary 14*: *Let* $p \in \mathcal{P}$, *and let the corresponding information* $I(X, Y; W) = I^*$. *Then, for arbitrary* $\Delta_1, \Delta_2 \geqq 0$, *the rate-triple* $[I^*, R_{X|W}(\Delta_1), R_{Y|W}(\Delta_2)]$ *is* $(\Delta_1, \Delta_2)$*-achievable. Thus,* $\mathcal{R}^{(p)}(\Delta_1, \Delta_2) \subseteq \mathcal{R}(\Delta_1, \Delta_2)$, *for all* $p \in \mathcal{P}$.

The direct-half now follows on noting that, if $S_1 \subseteq S_2$ and $S_2$ is closed, then the closure $S_1^c \subseteq S_2$. Thus, $\mathcal{R}^*(\Delta_1, \Delta_2) = [\cup_p \mathcal{R}^{(p)}(\Delta_1, \Delta_2)]^c \subseteq \mathcal{R}(\Delta_1, \Delta_2)$, which is what we had to prove.

It remains to prove Lemma 12. Since the proof is nearly identical to that of Lemma 9.3.1 in Gallager,[4] we will only outline the steps. Let $\epsilon > 0$ be arbitrary. For $\hat{\mathbf{w}} \in \mathcal{W}^n$, let $N_w(\hat{\mathbf{w}})$ = the number of occurrences of symbol $w \in \mathcal{W}$ in the $n$-vector $\hat{\mathbf{w}}$. Then define

$$T(\epsilon) = \left\{ \hat{\mathbf{w}} \in \mathcal{W}^n : \text{all } w \in \mathcal{W}, \left| \frac{N_w(\hat{\mathbf{w}})}{n} - p_w(w) \right| \leqq \epsilon \right\}.$$

Then, paralleling Gallager, there exists a mapping $g$ [as in (61a)] for which

$$M_0 \leqq 2^{n(I^* + \epsilon)}$$

and

$$\Pr\left\{ \frac{1}{n} i^{(n)}[\mathbf{X}, \mathbf{Y}; g(\mathbf{X}, \mathbf{Y})] \leqq I^* - \epsilon \quad \text{or} \quad g(\mathbf{X}, \mathbf{Y}) \notin T(\epsilon) \right\} \leqq P_t(A) + \exp\{-e^{n(\epsilon - \epsilon_2)}\} \triangleq \xi(n),$$

where $\epsilon_2 > 0$ is arbitrary and

$$A = \left\{ (\mathbf{X}, \mathbf{Y}, \mathbf{W}) : \text{either } \frac{1}{n} i^{(n)}(\mathbf{X}, \mathbf{Y}; \mathbf{W}) > I^* + \epsilon_2 \text{ or } \frac{1}{n} i^{(n)}(\mathbf{X}, \mathbf{Y}; \mathbf{W}) \leqq I^* - \epsilon, \text{ or } \mathbf{W} \notin T(\epsilon) \right\},$$

and $P_t(\cdot)$ is probability computed with respect to $p(x, y, w) \in \mathcal{P}$. By the weak law of large numbers, if $\epsilon_2 < \epsilon$, then $\xi_n \to 0$, as $n \to \infty$.

Let the code whose existence we have just asserted be $\{\mathbf{w}_i\}_1^{M_0}$. There must be at least one code vector, say, $\mathbf{w}_1$, which belongs to $T(\epsilon)$. Now

$$\Pr\{g(\mathbf{X}, \mathbf{Y}) \notin T(\epsilon)\} \leqq \xi(n).$$

If $\mathbf{x}, \mathbf{y}$ are such that $g(\mathbf{x}, \mathbf{y}) \notin T(\epsilon)$, change $g(\mathbf{x}, \mathbf{y})$ to $\mathbf{w}_1$. The new code has $g(\mathbf{x}, \mathbf{y}) \in T(\epsilon)$ and

$$\Pr\left\{\frac{1}{n} i^{(n)}[\mathbf{X}, \mathbf{Y}; g(\mathbf{X}, \mathbf{Y})] \leqq I^* - \epsilon\right\} \leqq 2\xi(n) \xrightarrow{n} 0.$$

Thus, this new code satisfies conditions $(i)$, $(ii)$, and $(iii)$ of Lemma 12.

## IV. ACKNOWLEDGMENT

Useful discussions with T. Cover, J. Wolf, D. Slepian, M. Kaplan, and H. Witsenhausen are acknowledged with thanks.

## APPENDIX

### A.1 Proof of the convexity of $\mathfrak{R}(\Delta_1, \Delta_2)$

Let $\Delta_1, \Delta_2$ be given and held fixed. Write $\mathfrak{R}(\Delta_1, \Delta_2)$ as $\mathfrak{R}$.

*Theorem 15*: $\mathfrak{R}$ *is convex.*

*Proof*: The theorem follows by a "time-sharing" argument. Let $\mathbf{R}^{(1)}$, $\mathbf{R}^{(2)} \in \mathfrak{R}$, and $0 \leqq \theta \leqq 1$. We must show that

$$\mathbf{R} \in \mathfrak{R}, \tag{72a}$$

$$\mathbf{R} = \theta\mathbf{R}^{(1)} + (1 - \theta)\mathbf{R}^{(2)}. \tag{72b}$$

Let $(g_E, g_D^{(X)}, g_D^{(Y)})$ and $(h_E, h_D^{(X)}, h_D^{(Y)})$ be codes with parameters $(n_1, M_0^{(1)}, M_1^{(1)}, M_2^{(1)}, \Delta_X^{(1)}, \Delta_Y^{(1)})$ and $(n_2, M_0^{(2)}, M_1^{(2)}, M_2^{(2)}, \Delta_X^{(2)}, \Delta_Y^{(2)})$, respectively, where $\Delta_X^{(1)}, \Delta_X^{(2)} \leqq \Delta_1$, $\Delta_Y^{(1)}, \Delta_Y^{(2)} \leqq \Delta_2$. Say $\theta = A/B$, where $A, B$ are integers, $0 \leqq A \leqq B \leqq \infty$. We show how to construct a code $(n, M_0, M_1, M_2, \Delta_X, \Delta_Y)$, where

$$\frac{1}{n}\log M_i = \theta\left(\frac{1}{n_1}\log M_i^{(1)}\right) + (1 - \theta)\left(\frac{1}{n_2}\log M_i^{(2)}\right), \tag{73}$$

$(i = 0, 1, 2)$, and $\Delta_X \leqq \Delta_1$, $\Delta_Y \leqq \Delta_2$. This will establish (72) for rational $\theta$. Since the region $\mathfrak{R}$ is closed, (72) must hold for all $\theta$, establishing Theorem 15.

We now define a code with block length $n = cn_1 + dn_2$, where $c = An_2$, $d = (B - A)n_1$. Let $(\mathbf{x}, \mathbf{y}) \in \mathfrak{X}^n \times \mathfrak{Y}^n$ be a sequence of $n$ pairs. Partition this sequence into $c$ blocks of $n_1$ pairs and $d$ blocks of $n_2$ pairs. Encode-decode the first $c$ blocks using encoder-decoder $(g_E, g_D^{(X)}, g_D^{(Y)})$, and encode-decode the remaining $d$ blocks using encoder-decoder $(h_E, h_D^{(X)}, h_D^{(Y)})$. Denote this combination encoder-decoder by $(f_E, f_D^{(X)}, f_D^{(Y)})$. Consider $f_E(\mathbf{x}, \mathbf{y}) = (S_0, S_1, S_2)$. The quantity

$S_i (i = 0, 1, 2)$ takes values in a set with

$$(M_i^{(1)})^c \cdot (M_i^{(2)})^d \triangleq M_i$$

members. This set can, of course, be put in 1-1 correspondence with $I_{M_i}$. Thus, for $i = 0, 1, 2$,

$$\frac{1}{n} \log M_i = \frac{c}{n} \log M_i^{(1)} + \frac{d}{n} \log M_i^{(2)} = \frac{\theta}{n_1} \log M_i^{(1)} + \frac{(1-\theta)}{n_2} \log M_i^{(2)},$$

which is (73). Further, the new code has $\Delta_X \leqq \Delta_1$, and $\Delta_Y \leqq \Delta_2$, so that the lemma follows.

### A.2 Proof of Theorem 1

Again let $\Delta_1, \Delta_1 \geqq 0$ be given and fixed, and write $\mathfrak{R}(\Delta_1, \Delta_2) = \mathfrak{R}$, and $\overline{\mathfrak{R}}(\Delta_1, \Delta_2) = \overline{\mathfrak{R}}$.

We first establish part (*ii*) of the theorem. Let $\mathbf{R} \in \mathfrak{S}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \mathfrak{A}'$. If $\mathbf{R} \notin \overline{\mathfrak{R}}$, then there exists an $\hat{\mathbf{R}} = (\hat{R}_0, \hat{R}_1, \hat{R}_2) \in \mathfrak{R}$, such that $\hat{R}_i \leqq R_i$, $i = 0, 1, 2$, and at least one of these inequalities holds strictly. Thus,

$$C(\boldsymbol{\alpha}, \hat{\mathbf{R}}) - C(\boldsymbol{\alpha}, \mathbf{R}) = (\hat{R}_0 - R_0) + \alpha_1(\hat{R}_1 - R_1) + \alpha_2(\hat{R}_2 - R_2) < 0. \quad (74)$$

The inequality follows from $\alpha_1, \alpha_2 > 0$. This contradicts $\mathbf{R} \in \mathfrak{S}(\boldsymbol{\alpha})$. Thus $\mathbf{R} \in \overline{\mathfrak{R}}$, which establishes part (*ii*).

It remains to establish part (*i*). We must first obtain some preliminary facts.

*Lemma 16*: *Let* $(R_0, R_1, R_2) \in \mathfrak{R}$. *Then*

(*a*) *for* $a_i \geqq 0$ $(i = 0, 1, 2)$, $(R_0 + a_0, R_1 + a_1, R_2 + a_2) \in \mathfrak{R}$,

(*b*) *for* $0 \leqq \theta \leqq 1$, $[(1-\theta)R_0, R_1 + \theta R_0, R_2 + \theta R_0] \in \mathfrak{R}$,

(*c*) *for* $0 \leqq \theta_1, \theta_2 \leqq 1$,
$[R_0 + \theta_1 R_1 + \theta_2 R_2, (1-\theta_1)R_1, (1-\theta_2)R_2] \in \mathfrak{R}$.

*Proof*:

(*a*) follows immediately from the definition of $\mathfrak{R}$.

(*b*) follows on noting that data sent through the common channel can be transmitted instead through *each* private channel.

(*c*) follows on noting that any data transmitted through either private channel can be transmitted instead over the common channel.

Next, for $R_1, R_2 \geqq 0$ write $\mathbf{r} = (R_1, R_2)$, and define the function

$$F(\mathbf{r}) = F(R_1, R_2) = \min\{R_0 : (R_0, R_1, R_2) \in \mathfrak{R}\}. \tag{75}$$

The minimum exists because $\mathfrak{R}$ is closed. Clearly, $(R_0, R_1, R_2) \in \overline{\mathfrak{R}}$ only if $R_0 = F(R_1, R_2)$.

*Lemma 17*: $F(\mathbf{r})$ *is convex.*

*Proof*: Let $\mathbf{r}^{(1)}, \mathbf{r}^{(2)}$ be arbitrary. Then $[F(\mathbf{r}^{(1)}), \mathbf{r}^{(1)}], [F(\mathbf{r}^{(2)}), \mathbf{r}^{(2)}] \in \mathfrak{R}$. Since $\mathfrak{R}$ is convex, for $0 \leqq \theta \leqq 1$,

$$\begin{aligned}\theta[F(\mathbf{r}^{(1)}), \mathbf{r}^{(1)}] + (1-\theta)[F(\mathbf{r}^{(2)}), \mathbf{r}^{(2)}] \\ = [\theta F(\mathbf{r}^{(1)}) + (1-\theta)F(\mathbf{r}^{(2)}), \theta\mathbf{r}^{(1)} + (1-\theta)\mathbf{r}^{(2)}] \in \mathfrak{R}.\end{aligned}$$

Thus, by the definition of $F(\cdot)$,

$$F[\theta\mathbf{r}_1^{(1)} + (1-\theta)\mathbf{r}^{(2)}] \leqq \theta F(\mathbf{r}^{(1)}) + (1-\theta)F(\mathbf{r}^{(2)}),$$

establishing the lemma.

Now it follows from the convexity of $F(\cdot)$ that, for arbitrary $\mathbf{r}^* = (R_1^*, R_2^*)$, $R_1^* R_2^* \geqq 0$, there exists constants $\alpha_i = \alpha_i(\mathbf{r}^*)$, $i = 1, 2$, such that, for all $\mathbf{r}$,

$$F(\mathbf{r}) - F(\mathbf{r}^*) \geqq \sum_{i=1}^{2} \alpha_i(R_i^* - R_i). \tag{76}$$

This is a statement of the well-known fact that any convex curve lies above a plane of support. Here the curve is the locus of points in $(R_0, R_1, R_2)$-space given by $R_0 = F(\mathbf{r}) = F(R_1, R_2)$, and the plane is the locus of points $R_0 = F(\mathbf{r}^*) + \sum_{i=1}^{2}\alpha_i(R_i^* - R_i)$. Note that the curve and the plane coincide at $\mathbf{r} = \mathbf{r}^*$.

Now let $\mathbf{R}^* = (R_0^*, R_1^*, R_2^*) \in \overline{\mathfrak{R}}$. Then $R_0^* = F(R_1^*, R_2^*)$. Let $\mathbf{R} = (R_0, R_1, R_2)$ be any triple in $\mathfrak{R}$. Then with $\mathbf{r} = (R_1, R_2)$, (76) yields

$$F(r) + \alpha_1 R_1 + \alpha_2 R_2 \geqq R_0^* + \alpha_1 R_1^* + \alpha_2 R_2^*.$$

Since, by definition of $F(\cdot)$, $F(\mathbf{r}) \leqq R_0$, we have:

$$\begin{aligned}R_0^* + \alpha_1 R_1^* + \alpha_2 R_2^* &= \min_{\mathbf{R}\in\mathfrak{R}} (R_0 + \alpha_1 R_1 + \alpha_2 R_2) \\ &= \min_{\mathbf{R}\in\mathfrak{R}} C(\boldsymbol{\alpha}, \mathbf{R}) = T(\boldsymbol{\alpha}),\end{aligned}$$

where $C(\boldsymbol{\alpha}, \mathbf{R})$ and $T(\boldsymbol{\alpha})$ are defined by (12) and (13), respectively. Thus, we have shown that, if the triple $\mathbf{R}^* \in \overline{\mathfrak{R}}$, then $\mathbf{R}^* \in \mathfrak{s}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ need not necessarily belong to $\mathfrak{A}$.

Now suppose that $\mathbf{R}^* = (R_0^*, R_1^*, R_2^*) \in \overline{\mathcal{R}}$, and $\mathbf{R}^* \in \mathcal{S}(\boldsymbol{\alpha})$ where, say, $\alpha_1 < 0$. From Lemma 16($a$) (with $a > 0$), $\hat{\mathbf{R}} = (R_0, R_1^* + a, R_2^*) \in \mathcal{R}$, and

$$C(\boldsymbol{\alpha}, \hat{\mathbf{R}}) < C(\boldsymbol{\alpha}, \mathbf{R}^*),$$

which implies $\mathbf{R}^* \notin \mathcal{S}(\boldsymbol{\alpha})$, a contradiction. Thus, $\alpha_1$ (and similarly $\alpha_2$) $\geqq 0$. Next, suppose $\mathbf{R}^* \in \overline{\mathcal{R}}$, and $\mathbf{R}^* \in \mathcal{S}(\boldsymbol{\alpha})$ where, say, $\alpha_1 > 1$. Then, from Lemma 16($c$), $\hat{\mathbf{R}} = (R_0^* + R_1^*, 0, R_2^*) \in \mathcal{R}$, and

$$C(\boldsymbol{\alpha}, \hat{\mathbf{R}}) < C(\boldsymbol{\alpha}, \mathbf{R}^*),$$

again a contradiction. Thus, $\alpha_1$, and similarly $\alpha_2 \leqq 1$. Finally, suppose that $\mathbf{R}^* \in \overline{\mathcal{R}}$ and $\mathbf{R}^* \in \mathcal{S}(\boldsymbol{\alpha})$, where $\alpha_1 + \alpha_2 < 1$. By Lemma 16($b$), $\hat{\mathbf{R}} = (0, R_1^* + R_0^*, R_2^* + R_0^*) \in \mathcal{R}$, and

$$C(\boldsymbol{\alpha}, \hat{\mathbf{R}}) \leqq C(\boldsymbol{\alpha}, \mathbf{R}^*),$$

a contradiction. Thus, $\alpha_1 + \alpha_2 \geqq 1$. We conclude that

$$\overline{\mathcal{R}} \subseteq \bigcup_{\boldsymbol{\alpha} \in \mathcal{A}} \mathcal{S}(\boldsymbol{\alpha}), \tag{77}$$

where $\mathcal{A}$ is the set of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ that satisfy $0 \leqq \alpha_1, \alpha_2 \leqq 1$, and $\alpha_1 + \alpha_2 \geqq 1$. This is part ($i$). This completes the proof of Theorem 1.

### *A.3 Proof of Lemma 5*

Let $\{\Delta_k\}_1^n$ be given, and, for $k = 1, 2, \cdots, n$, let $q_{tk}(\hat{x} \mid x, w)$, $\hat{x} \in \hat{\mathfrak{X}}$, $w \in \mathcal{W}_k$ be a test channel for which

$$\sum_{w \in \mathcal{W}_k} \sum_{x, \hat{x}} d(x, \hat{x}) q_{tk}(\hat{x} \mid x, w) p_k(x, w) \leqq \Delta_k, \tag{78a}$$

and

$$I(X; \hat{X} \mid W_k) \leqq R_{X \mid W_k}(\Delta_k) + \epsilon, \tag{78b}$$

where $\epsilon > 0$ is arbitrary. For $w \in \mathcal{W} = \sum_{k=1}^n \mathcal{W}_k$, $x \in \mathfrak{X}$, $\hat{x} \in \hat{\mathfrak{X}}$, define the test channel

$$q_t^*(\hat{x} \mid x, w) = q_{tk}(\hat{x} \mid x, w), \quad \text{for } w \in \mathcal{W}_k, \ 1 \leqq k \leqq n.$$

Then

$$\sum_{x,y,w} d(x, \hat{x}) q_t(\hat{x} \mid x, w) p^*(x, w) = \sum_{k=1}^n \sum_{w \in \mathcal{W}_k} \sum_{x,y} \frac{1}{n} d(x, \hat{x}) q_{tk}(\hat{x} \mid x, w) p_k(x, w) \leqq \frac{1}{n} \sum_{k=1} \Delta_k.$$

Thus, corresponding to the distribution $p^*(x, w) \cdot q_i^*(\hat{x} \mid x, w)$,

$$I(X, \hat{X} \mid W) \geqq R_{X \mid W}\left(\frac{1}{n} \sum_k \Delta_k\right). \tag{79}$$

However, by a straightforward calculation,

$$I(X, \hat{X} \mid W) = \frac{1}{n} \sum_{k=1}^{n} I(X, \hat{X} \mid W_k) \leqq \frac{1}{n} \sum_{k=1}^{n} R_{X \mid W_k}(\Delta k) + \epsilon. \tag{80}$$

The inequality follows from (78b). Combining (79) and (80) and letting $\epsilon \to 0$, we have Lemma 5.

## REFERENCES


1. T. Berger, *Rate-Distortion Theory*, Englewood Cliffs, N. J.: Prentice-Hall, 1971.
2. D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," IEEE Trans. on Information Theory, *IT-19*, July 1973, pp. 471–480.
3. A. D. Wyner, "The Common Information of Two Dependent Random Variables," to appear in IEEE Trans. on Information Theory.
4. R. G. Gallager, *Information Theory and Reliable Communication*, New York: John Wiley, 1968.
5. R. M. Gray, "A New Class of Lower Bounds to Information Rates of Stationary Sources via Conditional Rate-Distortion Functions," IEEE Trans. on Information Theory, *IT-19*, July 1973, pp. 480–489.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 9, November 1974
*Printed in U.S.A.*

# Interfacial Dopants for Dual-Dielectric, Charge-Storage Cells

By D. KAHNG, W. J. SUNDBURG, D. M. BOULIN, and J. R. LIGENZA

(Manuscript received February 19, 1974)

*When a suitable interfacial dopant, such as W, is introduced at the interface between the dielectrics of a DDC cell, the write-erase characteristics of the cell are greatly improved. The useful range of the dopant concentration is determined to lie between about $10^{14}$ to $10^{15}$ atoms/cm$^2$. The interfacial dopant allows the fabrication of a DDC cell with relatively thick $SiO_2$ layers ($>50$ Å). The result is a substantially permanent memory cell that can still be subjected to electrical write-erase at reasonable gate-voltage conditions.*

## I. INTRODUCTION

There has been an increasing interest in recent months in the dual dielectric metal insulator semiconductor (MIS) cell as a nonvolatile semiconductor memory element. A true nonvolatile semiconductor memory could replace the omnipresent magnetic memory because associated with it one also expects fast access capability as well as interface compatibility with other semiconductor logic circuits. Key features in such a semiconductor memory cell, then, are: true nonvolatility, high-speed access capability, and ease of write-erase operations. The development of the dual-dielectric charge-storage (DDC) cells have followed two parallel paths, both enjoying a limited success yielding commercial products. The first centers around the concept of the *floating gate*,[1] an artificially created metallic charge-storage site located at the dual dielectric interface. The second uses the naturally occurring interfacial states existing at the dual-dielectric interface as the charge storage sites, as in metal-nitride-oxide-semiconductor (MNOS) memory transistors.[2] The advantages and disadvantages of these two approaches to the realization of DDC cells is reviewed

briefly. And then the concept of the interfacial dopants, the heart of this paper, emerges as a particularly beneficial compromise between these two concepts, resulting in an optimum DDC cell, with true nonvolatility, yet with undemanding write-erase conditions.

### *1.1 Floating-gate concept*

The early DDC cells were built around a floating gate introduced as the charge-storage site located between the two dissimilar dielectric layers. The charge exchange between the floating gate and Si in this structure was achieved via electron tunnelling through a thin $SiO_2$ layer grown on the Si substrate. The advantage in the floating-gate concept lies in that the potential well can be tailored to the write-erase needs. In practice, a difficulty was encountered in achieving true nonvolatility in these cells, at least in large-volume-memory situations, because any single shorting path between the floating gate and either the Si substrate or the external gate was sufficient to cause rapid loss of the entire charge on the floating gate. It has been suggested[3] that this shortcoming may be overcome by replacing the continuous-metal floating gate with mutually isolated small metal islands separated by distances shorter than the Debye length at the Si surface. Since then, attempts have been made[4] to fabricate and characterize dual-dielectric MIS cells with metal islands at the dielectric-layer interface. In general, the nonvolatility has not been exceptionally good (the retention time being on the order of several hours) presumably because the existence of the metal islands have degraded the quality of the dielectric layers, leading to higher leakage. The metal islands are also expected to give rise to a field-enhancement effect, causing more rapid stored-charge decay.

A more successful approach[5] to achieve true nonvolatility has been to insert a continuous floating gate between thick (approximately 1000 Å) $SiO_2$ layers. However, due to its large oxide thickness, this structure would require prohibitively large gate voltages to effect write-erase if tunnelling were to be used. Thus, injection of electrons from Si into the floating gate is accomplished by use of hot electrons created by biasing a nearby pn junction into avalanche.[6] This approach, although it gives rise to excellent nonvolatility, suffers from the lack of a ready electrical means to eject electrons from the floating gate. The erase operation is only possible either by thermal means or by ultraviolet irradiation, thus severely limiting the versatility of these devices. Although some arrangements which permit electrical erasing have been proposed and tested,[7] they usually require increased com-

plexity in the cell structure as well as high gate voltages, and therefore do not appear practical at present.

### *1.2 MNOS cells*

Another DDC cell structure uses the naturally occurring interface states at the dividing planes of the two dissimilar dielectric layers. The structure commonly referred to as the MNOS cell is the outgrowth of $Si_3N_4$-$SiO_2$-Si MIS studies carried out initially by Szedon et al.[8] Since then, there has been a considerable amount of work done in this area in the past few years, culminating in experimental memories as large as 2048 bits.[9] It has been well established that ease of write-erase operations in these devices requires use of very thin (a little less than 30 Å) $SiO_2$ layers on Si so that the charge exchange between the interfacial storage sites and Si may proceed via direct tunnelling across the $SiO_2$ layers.

Figure 1a, a schematic energy-band diagram of an MNOS structure under bias, illustrates the direct tunnelling mechanism. Electrons from the Si conduction band tunnel through the energy barrier associated with the thin oxide layer to final states at the oxide-nitride interface. Some of these electrons may further penetrate into the nitride layer by hopping along the bulk traps. However, the majority of them are expected to reside at or near the dual-dielectric interface. If the oxide thickness is larger than about 30 Å, an appreciable current flow is possible only when the electrons originating from the Si conduction band tunnel to the conduction band of the oxide and then proceed to the dual-dielectric interface (as in the Fowler-Nordheim tunnelling in Fig. 1b). In this case, however, most of the electrons reach the nitride as hot electrons and only a small fraction of them are captured by the interface states. Since the Fowler-Nordheim tunnelling probability is significantly smaller than the direct tunnelling case at a given field, a higher field must be applied to induce a comparable current density. This would raise the field in the nitride and further impede electron trapping in the nitride. It is clear from the above that the conventional MNOS cells are expected to work well only with thin oxide layers. A similar argument is expected to hold true in the case of hole tunnelling in MNOS structures, although the discussion so far has been given in terms of electron tunnelling for simplicity.

A major drawback in the MOS cells is, however, a direct consequence of the thin $SiO_2$ layer. The nonvolatility is limited by charge leakage through the thin $SiO_2$ layer by direct back-tunnelling. Although some claims have been made that nonvolatility of 10 to 100 years is possible

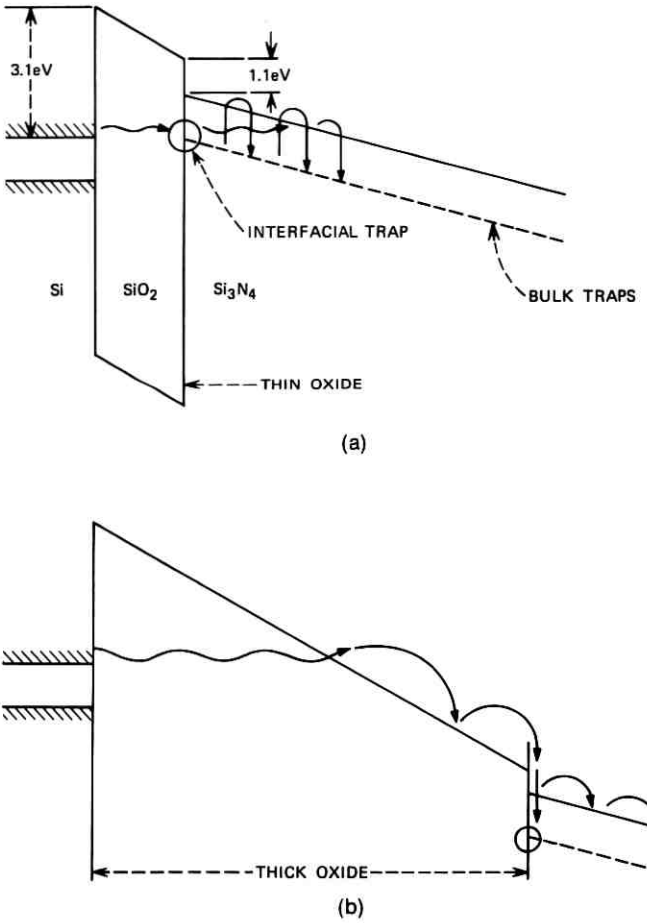



Fig. 1—Energy-band diagram of charging of the dual dielectric interfacial region. (a) Direct tunnelling of electrons. (b) Fowler-Nordheim tunnelling of electrons.

with MNOS structures, it appears that this feat is only achievable by use of thicker $SiO_2$ layers with the ensuing penalty in write-erase gate-voltage pulse both in height and width. The MNOS structures operable with a short write-erase time, say in 1-$\mu$s range, appear to have a non-volatility usually considerably less than 1 year.

Outer dielectric layers other than $Si_3N_4$ have also been used in dual-dielectric charge-storage cells. A notable example is the use of $Al_2O_3$. In 1970, Nakanuma et al.,[10] for example, showed that $Al_2O_3$ is a suitable outer layer. The nonvolatility of these cells is again critically dependent on $SiO_2$ layer thickness. Cells with $SiO_2$ layers about 100 Å in thickness

were fabricated and showed good charge retention. However, their write-erase gate voltages had to be 50 volts or more and of 1 ms or so in duration. Also, the ejection of electrons from the interfacial states during the erase operation appears to be rather sluggish, possibly because the density of states is not high enough and/or the barrier height appears to be too high. Some attempts to increase the interfacial state density by, for example, varying the $Al_2O_3$ deposition conditions appear to result in a lack of reproducibility and/or loss of good nonvolatility.

### *1.3 Interfacial dopants*

It is apparent from the foregoing that while a true nonvolatility in a dual-dielectric cell can only be expected with a structure utilizing relatively thick (greater than 50 Å) $SiO_2$ layers, this has led to write-erase operations requiring high voltage and/or long pulses. These problems have arisen from the fact that the naturally occurring interfacial states are not really ideal for the various roles they have to play. Attempts to increase their density, which allows shorter time for write-erase, usually result in lossy outer dielectric layers. Although no specific description is available in the literature, past attempts in this regard appear to be comprised of variations in outer-layer-deposition conditions to achieve a certain degree of off-stoichiometric outer layers, at least in the vicinity of the interface.

It is reasonable to expect that if a method for controlling interfacial states without degrading the dielectric properties of the dual-dielectric layers were available, one might overcome the aforementioned difficulties. We wish to show in this paper that, indeed, such a means has been found. Interfacial states induced by certain dopants deliberately located at the interface do have beneficial effects such that considerable improvement in write-erase operations have resulted in conjunction with the use of relatively thick (greater than 50 Å) $SiO_2$ layers, which are essential for nonvolatility.

Interfacial dopant-induced states of a sufficient density will increase the capture probability of the incoming electrons. This is schematically illustrated in Fig. 2. Furthermore, the energy levels associated with these dopant-induced states will depend on the choice of the dopant. This would allow a suitable condition for the erase operation. The energy levels should not be excessively deep, since this would require a prohibitively high field for ejection of stored electrons. The energy levels should not be too shallow, since the stored charge might easily decay through thermal activation. The naturally occurring interfacial

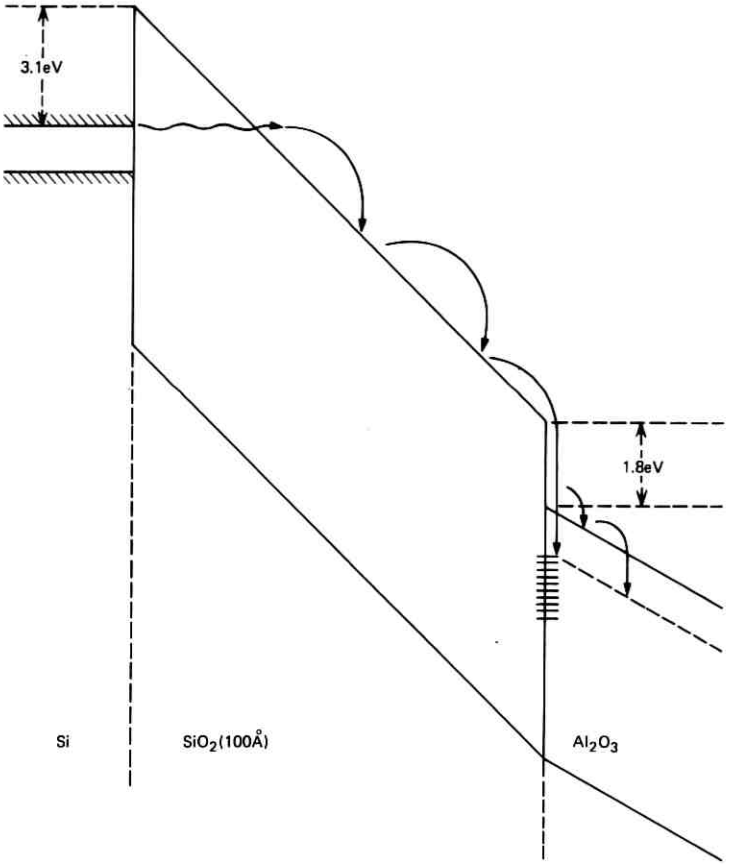



Fig. 2—Energy-band diagram of dopant-induced interfacial states. These states efficiently capture electrons arriving at the dual-dielectric interface following Fowler-Nordheim tunnelling through a relatively thick energy barrier.

states clearly do not possess a comparable versatility as charge-storage sites.

We find that there exists a range of dopant concentration which gives rise to optimum benefits. The upper limit of this range is dictated by the considerations that island formation by the dopant, either during its deposition or during the subsequent outer-layer deposition, is not really desirable since the dielectric property of the dual-dielectric layers is usually degraded. Furthermore, the initial deposition of the outer layers is strongly influenced by the presence of nonuniformity and may result in lossy layers. The upper limit also depends on the choice

of an outer layer and a dopant. Thus, $Al_2O_3$ with a nonoxidizing dopant such as Ir appears to limit the dopant concentration below $1 \times 10^{15}/cm^2$. On the other hand, $Si_3N_4$ with W may tolerate up to $5 \times 10^{15}/cm^2$. The lower limit is naturally dictated by the ease of write-erase operations and appears to lie at around $1 \times 10^{14}/cm^2$.

### *1.4 Selection of dopants*

It is desirable that the dopants selected be physically confined at the dual-dielectric-layer interface. This means that both during dopant deposition and during subsequent outer-layer deposition, any migration of the dopant is undesirable. The preferred method for depositing the dopant on the surface of the $SiO_2$ layer is thermal evaporation. Since good-quality outer layers are, at present, only obtainable through chemical vapor deposition (CVD) techniques at somewhat elevated temperatures of about 900°C, it is important that the deposited dopant does not diffuse into the $SiO_2$ layer and the outer layer at these temperatures. It is equally important that the deposited dopant does not evaporate away during the first phase of the outer layer growth. The vapor pressures of W, Pt, Ir, Ta, and Nb at 900°C are all lower than $10^{-11}$ torr.[11] On the other hand, Al has a vapor pressure of $1.5 \times 10^{-5}$ torr;[11] thus, considerable amounts of Al could be lost. At any rate, control of the final dopant concentration may be more difficult in the case of Al.

The vapor pressures of the dopant oxide may also be an important quantity to consider especially when the outer layer is an oxide layer. For example, tungsten oxide has a vapor pressure of approximately $2 \times 10^{-6}$ torr.[12] The diffusion constant of Pt into $SiO_2$ at 900°C is reported[13] to be $7.3 \times 10^{-17}$ cm²/sec. Compare this with the diffusion constant of Al into $SiO_2$ at 990°C of $8.2 \times 10^{-14}$ cm²/sec[14] and that of P into $SiO_2$ at 900°C of $9 \times 10^{-15}$ cm²/sec.[15]

It is evident from the above that indeed the suitable dopants available are quite limited in number from the thermodynamic properties alone. The dopant-induced interfacial states should also possess suitable properties. This might narrow the choice down further. We discuss this aspect in more detail when we examine the electrical behavior of these cells, since it is in this aspect that the dopant-induced states come into strong play.

## II. CELL FABRICATION

Our experimental vehicles have been capacitors and IGFETs built on an Si substrate about 5-ohm-cm n- or p-type, oriented ⟨111⟩ or

⟨100⟩. The usual mechanical-chemical polish was given before the first thin $SiO_2$ layer growth was carried out in dilute $O_2$ in He at about 1100°C. Prior to this important thin-$SiO_2$-growth step, the wafers were oxidized in 100 percent $O_2$ to a thickness of about 1000 Å. This initial oxide was kept until the wafer was ready to receive the treatment for the thin oxide layer when this thick oxide layer was stripped in a buffered HF and thoroughly rinsed in deionized water. The thin-oxide-growth time was kept at 10 minutes. The partial pressure of $O_2$ was varied to give rise to the desired first $SiO_2$ layer thickness of 50 Å to 150 Å. The dielectric breakdown strengths of the thin-oxide layers were usually 7 to 8 $\times$ $10^6$ V/cm defined as the dc field at which the current density reaches 5 $\times$ $10^{-8}$ A/cm$^2$.

The dopant evaporation was performed with an E-gun. A quartz oscillator monitor was located about 5 cm away from the source and the samples were located at various distances away from the source depending on the amount of dopant desired. A shutter was employed to initiate the dopant flux as well as to shut it off. An exposure time of 2 minutes was generally used. The monitor had a capability of measuring 5 $\times$ $10^{16}$/cm$^2$ of deposited dopant to within 50-percent accuracy. A dopant concentration of 5 $\times$ $10^{14}$/cm$^2$ could easily be obtained on the sample surface when the sample was located at about 50 cm away from the source.

The dopant-covered samples were then ready for the outer-dielectric-layer deposition, which was either an $Al_2O_3$ deposition using the well-known techniques of $AlCl_3$-$CO_2$-$H_2$ CVD or an $Si_3N_4$ deposition via $SiH_4$-$NH_3$ CVD. A ratio of $SiH_4$ to $NH_3$ of about 0.01 was used to obtain $Si_3N_4$ layers of least conductance.[16] We feel that freedom in choosing CVD conditions for optimum results in outer layers, such as low conductance and/or high-dielectric constant, need not be surrendered in our case since the interfacial states configuration is more or less independent of the chemical nature of the outer layer when the interfacial dopants are employed. This is important because a true nonvolatility requires not only relatively thick $SiO_2$ layers but also good insulating outer layers. Outer-layer thickness in the 300-Å to 700-Å range were investigated in this study.

Some of the doped and undoped dual-dielectric layers were subjected to Rutherford back-scattering experiments[17] to ascertain dopant location and its concentration. The dopant concentration was within a factor of two of the estimated values from the indication on the monitor. The dopant location is judged to be at the dielectric interface, to the degree that this type of probe can be used to certify this.

The dual-dielectric layers were finally contacted with Al by evaporation from a devitrified carbon crucible through masks defining 15-mil-diameter circular areas for test capacitors. In addition, we have fabricated IGFETs with the dual-dielectric gate insulator for examination, more or less following a standard procedure. Again Al was used as the gate electrode.

### III. MEMORY CHARACTERISTICS

IGFET threshold voltages are the most convenient parameter to assess memory behavior of the dual-dielectric charge-storage cells. Figure 3 shows write characteristics of a cell with interfacial dopant W

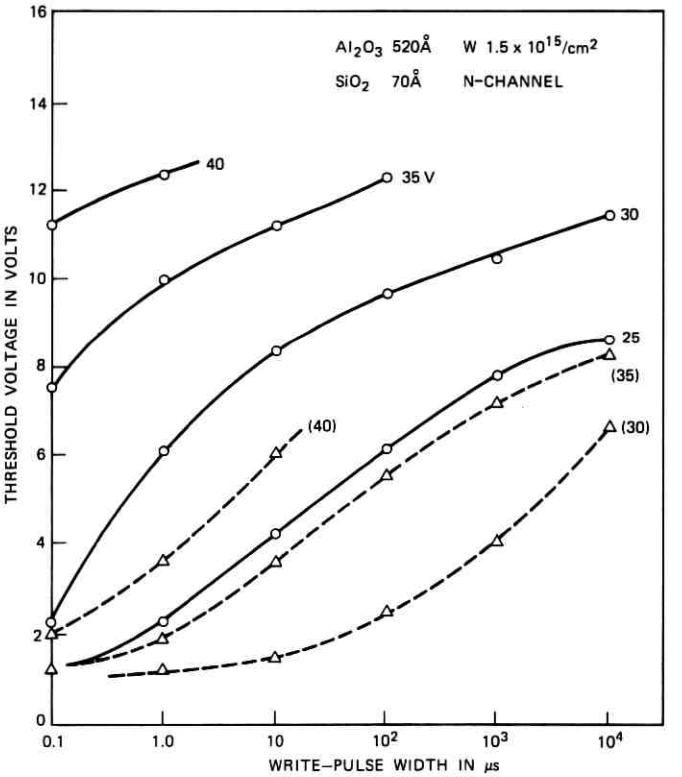



Fig. 3—Shifts in threshold voltage in n-channel IGFETs with dual-dielectric gate insulation consisting of 70-Å-thick $SiO_2$ and 520-Å-thick $Al_2O_3$ after application of positive-gate-voltage pulses. Solid lines apply to structures containing $1.5 \times 10^{15}$/cm² of W at dual-dielectric interface; dashed lines apply to structures with no interfacial dopant. The initial threshold voltages were at 1 volt.

of $1.5 \times 10^{15}/cm^2$ in concentration. The n-channel IGFET threshold voltages attained after positive-gate-voltage pulses (with respect to the Si substrate) were applied are plotted as a function of pulse width. The threshold voltage was initially at about 1 V. After each pulsing, the threshold voltage shifted to a larger positive value as one might expect since the pulsing resulted in electron injection from the Si substrate into the storage states. The $SiO_2$-layer thickness was about 70 Å and the $Al_2O_3$-outer-layer thickness was about 520 Å. It is evident that this structure allows shifts in threshold voltages of as much as 7 V with only a 35-V, 100-ns pulse. A similar shift can be obtained with a 30-V pulse when the pulse width is increased to 100 $\mu$s. For comparison, plots are shown obtained with cells located on the same wafer identical to those discussed above, except that the W dopant was not present. These curves are shown in Fig. 3 in dashed lines. With 35 volts, it is necessary to use a 10-ms pulse width to produce a similar shift in threshold voltages.

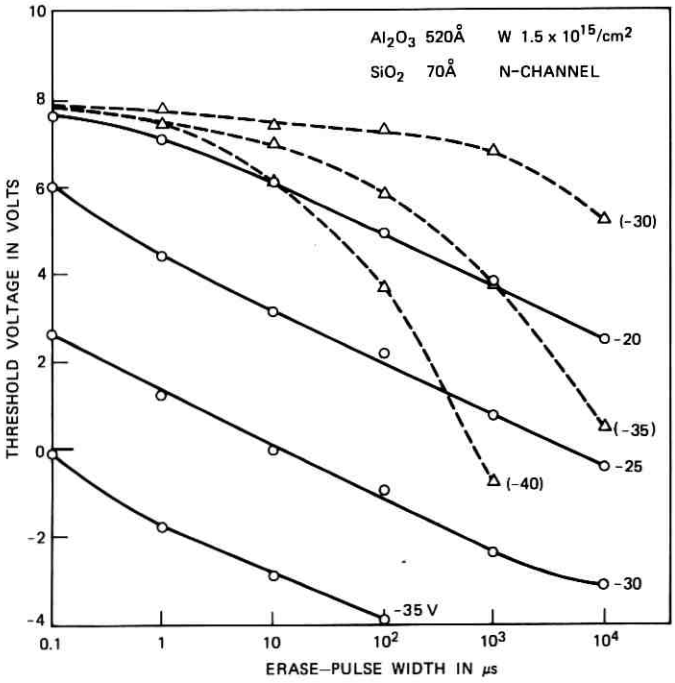



Fig. 4—Shifts in threshold voltage, initially at 8 V, after application of negative-gate-voltage pulses to the IGFETs described in Fig. 1.

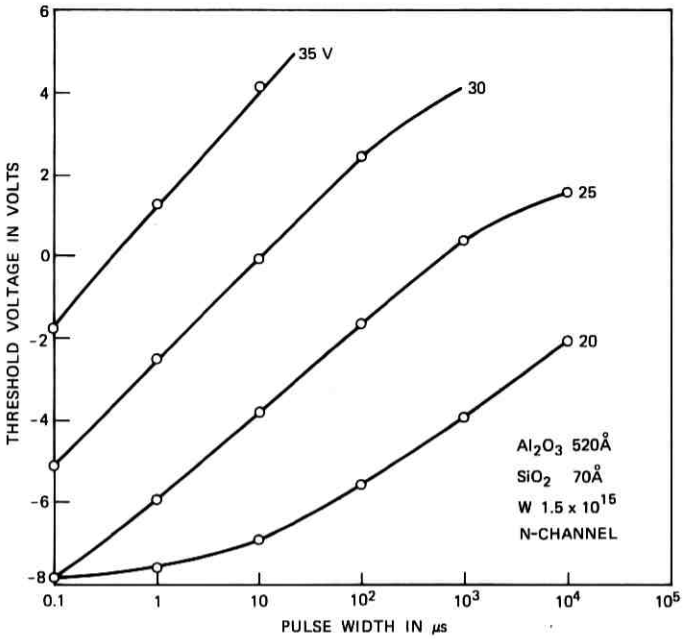



Fig. 5—Shifts in threshold voltages, initially at −8 V, after application of positive-gate-voltage pulses to the IGFETs described in Fig. 1.

Even more spectacular differences between memory cells with and without the interfacial dopant are evident when one examines the erase characteristics shown in Fig. 4. The initial threshold voltage was at 8 V prior to application of erase-voltage pulses. The usual sluggish erase behavior of an $Al_2O_3$-$SiO_2$-Si structure has clearly disappeared with similar structures with W as the interfacial dopant. An important additional feature of cells with interfacial dopant W is the fact that $Al_2O_3$-$SiO_2$-Si structures allow *positive* charging as well, which is usually not the case without the interfacial dopant. Figure 5 shows the erase characteristics after positive charging to the extent of −8 V in threshold voltage.

Dual-dielectric charge-storage cells with thicker insulator layers show similar improvements when the interfacial dopants are used. Figure 6 shows the erase characteristics of cells with a 100-Å-thick $SiO_2$ layer and 570-Å-thick $Al_2O_3$ layer. Again curves pertaining to cells with and without interfacial dopant W are shown for comparison. As expected, the pulse voltages required to induce comparable threshold-voltage shifts are increased from the values needed with thinner structures

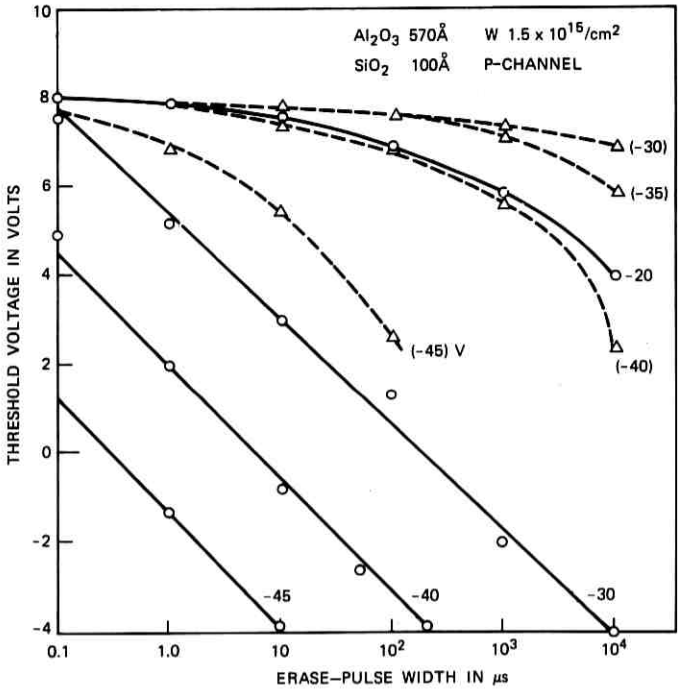



Fig. 6—Shifts in threshold voltage, initially at 8 V, after application of negative-gate-voltage pulses to p-channel IGFETs with dual-dielectric gate insulation consisting of 100-Å-thick $SiO_2$ and 570-Å-thick $Al_2O_3$. Solid lines apply to structures containing $1.5 \times 10^{15}/cm^2$ of W at dual-dielectric interface; dashed lines apply to structures with no interfacial dopant.

(Fig. 4). However, dramatic improvements in erase characteristics in structures with the interfacial dopant are clearly evident. Here again, the initial threshold voltage of 8 V was used prior to the erase operation. The write characteristic also shows similar improvements.

One does not expect a strong dependence of the erase characteristics on the initial amount of charge in storage (although the self-induced field is linearly super-imposed on the externally applied field) because the induced field is very small compared to the externally applied field. This is shown in Fig. 7 where the threshold voltages after application of the erase pulse are plotted as a function of pulse height with the initial threshold voltages as a parameter. A pulse width of 200 $\mu$s was used for this experiment. As can be seen, the erase curves quickly converge with each other. For a $-30$-V pulse, the resulting threshold voltages are identical for all practical purposes. This is a useful feature

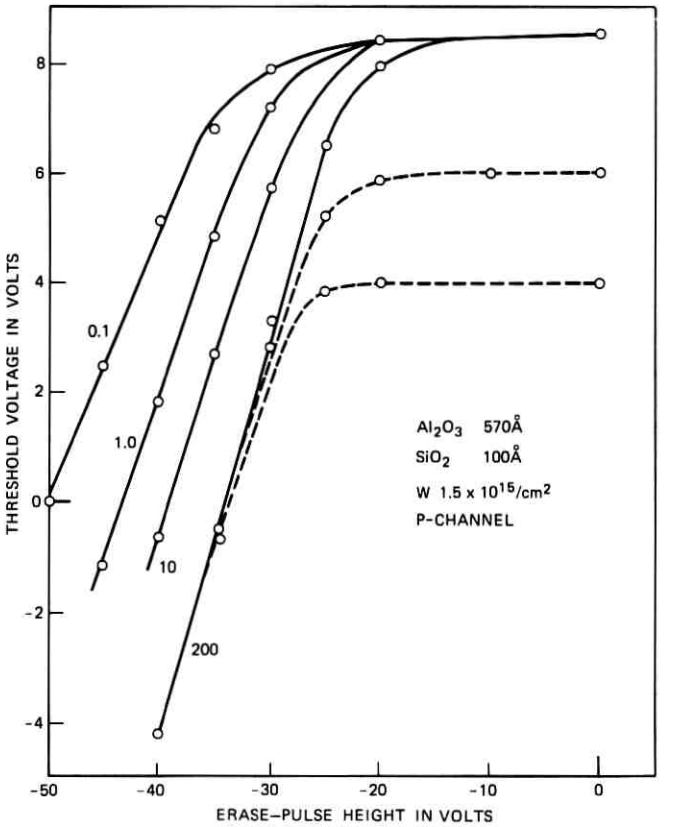



Fig. 7—Shifts in threshold voltages, initially at three different values, in IGFETs described in Fig. 4 after application of negative-gate-voltage pulses of 200-$\mu$s duration and various pulse heights. Also shown are shifts in threshold voltages, initially at 8 V, after application of pulses of 10 $\mu$s, 1.0 $\mu$s, and 0.1 $\mu$s in duration.

in that the various initial threshold voltages may represent cells that have gone through various amounts of information storage time after the simultaneous write operation. Figure 7 also shows erase curves using shorter pulse widths for comparison.

The interfacial-dopant concentration is expected to have a lower limit for which the beneficial aspect of the interfacial dopant is not so evident. This lower limit is established to be about $1 \times 10^{14}$/cm$^2$, as can be seen from Fig. 8. Here the amount of interfacial dopant W is $1.5 \times 10^{14}$/cm$^2$. When compared to the mate cell with no interfacial dopant, the erase curves show some effect, but it is not as pronounced

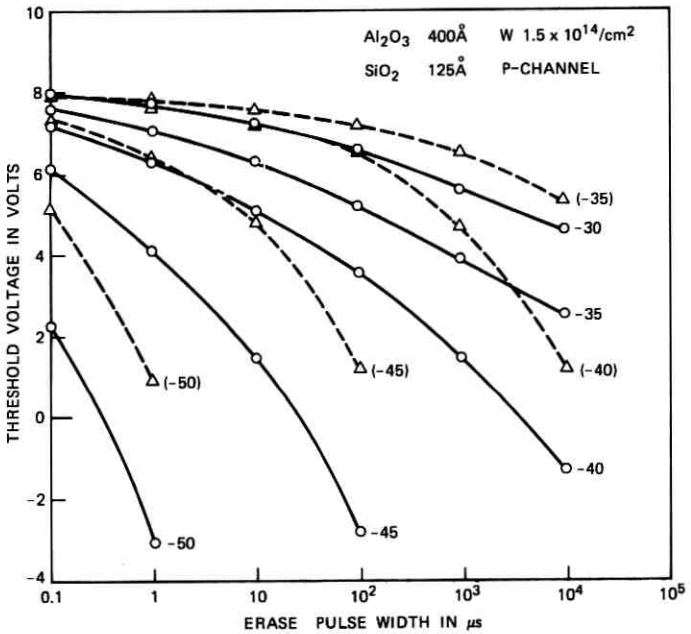



Fig. 8—Shifts in threshold voltage, initially at 8 V, after application of negative-gate-voltage pulses in p-channel IGFETs with dual-dielectric gate insulation of 125-Å-thick $SiO_2$, 400-Å-thick $Al_2O_3$. Solid lines apply to structure with $1.5 \times 10^{14}/cm^2$ of W at dual-dielectric interface; dashed lines apply to structures with no interfacial dopant.

as in earlier comparisons with interfacial-dopant concentration at a larger value.

It is well known that MNOS cells do not function well when the $SiO_2$ layer thickness is larger than 50 Å. This is not the case when the interfacial dopant is introduced. Figure 9 shows write-erase characteristics of MNOS cells with a 500-Å-thick $Si_3N_4$ layer and a 100-Å-thick $SiO_2$ layer. Not only does this cell function well with $1.5 \times 10^{15}/cm^2$ of W interfacial dopant but it also shows negative charging. It is virtually impossible to charge MNOS cells negative with any $SiO_2$ layer thickness when no interfacial dopants are present. The mate cell with no interfacial dopant (see Fig. 9) shows some charging with positive charges. However, it is not possible to erase this cell before a catastrophic breakdown sets in. In Fig. 9, an initial value of 8 V in threshold voltage is again used for erase curves and 1 V for the write curves. For cells with no interfacial dopant, the initial threshold voltage was near zero.

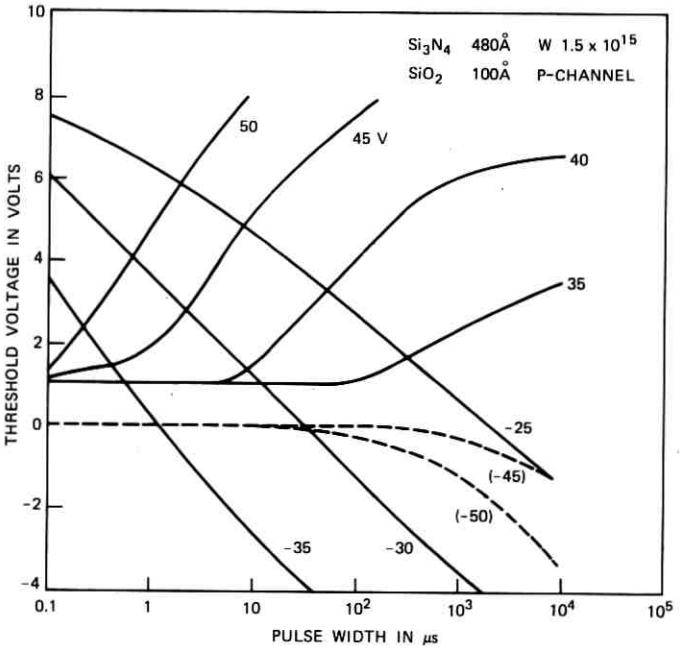



Fig. 9—Shifts in threshold voltages—initially at 1 V after application of positive-gate-voltage pulses, and initially at 8 V after application of negative-gate-voltage pulses—in IGFETs with dual-dielectric gate insulation consisting of 100-Å-thick $SiO_2$ and 480-Å-thick $Si_3N_4$ and containing $1.5 \times 10^{15}/cm^2$ of W at dual-dielectric interface. Dashed lines show shifts in threshold voltages, initially at 0 V, after application of negative-gate-voltage pulses in same IGFETs but with no interfacial dopant.

We have also examined dual-dielectric charge-storage cells with interfacial dopants other than W. Ta induces storage states with a behavior indicative of two distinct energy levels for electron trapping. The shallower level can be filled with electrons and emptied as well. However, the second deeper level can only be filled, or at least it was impossible to completely empty this level. This behavior was observed for both $Si_3N_4$ and $Al_2SO_3$ outer layers. We do not understand the behavior of Ta-induced states well enough at present to warrant further discussion at this time.

We have also examined Ir-induced interfacial states. Their behavior is fairly close to that of W-induced states. Exact comparison, however, requires a further study. It could be conjectured that the dopant-induced states may show in their behavior marked correlation with the relative oxygen affinity of the dopant employed. Ir and W more or less

lack the ability to steal oxygen from $SiO_2$ or $Al_2O_3$ in our thermodynamic environment, while Ta is expected to be oxidized.

## IV. SUMMARY AND CONCLUSION

It is shown in this paper that when suitable interfacial dopants such as W are introduced in a well-defined concentration range, the write-erase characteristics of dual-dielectric charge-storage cells are enormously improved. The upper limit of the dopant is dictated by its influence on the dielectric properties of the outer layer and is determined to lie at about $10^{15}/cm^2$ for $Al_2O_3$. Cells with $5 \times 10^{15}/cm^2$ interfacial dopants showed marked increase in charge leakage. The lower limit is determined to lie at about $10^{14}/cm^2$ by comparison with cells with no interfacial dopants.

The improvement in write-erase characteristics of these cells is of such a magnitude as to allow using relatively thick $SiO_2$ layers (greater than 50 Å) in these cells, which is mandatory for long memory-retention time ( longer than 20 years at 100°C). (Detailed studies of retention time will be published separately.[18]) This study indicates that cells with thinner outer layers (approximately 300 Å of outer layer and approximately 60 Å of $SiO_2$) that would operate at gate-pulse voltages in the 25-V, 1-$\mu$s range, should be feasible.

## V. ACKNOWLEDGMENTS

The authors are grateful to R. C. Beairsto for evaporation of impurities, to L. D. Molnar for $Al_2O_3$ depositions, and to W. R. Costello and F. V. Burckbuchler for $Si_3N_4$ depositions. We also wish to thank T. Buck for the Rutherford back-scattering experiments for quantitative analysis of the interfacial dopants and C. C. Chang for impurity analysis using Auger spectroscopy.

## REFERENCES

1. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," B.S.T.J., *46*, No. 6 (July-August 1967), pp. 1288–1295; D. Kahng, "Semipermanent Memory Using Capacitor Charge Storage and IGFET Read-out," pp. 1297–1300.
2. H. A. R. Wegener, U. S. Patent 3,590,337.
3. D. Kahng, U. S. Patent 3,500,142.
4. R. B. Laibwitz and P. J. Stiles, "Charge Storage on Small Metal Particles," Appl. Phys. Lett., *18*, No. 7 (1 April 1971), pp. 267–269.
5. D. Frohman-Bentchkowsky, "Memory Behavior in a Floating-Gate Avalanche-Injection MOS(FAMOS) Structure," Appl. Phys. Lett., *18*, No. 8 (15 April 1971), pp. 332–334.
6. E. H. Nicollian, A. Goetzberger, and C. N. Berglund, "Avalanche Injection Currents and Charging Phenomena in Thermal $SiO_2$," Appl. Phys. Lett., *15*, No. 6

(15 September 1969), p. 174; E. H. Nicollian and C. N. Berglund, "Avalanche Injection of Electrons into Insulating $SiO_2$ Using MOS Structures," J. Appl. Phys., *41* (June 1970), pp. 3052–3057.
7. Y. Tarui, Y. Hayashi, and K. Nagai, Conf. Solid State Devices, Proc. of Third Conf, 1971, Tokyo, Japan, p. 155; H. Iizuka, T. Sato, F. Masnoka, K. Obuchi, H. Hara, H. Tango, M. Ishikawa, and Y. Takeishi, Fourth Conf., 1972, p. 158.
8. J. R. Szedon, "Charge Instability in Metal-Silicon Nitride-Silicon Oxide-Silicon Structures," IEEE Solid-State Device Res. Conf., Evanston, Ill., June 1966; T. L. Chu, J. R. Szedon, and C. H. Lee, "The Preparation and C-V Characteristics of Si-$Si_3N_4$ and Si-$SiO_2$-$Si_3N_4$ Structures," Solid-State Elect., *10*, No. 9 (September 1967), pp. 897–905; T. L. Chu, "Films of Silicon Nitride-Silicon Dioxide Mixtures," J. Electrochem. Soc., *115*, No. 3 (March 1968), pp. 318–322.
9. See, for instance, extended abstracts on "Electrically Alterable Nonvolatile Semiconductor Memories," Session 4, 1972 Weson Technical Papers.
10. S. Nakanuma, "A Read Only Memory Using MAS Transistors," ISSCC Digest Tech. Papers, February 1970, pp. 68–69.
11. R. E. Honig, "Vapor Pressure Data for the Solid and Liquid Elements," RCA Rev., *23*, No. 4 (December 1962), pp. 567–586.
12. J. L. Margrave, *The Characterization of High-Temperature Vapors*, New York: John Wiley, 1967.
13. A. L. Tyler, private communication.
14. G. H. Frischat, "Evidence for Calcium and Aluminum Diffusion in $SiO_2$ Glass," J. Am. Ceram. Soc., *52*, No. 11 (November 1969), p. 625.
15. C. T. Sah, H. Sello, and D. A. Trewere, "Diffusion of Phosphorus in Silicon Oxide Film," Phys. Chem. Solids, *11*, Nos. 3/4 (October 1959), pp. 288–298.
16. G. A. Brown, W. C. Robinette, Jr., and H. G. Carlson, "Electrical Characteristics of Silicon Nitride Films Prepared by Silane-Ammonia Reaction," J. Electrochem. Soc., *115*, No. 9 (September 1968), pp. 948–955.
17. T. Buck and J. M. Poate, J. Vac. Sc. Tech., *11* (1974), p. 289.
18. K. K. Thornber, D. Kahng, and C. T. Neppell, "Bias-Temperature-Stress Studies of Charge Retention in Dual-Dielectric, Charge-Storage Cells," B.S.T.J., this issue, pp. 1741–1770.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 9, November 1974
*Printed in U.S.A.*

# Bias-Temperature-Stress Studies of Charge Retention in Dual-Dielectric, Charge-Storage Cells

By K. K. THORNBER, D. KAHNG, and C. T. NEPPELL

(Manuscript received February 19, 1974)

*We present a simple, relatively efficient method to predict the nonvolatility of dual-dielectric, charge-storage cells. Using this method, charge-retention times of several hundred years can be predicted unambiguously from experiments of several days' duration made at elevated temperatures and with externally applied, accelerating fields. The method is first presented in the context of a simple, physically reasonable model of the bias and temperature dependence of the characteristic relaxation time of the device. It is then used to analyze a particular device structure. At 80°C, we predict that this device can maintain a flatband-voltage shift in excess of 4 volts for approximately 500 years. Our analysis suggests that this method can be applied to a variety of dual-dielectric, charge-storage cells to predict their nonvolatilities.*

## I. INTRODUCTION

The retention time[1–9] of charge in dual-dielectric, charge-storage cells[10–12] (DDC's) is of central importance in evaluating and comparing these devices. An ideal device would hold its charge at a constant level indefinitely. While, of course, such an ideal device is physically impossible, nonideal, charge-storage memory cells have been fabricated that are expected to have minimum charge-retention times on the order of tens of years. And, as such devices are improved, even longer retention times can be expected. The question that naturally arises is how to evaluate such devices in a few days to predict their charge-retention times.

The primary purpose of this paper is to describe a method of bias-temperature stressing in which the decay of the stored charge is greatly enhanced and, as a result, from which one can predict the charge-

retention time of the device under normal operating conditions (zero bias and room, or somewhat higher, temperature). We stress at the outset, moreover, that our primary concern is charge retentivity and *not* device reliability. Bias-temperature-stress methods are often used to good advantage in reliability studies to artificially reduce the lifetime of the device. Here, we are using similar methods to artificially reduce the retention time of charge stored on the device. For reliability studies, one makes use of models in which the device lifetime is governed by temperature and applied bias. Owing to the complexity of the aging, however, these models are often quite tentative. Here we use a well-defined model in which the rate of decay of the stored charge is governed by temperature and bias. As in reliability studies, experimental results are used to determine the parameters of the model. While the bias dependence of the rate of decay of stored charge has been noted previously,[7] the work we report on here is the first in which bias and temperature stressing have been used simultaneously to predict charge retention under normal operating conditions. We feel obligated, therefore, to elaborate our approach and findings in some detail.

At the heart of our method of prediction are the observations (*i*) that there exists an "envelope" function that sets an upper bound on the total stored charge that can be present in the device at any given time, independent of initial conditions, (*ii*) that this envelope function is determined primarily by the characteristic relaxation of the device, and (*iii*) that this relaxation is a strong function of temperature and electric field. By focusing attention on the envelope function, which properly indicates the long-time decay of the stored charge, we avoid incorrect predictions of decay time based on extrapolations of the initial portion of the charge-decay curve. (If the charge decay is plotted versus log time, such predictions are overoptimistic; if it is plotted versus time, such predictions are overpessimistic.) From measured values of the characteristic relaxation obtained at elevated temperatures and under applied bias, relaxation times of interest at room temperature and under zero bias can be predicted. Owing to the nonlinearities inherent in the charge decay, these zero-bias, room-temperature relaxation times are functions of the initial stored charge; the larger the initial stored charge, the smaller the relaxation time. For example, for one version of our device[13] we find that, for an initial charge corresponding to a flatband voltage of 10 V, a relaxation time of $3 \cdot 10^4$ years is predicted; for 7 V, we predict $6 \cdot 10^4$ years. Using our results, predictions of relaxation times can also be made for devices

operated under bias and at other than room temperature. For example, at 80°C under zero bias we predict for 10 V a relaxation time of 100 years and for 7 V 300 years. Linear extrapolations of the initial portion ($t \ll \tau_{relax}$) of the charge-decay curve would result in misleading estimates of charge-retention times on the order of $10^{20}$ years.

It is important at the outset to stress that we are concerned in this paper with devices[10] with "thick" oxide layers; that is, with oxide layers between the semiconductor and the charge-storage sites sufficiently thick that the rate of decay of charge via back-tunneling through the oxide layer is small compared to the rate of decay of charge through the insulator layer between the storage sites and the gate. While one can accelerate the back-tunneling charge decay of thin-oxide devices[7] by applying an electric field, we believe that, unless one has an exceptionally good means of characterizing the specific tunneling processes of interest, such information is of little direct value in estimating the decay in the absence of such field, i.e., under normal operating conditions. The applied voltage, of course, can accelerate or decelerate decay through the oxide layer or the insulator layer, depending upon its polarity. If electrons are stored, a positive gate voltage enhances the decay in thick-oxide devices because the decay is through the insulator, whereas the same voltage reduces the decay in thin-oxide devices[7] where the primary decay is back through the oxide. One reason for studying thick-oxide devices is to determine the limits on the retention time of charge-storage devices imposed by charge decay through the insulator layer. Such limits are of considerable interest, especially for devices whose back-tunneling decay is sufficiently low that, based on this decay mechanism alone, one might exaggerate the device's charge-storage capabilities.

In this paper we first discuss in general terms the physical processes that lead to the decay of the stored charge. A relatively thick oxide layer is used so that the primary discharging current is through the insulator layer. This current through the insulator is found to be characterizable as a thermally activated flow of charge via defects with a very low, but strongly field-dependent, mobility. We then discuss how this strong temperature and electric-field dependence of the decay current can be used to controllably accelerate the discharging of the DDC. A simple, physically reasonable, empirical model is introduced to explain the experimentally measured decay of the flatband-voltage shift in time with temperature and externally applied bias as parameters. The mathematical expression for the decay current is sufficiently simple that all quantities of interest, particularly device relaxation

times, can be calculated analytically. From the high-temperature, high-accelerating-field results, it is possible to predict low-temperature, zero-bias behavior. We can thus unambiguously determine nonvolatility on the order of tens or hundreds of years on the basis of experiments performed in several days or weeks.

Before proceeding, it must be emphasized that the method of bias-temperature stressing that we develop here is not limited to devices which behave according to our simple empirical model. Indeed, as we point out in Appendix A, the central ideas at the heart of the method are valid, within certain limits, to a variety of models: the method is, within certain bounds, insensitive to the details of the actual decay process. We discuss the physical processes that we believe are operative in our devices first, however, because this will permit the reader to familiarize himself with the actual devices to which our method is then applied. Although our results indicate the general features of a particular, detailed decay model, we emphasize that we are not primarily concerned with establishing the existence of such a model.

## II. PHYSICAL PROCESSES

The structure and fabrication of the memory devices studied here have been described in detail elsewhere.[13] For our purposes here, it suffices to note that a typical device consists of the following layers (see Fig. 1): a metallic contact (taken fixed at zero voltage), n- (or p-) type silicon, 70 to 200 Å of $SiO_2$ (referred to as the oxide layer, or simply the oxide), a set of suitable dopant-induced storage states (referred to as the storage states), 400 to 500 Å of $Al_2O_3$ or $Si_3N_4$ (referred to as the insulating layer, or simply the insulator), and a metallic contact (referred to as the gate). Inclusion of the specific states, which are the interfacial dopant-induced states, makes it possible to store either electrons or holes, whichever is desired, as well as to provide very well-defined storage sites for the elementary charges. The oxide and the insulator are sufficiently thick that tunneling from one side of either to the other side can be neglected.

The device can be charged either negative or positive as follows. Negative charge can be stored by driving the gate to a high positive voltage ($\approx 50$ V) for a short period of time ($\approx 100$ $\mu$s). Some electrons in the Si tunnel into the conduction band of the oxide, traverse the oxide, and then are trapped in the storage states or pass completely through the device to the gate electrode. Positive charge is stored by driving the gate negative to force electrons out of the storage states, through the oxide, and into the semiconductor.

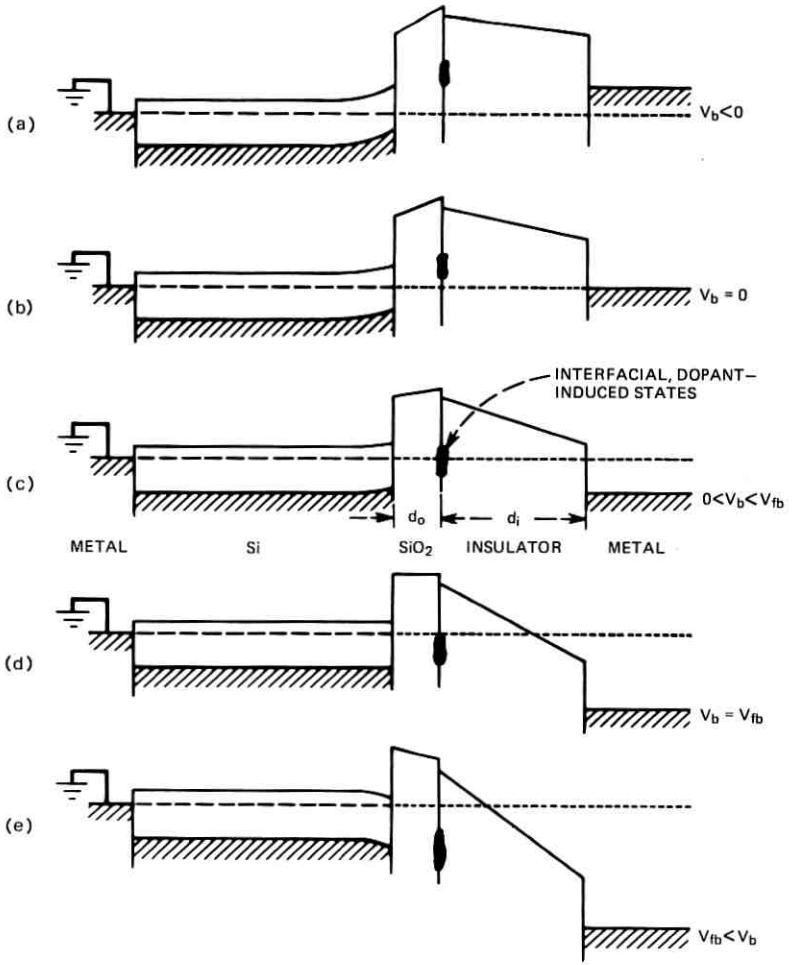



Fig. 1—Schematic of the structure of the MIOS memory device under five different, applied-bias conditions. The location of the interfacial, dopant-induced, charge-storage states is indicated. They are assumed to be charged negative.

Our experimental results indicate that these multilayered memory devices discharge as follows. (Refer to Fig. 1 in which a device is shown under five different biases. Note, in particular, Fig. 1a, in which the semiconductor is inverted near the $SiO_2$.) The (negative) stored charge produces large electric fields in both the oxide and the insulator. In the first few hundred milliseconds, some fraction of the charge is lost through thermionic emission or tunneling into the conduction bands of

the oxide and/or the insulator. For most of the stored charge, however, the trapping levels are sufficiently deep that these processes are very improbable. Charge can tunnel into traps in the insulator, and through such imperfections a discharging current can pass into the conduction band of the insulator and flow to the gate. Some charge can also pass completely through the insulator in the forbidden band by means of these traps. In addition, in a similar manner, a discharging current can be present in the oxide and flow to the Si.

The size of the discharging current is expected to be strongly field dependent.[1–9] In addition, we note that increasing the (positive) gate voltage enhances the electric field in the insulator and decreases this field in the oxide. The manyfold increase in the decay current which occurs when a positive voltage is applied to the gate thus implies that the primary discharging current is passing through the insulator. Inversely, when a negative voltage is applied to the gate, a decrease in the decay current is observed. In this case, the change in the applied fields is such as to increase the current through the oxide and decrease the current through the insulator. A detailed study of the reverse gate-bias decay was not undertaken because of the long time intervals required even at high temperatures. For the present, it is sufficient to note that the total decay current increases with positive voltage and decreases with negative voltage applied to the gate.

The physical details of the transport of the charge within the insulator can be narrowed down as follows. The strong temperature enhancement of the decay of the charge indicates that thermal activation rather than tunneling plays the dominant role in enabling charge to be transferred toward the gate. The question of whether the charge passes through the insulator to the gate entirely within the forbidden band, or partially in the conduction band, is more difficult. The relatively low thermal activation energies observed ($\approx 0.6$ eV) suggest that the rate limiting portion of the transport is not associated with the conduction band. If it were, much higher decay currents would be expected. We infer, therefore, that the current is controlled by the traps in the bulk of the insulator.

For charge transport from one trap to the next, one might expect the current to be proportional to $\exp[(E_o + q\mathcal{E}a/2)/kT]$, according to Poole's law.[14] (Here $E_o$ is an ionization energy, $\mathcal{E}$ is the electric field in the insulator, and $a$ is the intertrap spacing.) However, Poole's law, or its extension and modification by Frenkel[15,16] and others[17–19] include only the thermal excitation of carriers and do not include velocity-versus-field effects connected with the transport from one site

to the next.[20] Current density is proportional to $qnv$, where $n$ is the number of charges per unit volume and $v$ is the carrier velocity. One expects $n$ to be thermally activated; however, in an insulator, $v$ can be a rapidly varying function of applied fields,[21–23] whether electron transport is in the conduction band or in a defect "band," as seems to be the case here. Thus, whereas for pure Frenkel-Poole behavior the excitation of electrons is the rate-limiting, field- and temperature-modulated process, for our devices both $n$ and $v$ are so modulated.

Our experimental results indicate that the discharging current is proportional to the empirical expression

$$\mathcal{E}_o(T) \exp(-E_a/kT) \exp[+\mathcal{E}/\mathcal{E}_o(T)], \tag{1}$$

where $\mathcal{E}_o(T)$ is a very slowly varying, *decreasing* function of the temperature $T$. To the accuracy of the experiments, the activation energy $E_a$ is independent of the applied field. The decrease of $\mathcal{E}_o(T)$ with temperature could be due to an increased overlap between polarized electronic states because of increased lattice vibrations. The electric field dependence can in general be expected to be more complicated than that given in eq. (1). Fortunately, eq. (1) suffices for our purposes, partially because the decay curves are relatively independent of the detailed dependence of the discharging current on electric field. See Appendix A for a discussion of more complicated field dependences.

One final point should be made regarding the decay current. As part of the assumption that the rate-limiting process controlling the charge decay is the trap-controlled current within the bulk of the insulator, we have assumed that the number of carriers in transit is independent of time. In other words, as a trap in close proximity to the storage states loses its charge to a trap somewhat closer to the gate, the emptied trap is quickly refilled with a charge from a storage state. Thus, the average time for a charge to pass from a storage state to an empty trap is much less than the average time for a charge to pass from one trap to the next. This results from the high density of storage states, and makes the effective number of carriers available for transit independent of the magnitude of the stored charge and hence independent of time. This is reasonable throughout most of the decay owing to the large number of stored electrons. Near the end of the decay, however, when the amount of stored charge is much less than its initial value, the rate-limiting process may become the transfer of charge from the storage states into the insulator. We ignore this effect, since, by the time it becomes important, the fundamental relaxation time of the device will be well determined.

From the foregoing discussion, it is evident that we can artificially enhance the decay greatly by increasing the temperature and the electric field in the insulator. In so doing, however, we must be very careful that we do not excessively enhance charge-transport mechanisms that under normal operating conditions would play no significant role. For example, by applying too large a potential at the gate, the field in the oxide will be enhanced to such an extent that the storage layer will be slowly charged by electrons tunneling from the Si into the conduction band of the oxide. Excessive temperatures, on the other hand, may enhance decay into the insulator's conduction band, exaggerating this mode of charge transport. These effects are often easily detected experimentally, and they give rise to upper limits on the increases in temperature and electric field possible to enhance charge decay for measurement purposes.

Although the decay of charge from the storage states can be enhanced by increasing either the temperature or the insulator electric field, the enhancement from either is insufficient to bring the decay times into the range of days. However, this can be achieved by combining higher temperature with higher fields. Of course, we again must be careful to avoid introducing decay processes that under normal conditions would be insignificant. (An example of this would be Schottky emission.) Nonetheless, we find that we can obtain the entire charge decay curve within a few days (at most) without introducing extraneous decay mechanisms. From these curves, we can predict the normal decay curve and hence the charge-retention time of the memory element. In this way we avoid having to determine this retention time by extrapolating the initial portion of the time dependence of the charge decay. In the next section we see how the nonlinear dependence of the rate of decay on the quantity of stored charge leads to characteristic, charge-decay curves from which one can predict the nonvolatility of DDC's.

In those cases where certain alternative conduction mechanisms are activated, this method may or may not work. In some devices, the charge decay is via Fowler-Nordheim tunneling from the storage states into the conduction band of the oxide or the insulator, or it is via direct tunneling from these states into the semiconductor. Neither of these decay currents will be affected by changes in the temperature; however, both such currents will be strongly modulated by applied bias. In the case of Fowler-Nordheim tunneling, our method is directly applicable. For direct tunneling, however, in which the primary decay is via Si-$SiO_2$ interfacial states, we must be very careful as we increase

the bias not to enhance the tunneling into semiconductor states in the valence or conduction band, which under normal conditions would play no important role in the discharging of the DDC. Perhaps one can modulate the decay by altering the thickness through which the electrons tunnel, either by physically squeezing the sample or by constructing samples of varying thicknesses.

### III. EMPIRICAL MODEL AND MEASUREMENT TECHNIQUE

Let us assume that we have some (negative) charge stored in the interfacial, dopant-induced states. The quantity of charge stored in these states is proportional to the measurable flatband voltage $V_{fb}$. Also proportional to $V_{fb}$ is the field in the insulator under zero-bias conditions. In the presence of a finite bias voltage $V_b$, an additional field proportional to the bias voltage will be linearly superimposed on the zero-bias field (with a *different* constant of proportionality). Thus, using expression (1) and expressing fields in terms of voltages, we write for the time rate of change of $V_{fb}$, the flatband voltage,

$$\frac{dV_{fb}}{dt} = -\frac{V_s(T)}{\tau(T)} \exp\left[\frac{V_{fb} + bV_b}{V_s(T)}\right], \tag{2}$$

where $V_s$ and $\tau$ are experimentally determined functions of temperature only and $b$ is a relative constant of proportionality (see Appendix B). Equation (2) presumes decay via the insulator; for decay through the oxide, one would have the factor $(V_{fb} - V_b)$ in place of the factor $(V_{fb} + bV_b)$. This empirically satisfactory expression is reasonable physically, as explained in Section II. It also permits a simple, analytic treatment of all significant features of the decay. Should the flatband voltage decay according to a relation other than (2), one can still employ bias-temperature stressing. This is indicated in Appendix A.

Before solving eq. (2), we should clarify two points. (*i*) We do not require eq. (2) to be valid for $t < 1$ minute. It may be valid there but, being interested in the long-time decay of the charge, all we need to do is to integrate eq. (2) from some time $t_1$, on the order of minutes after charging, when the experiments are begun. (*ii*) Since $dV_{fb}/dt$ does not vanish as $(V_{fb} + bV_b) \to 0$, eq. (2) is clearly not valid as $t \to \infty$. As remarked in Section II, this difficulty results from assuming that there are always a large number of stored charges, whereas in fact near the end of the decay this number becomes small. Again, this difficulty need not concern us, since we do not have to consider the leakage after the bulk of the stored charge has decayed away. (See Appendix A for a

more general treatment.) With these limitations in mind, we proceed to solve eq. (2).

The solution to eq. (2) follows at once after a separation of variables and an integration from $(t_1, V_1)$ to $(t, V_{fb})$. The resulting $V_{fb}(t)$ for the decay of the flatband voltage as a function of time $t$ at temperature $T$ and applied bias $V_b$ is given by

$$V_{fb}(t) = V_s(T) \log_e \{(\exp [-V_1/V_s(T)] + [(t - t_1)/\tau(T)] \cdot \exp [bV_b/V_s(T)])^{-1}\}. \quad (3)$$

For a single value of $T$, this function is sketched in Fig. 2 as a function of $\log_{10} (t/t_1)$ for several $V_b$ and for hypothetical, illustrative values of $V_s$, $\tau$, and $b$. The purpose of plotting as a function of $\log_{10} (t/t_1)$ is to call attention to the *actual* long-time behavior of $V_{fb}(t)$. For $t_1 < t \ll \tau(T) \exp [-(V_1 + bV_b)/V_s(T)]$, $V_{fb}(t)$ is relatively flat. However, for $t \gg \tau(T) \exp [-(V_1 + bV_b)/V_s(T)]$, $V_{fb}(t)$ is given approximately by $V_e(t)$ where, if $\dot{t}_1$ is any convenient scale factor,

$$V_e(t) = V_s(T) \log_e [\tau(T)/\dot{t}_1] - bV_b - V_s(T) \log_e (t/\dot{t}_1). \quad (4)$$

The function $V_e(t)$ is the previously mentioned envelope function for each $T$ and $V_b$. It is shown dashed in Fig. 2. (Note $\log_e x = \log_{10} x / \log_{10} e$.) Although $\dot{t}_1$ may be assigned the same value as $t_1$, the two

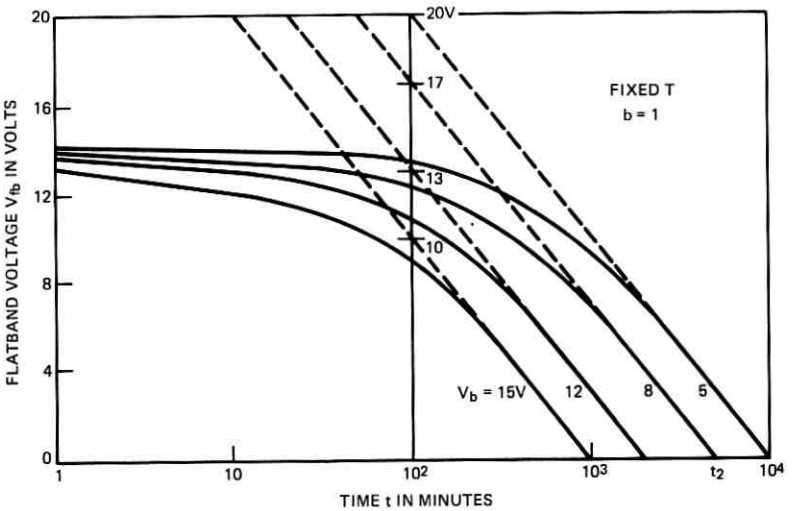



Fig. 2—Schematic illustrating a possible decay of the flatband voltage $V_{fb}(t)$ as a function of $\log_{10} (t/t_1)$ in a hypothetical device for four different applied biases $V_b$ at a fixed temperature $T$. For simplicity, we set $b = 1$. Also shown are the envelope functions $V_e(t)$ associated with each $V_b$.

must be distinguished. Our $t_1$ refers to an initial condition, while $V_e(t)$ is independent of the initial conditions, $t_1$ and $V_1$, and satisfies $V_{fb}(t) < V_e(t)$ (see Appendix C) and

$$V_{fb}(t) \approx V_e(t), \quad t \gg \tau \exp\left[-(V_1 + bV_b)/V_s\right]. \tag{5}$$

When $V_e(t)$ is plotted versus $\log_{10}(t/t_1)$, $V_s$ is the slope and $\tau$ is the intercept $[V_e(\tau) = 0]$ for zero bias ($V_b = 0$). From the experimental $V_{fb}(t)$ curves obtained for several $V_b$ at a given $T$, it is possible to readily determine $\tau(T)$, $V_s(T)$, and $b$, the three parameters in eq. (4). From the envelope function $V_e(t)$ drawn tangent to $V_{fb}(t)$ as shown in Fig. 2, we obtain $V_s(T)$ from the slope of $V_e(t)$, $b$ from the difference between $V_e(t)$ (for fixed $t$) at different $V_b$, and $\tau(T)$ from the value $t_2$ of $t$ where $V_e(t) = 0$:

$$\log_e\left[\tau(T)/t_1\right] = \log_e\left(\frac{t_2}{t_1}\right) + \frac{bV_b}{V_s(T)}. \tag{6}$$

In Fig. 2, $t_2$ for $V_b = 8$ V is shown ($t_2 = 5 \cdot 10^{+3}$ minutes). It is a test of the form of eq. (2) that $V_s(T)$, $b$, and $\tau(T)$ be independent of $V_b$ to within experimental accuracy. That this is indeed the case is discussed in Section IV. We now repeat the above for other temperatures in the range $150°C \leqq T \leqq 300°C$. We find that both $\tau(T)$ and $V_s(T)$ appear to be "thermally activated":

$$\tau(T) = \tau_o \exp(E_a/kT), \quad V_s(T) = V_{so} \exp(E_s/kT). \tag{7}$$

Since the slope $V_s$ may arise from a combination of physical processes, identifying it with a thermally activated process may be somewhat misleading. (The parameter $b$ is a weakly decreasing function of $T$. Since we are interested only in predicting zero bias ($V_b = 0$) behavior, we need not concern ourselves further with extrapolating $b$ to room temperature. In other words, we can for each temperature use the $b$ measured at that temperature to extrapolate to zero bias. Once we know the zero-bias result, we can extrapolate to the desired temperature. Since $b$ enters only as $bV_b$ and $V_b$ is zero, this latter extrapolation can be performed without knowledge of b.) The quantities $\tau_o$, $E_a$, $V_{so}$, and $E_s$ are obtained from the usual $(1/T)$ plots. We assume, and indeed from our discussion in Section II it is not unrealistic to expect, that we may extrapolate $\tau$ and $V_s$ to room temperature using (7). We must now relate these quantities to the characteristic relaxation time $\tau_{\text{relax}}$, the charge retention time, of the storage device.

In Fig. 3 we define $\tau_{\text{relax}}(\tau_r)$ graphically, again for hypothetical device parameters. It is simply the "roll-off" time of $V_{fb}(t)$: $V_e(\tau_{\text{relax}}) = V_1$, where $V_1 = V_{fb}(t_1)$. [As shown in Appendix C, $V_1 \approx V_{fb}(0)$

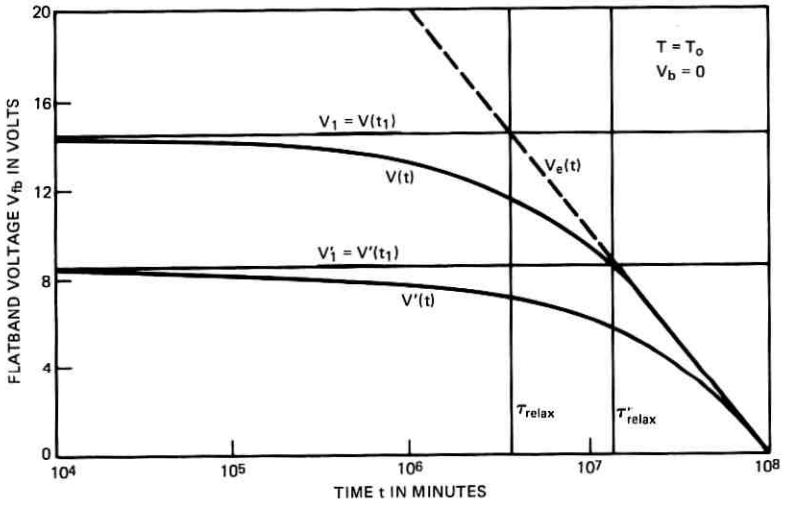



Fig. 3—Schematic illustrating a possible decay of the flatband voltage $V(t)$ in a hypothetical device for room temperature ($T = T_o$) and zero applied bias ($V_b = 0$) for two different quantities of stored charge, $V_1$ and $V_1'$. The size of the characteristic relaxation time $\tau_{relax}$ is indicated for both cases. Note that $V_1 - V_1^o = V_1' - V_1'^o = 0.69\ V_s$.

for $t_1 \ll \tau_{relax}$.] Solving (4) for $\tau_{relax}$ at room temperature $T_o$ for $V_b = 0$, we obtain

$$\tau_{relax} = \tau(T_o) \exp\left[-V_1/V_s(T_o)\right]. \tag{8}$$

Indicative of the nonlinearities inherent in the charge-decay process, $\tau_{relax}$ is a function of the initial stored charge, and hence the initial flatband voltage $V_1$, increasing as $V_1$ decreases. (Compare $\tau_{relax}$ and $\tau'_{relax}$ corresponding to different amounts of initial stored charge $V_1$ and $V_1'$, respectively, in Fig. 3.) Again we emphasize that $\tau_{relax}$ cannot be obtained from (linearly) extrapolating $V(t)$ based on its behavior for $t \ll \tau_{relax}$. Since $\tau_{relax}$ is on the order of $10^4$ years at room temperature for devices reported on here, it is doubtful whether room-temperature, zero-bias experiments alone will be able to predict $\tau_{relax}$.

Although $\tau_{relax}$ provides a measure of the charge retention time of the device, we must, in addition, know how far $V_{fb}$ has decayed below $V_1$ by the time $\tau_{relax}$ before full significance can be attached to this definition of characteristic decay time. Returning to eq. (3) and letting $t_1 = 0$, $V_1 = V_1^o$, $V_b = 0$, we obtain for $V_{fb}(t)$ under zero bias

$$V_{fb}(t) = V_s \log_e \left[ e^{+V_1^o/V_s} \left(1 + \frac{t}{\tau_{relax}}\right)^{-1} \right]. \tag{9}$$

In arriving at eq. (9), we have used eq. (8). Thus,

$$\begin{aligned} V_{fb}(\tau_{\text{relax}}) &= V_1^o - V_s \log_e 2 \\ &= V_1^o - 0.69 V_s. \end{aligned} \tag{10}$$

We thus obtain the very important result that, at the roll-off time $\tau_{\text{relax}}$, $V_{fb}$ has dropped $0.69V_s$ volts below its initial value. This amount is independent of $V_1^o$, the initial flatband voltage. Of course, $V_{fb}(\tau_{\text{relax}})$ must be several volts positive so that eq. (2) will still be valid. In addition, $0.69V_s$ thus defines the decay that must be taken into account when including this device in circuits. In determining $V_s$ from the envelope of a decay curve, one must be careful to convert from $\log_{10}$ to $\log_e$:

$$V_s = \frac{V_e(t_1) - V_e(t_2)}{\log_{10}(t_2/t_1)} \times 0.434,$$

where $0.434 = \log_{10} e$.

Another characteristic decay time is the time at which the flatband voltage drops below a certain margin voltage $V_m$. We refer to this time as the margin time, $\tau_{\text{margin}}$, or simply $\tau_m$. It follows at once from eq. (9) that $\tau_m$ as a function of $V_m$ and $V_1$ is given by

$$\tau_m = \tau(T)\{\exp[-V_m/V_s(T)] - \exp[-V_1/V_s(T)]\} \tag{11}$$

as a function of $T$ under zero bias conditions. A plot of $\tau_{\text{margin}}$ is presented in the next section.

## IV. EXPERIMENTAL RESULTS

Bias-temperature-enhanced, charge-relaxation experiments were carried out on the recently devised dual-dielectric, charge-storage cells.[13] Aluminum oxide ($Al_2O_3$, $\epsilon = 9$) was used as the insulator. The thickness $d_o$ of the $SiO_2$ ($\epsilon = 4$) was 125 Å, and that of the $Al_2O_3$, $d_i$, 550 Å. Thus, $b = C_o/C_i$ [see eq. (25)] is predicted to be given by

$$b = \frac{C_o}{C_i} = \frac{\epsilon_{SiO_2} A}{d_o} \frac{d_i}{\epsilon_{Al_2O_3} A} = 2.0.$$

Experimental values for $b$ ranging from 0.8 to 1.2 were obtained over the temperature range of $150°C \leqq T \leqq 300°C$. The reason for this discrepancy is not known. Some possible explanations are discussed in Appendix D. The area of the devices was about $1 \cdot 10^5$ $\mu m^2$, and the doping of the $n$-type silicon was about $6 \cdot 10^{15}$ per $cm^3$. Tungsten of a density of about 2 to $3 \cdot 10^{14}$ per $cm^2$ was used to produce the interfacial storage sites.

**Table I**

| Sample* | $T$ (°C) | $1/T \cdot 10^3$ | $V_b$ (volts) | $V_e$ (volts) | $t_2$ (minutes) | $b$ | $\tau(T)$ (minutes) |
|---|---|---|---|---|---|---|---|
| H15-4 | 150 | 2.36 | 15 | 2.44 | $2.6 \cdot 10^4$ | 1.2† | $3.75 \cdot 10^7$ |
| H15-1 | 200 | 2.11 | 10 | 1.93 | $1.3 \cdot 10^4$ | 1.12 | $4.3 \cdot 10^6$ |
| H15-1 | | | 15 | 1.92 | $7.7 \cdot 10^2$ | | |
| H15-4 | 250 | 1.91 | 7 | 1.76 | $2.7 \cdot 10^4$ | 1.03 | $1.7 \cdot 10^6$ |
| H15-4 | | | 10 | 1.76 | $4.7 \cdot 10^3$ | | |
| H15-1 | 300 | 1.74 | 5 | 1.53 | $4.6 \cdot 10^3$ | 0.8 | $6.2 \cdot 10^5$ |
| H15-1 | | | 10 | 1.52 | $3.3 \cdot 10^2$ | | |

* H15 is the wafer label and 1 or 4 is the chip label.
† Estimated value.

Table I presents a summary of the experimental results which we have analyzed most carefully. Four representative charge decay curves are plotted in Fig. 4. The envelope function $V_e(t)$ associated with each curve is shown as well. From the $V_e(t)$ we obtain the data given in Table I. One might do better by attempting to fit the actual experimental curves with eq. (3), but we have found $V_e(t)$ adequate for our purposes.

In studying Fig. 4 one may wonder why the slope of the decay curves seem to diminish at higher temperatures. In fact, as is clear from eqs.

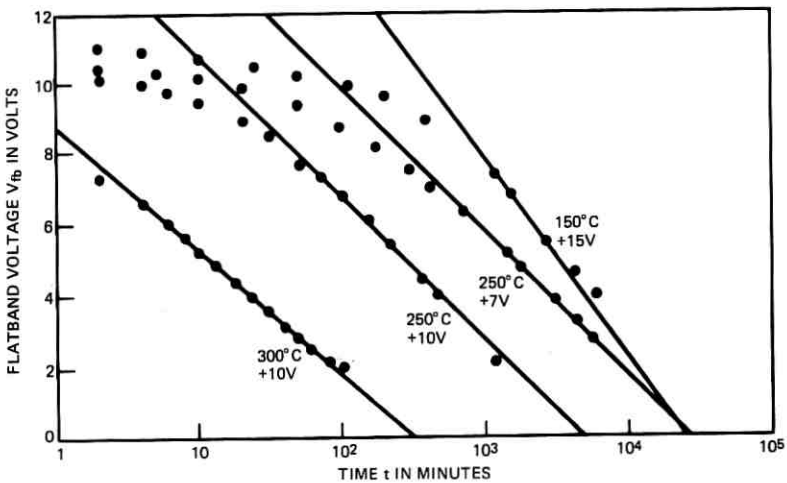



Fig. 4—Experimental results under four different conditions: $T = 150°C$, $V_b = 15$ V; 250°C, 7 V; 250°C, 10 V; 300°C, 10 V. In each case, the element was charged by setting $V_b = 50$ V for 100 $\mu$s.

(2) and (7), $|dV_{fb}/dt|$ increases as $T$ increases, as one would expect. One must remember that we are plotting $V_{fb}(t)$ versus $\log_{10}(t/t_1)$. Thus, the enhanced decay has shifted $V_{fb}(t)$ to lower $t$, reducing the apparent decay of $V_{fb}$ when plotted on a $\log_{10} t$ scale. The authors must admit to having expended some effort themselves getting used to plotting $V_{fb}$ versus $\log t$ rather than $\log V_{fb}$ versus $t$, as is common in many decay problems.

By plotting the values of $\tau$ and $V_s$ given in Table I as functions of $1/T$, the activation energies $E_a = 0.61$ eV for $\tau$ and $E_s = 0.065$ eV for $V_s$ can be determined [see Fig. 5 and eq. (7)]. From Fig. 5 we can also calculate $\tau_o$ and $V_{so}$ [eq. (7)]: $\tau_o = 2.0$ minutes and $V_{so} = 0.41$ V. Using the $E_a$, $\tau_o$, $E_s$, $V_{so}$ so determined, we use eq. (7) to extrapolate $\tau(T_o)$ and $V_s(T_o)$, where $T_o$ is room temperature (290°K, 17°C). We find $\tau(T_o) = 7.7 \times 10^{10}$ minutes and $V_s(T_o) = 5.5$ V. With these values, we can predict the room-temperature, zero-bias relaxation times using eq. (8).

The activation energy $E_a$ of 0.61 eV is comparable to activation energies ranging from 0.4 to 1.2 eV reported for Frenkel-Poole conduction through oxide layers.[24–28] We conclude, therefore, that $E_a$ may be interpreted as being associated with an activation process. Although $E_s$ is relatively small, being on the order of several kT's, its existence does result in an appreciable variation of $V_s$ with $T$ (as observed experimentally). We do not have a simple independent quantitative explanation of its magnitude and caution the reader against interpreting $E_s$ as being associated with a simple, thermally activated process.

As discussed in Section III, the charge-retention time of our dual-dielectric, charge-storage cell is well-characterized by the device "roll-off," or relaxation time, as defined in Fig. 3. Owing to the nonlinear dependence inherent in the charge decay, $\tau_{relax}$ depends upon the "initially" stored charge, which we denote in terms of flatband voltages by $V_1 = V_{fb}(t_1)$. In Fig. 6 we plot $\tau_{relax}$ as a function of $V_1$ for several temperatures of interest. For room temperature and a $V_1$ of 10 V, $\tau_{relax}$ is $3 \cdot 10^4$ years; for 7 V, $6 \cdot 10^4$ years. That $\tau_{relax}$ is a function of stored charge should be carefully noted. Lastly, we note that, for room temperature at $\tau_{relax}$, $V_{fb}$ has dropped $(0.69)(5.5) = 3.8$ V below $V_1$. At 80°C the drop is only 2.4 V owing to the reduced value, of $V_s$ (=3.4 V at 80°C). However, now $\tau_{relax}$ for 10 V is only 100 years, and for 7 V, 300 years.

In Fig. 7 we plot $\tau_{margin}$ as a function of $V_1$ for zero bias and for the same temperature used in Fig. 6. As is obvious physically, the larger

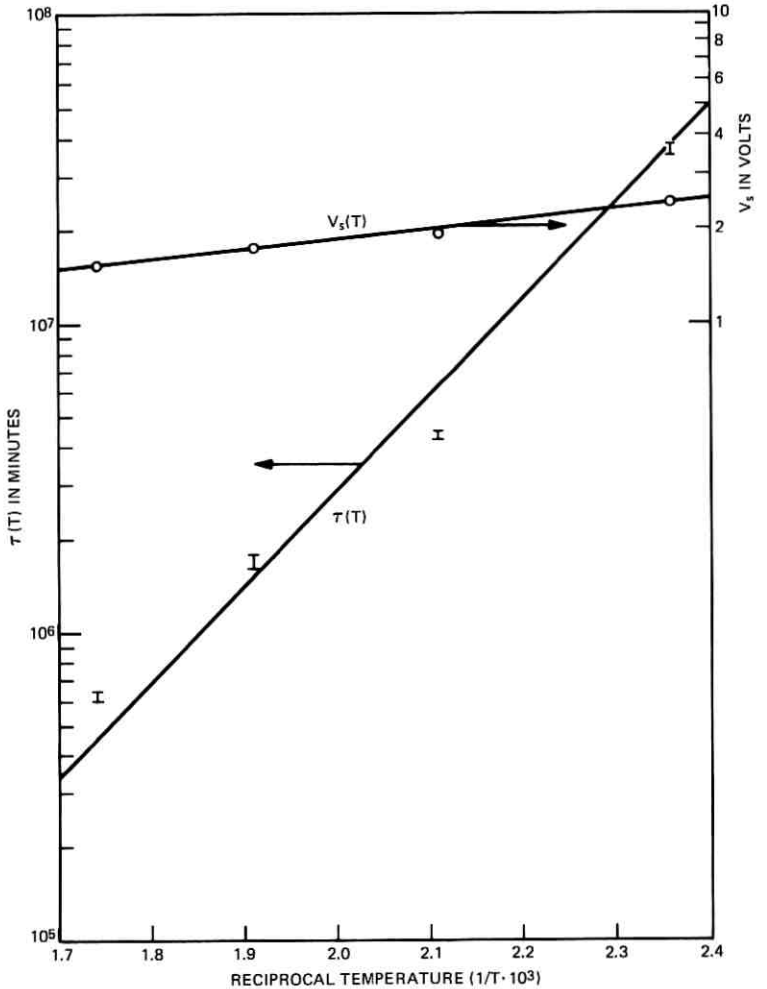



Fig. 5—Experimentally determined parameters $\tau$ and $V_s$ plotted versus $1/T$ to determine the activation energies $E_a$ and $E_s$.

the initial stored charge (initial flatband voltage $V_1$), the longer the time required for the flatband voltage to drop to the margin voltage $V_m$, here taken to be 4 V. The horizontal lines give the upper limit or largest possible $\tau_{\text{margin}}$ for the given temperature. For $V_1$ between 4 and 5 $V$, $\tau_m$ goes rapidly to zero, as is clear from Fig. 3. For $V_1$ less than 4 $V$, $\tau_m$ is obviously meaningless.

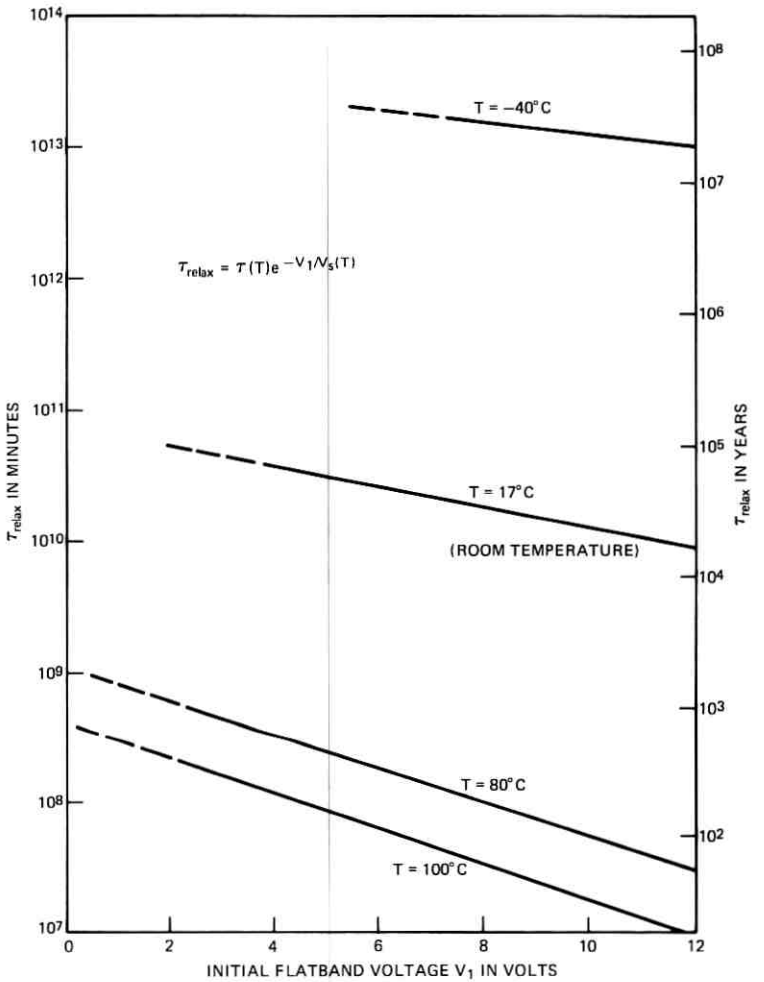



Fig. 6—The predicted roll-off or device relaxation time for zero-applied bias ($V_b = 0$) and temperatures of −40°C, 17°C, 80°C, and 100°C plotted as a function of "initial" flatband voltage $V_1 = V(t_1)$, where $t_1 = 1$ minute.

## V. CONCLUSION

In this paper we have discussed a method of predicting the retention time of charge in dual-dielectric, charge-storage cells. The method is based on the realization that an envelope function $V_e(t)$ exists, that this function is determined primarily by the characteristic relaxation of the device, and that this relaxation is a strong function of tempera-

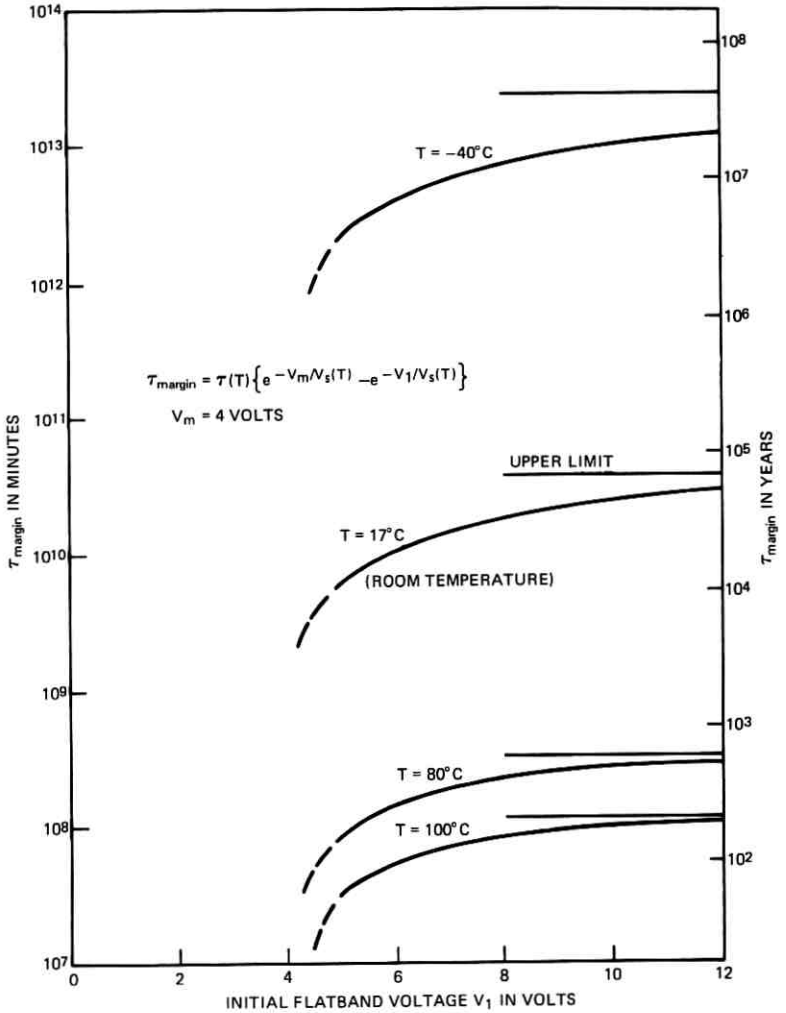



Fig. 7—The predicted margin time for zero-applied bias ($V_b = 0$) and temperatures of −40°C, 17°C, 80°C, and 100°C plotted as a function of "initial" flatband voltage $V_1 = V(t)$, where $t_1 = 1$ minute. The horizontal line above each curve is the upper limit for $\tau_{margin}$ as $V_1$ is taken larger and larger.

ture and applied bias. By choosing an appropriate function to represent the dependence of the discharging current on the applied bias and stored charge, it is possible to discuss all the interesting features of the time-dependent charge decay in terms of simple mathematical expressions. We were thereby able to predict, on the basis of experimental

data obtained under bias-temperature stressing of representative devices, the time dependence of the shift in the flatband voltage at zero bias and operational temperatures (room temperature to 80°C). In particular, we can characterize the nonvolatility of the DDC's in terms of a roll-off or relaxation time $\tau_r$ and a margin time $\tau_m$, both dependent on the level of initial charging. Although these times are on the order of $10^4$ years, they were determined from experimental data taken over a period of only a few weeks. This was possible because our method of bias-temperature stressing greatly enhances the dominant discharging processes, speeding up the charge decay without introducing significant decay from otherwise insignificant decay modes. Because of this, we feel that our method will greatly facilitate the realistic prediction of the nonvolatility of DDC's.

## VI. ACKNOWLEDGMENTS

We wish to thank D. M. Boulin for his assistance in preparing and charging the devices studied and W. J. Sundburg for his measurements of insulator and oxide layer thicknesses.

## APPENDIX A

Equations (1) and (2) in the text are admittedly oversimplified, despite the ease they render in interpreting the data. Nonetheless, the characteristic decay $V(t)$ discussed in the text is approximately obtained in a variety of cases, e.g., tunneling or Frenkel-Poole, even when it is known that the rate of decay is proportional to more complicated voltage dependences than exp $(V/V_s)$. It is the purpose of this appendix to outline why this is so. For simplicity, we assume that conditions are such that only one decay mode is important.

Usually we can write the decay of the stored charge, or, equivalently, that of the flatband voltage $V$, in the following form,

$$\frac{dV}{dt} = -[f(V + bV_b) - f(0)], \tag{12}$$

where, in turn,

$$f(V') = Ae^{g(V')}, \quad V' = V + bV_b, \tag{13}$$

$A$ being a dimensional constant and $g(V')$ a function of temperature and applied bias as well as of various powers of $V'$ (and possibly of $\log_e V'$). The point of writing $f(V')$ in the form given in (13) is to focus primary attention on the exponential nature of the functional depen-

dence of $f(V')$ on $V$. Thus, $f(V')$ can be approximated over a wider range of $V$ by a Taylor expansion of $g(V')$ in (13) than by a Taylor expansion of $f(V')$ directly.

Suppose at $t_1 \ll \tau_{\text{relax}}$, $V = V_1$. Then, assuming $f(0) \ll f(V')$ for $V$ of interest, we can readily integrate (12) if we approximate $g(V')$ by $g(V_1') + \gamma(V - V_1)$, where $\gamma$ may be chosen in various ways ($V_i' = V_i + bV_b$). For example, we may desire upper and lower bounds for $V(t)$, in which case we can use upper and lower bounds for $g(V')$. In some cases, $\gamma = g'(V_1')$ or $\gamma = [g(V_1') - g(V_2')]/(V_1 - V_2)$ give such bounds. In all cases, we obtain the characteristic decay of which eq. (3) is one example.

A somewhat more appealing approach that avoids approximating quantities in exponents is the following.[28,29] Again, assume $f(0) \ll f(V')$ to write the integral of (12) in the form

$$t_2 - t_1 = -\int_{V_1}^{V_2} \frac{dV}{f(V')} = -\frac{1}{A}\int_{V_1}^{V_2} \frac{dVe^{-g(V')}g'(V')}{g'(V')}. \tag{14}$$

According to the intermediate-value theorem, this becomes

$$\begin{aligned} t_2 - t_1 &= -\frac{1}{A}\frac{1}{g'(V_3')}\int_{V_1}^{V_2} dVe^{-g(V')}g'(V') \\ &= \frac{1}{A}\frac{1}{g'(V_3')}[e^{-g(V_2')} - e^{-g(V_1')}], \end{aligned} \tag{15}$$

where $V_2 < V_3 < V_1$. (Physically, we must have $g'(V') > 0$.) Thus, given $V_1$ and $t_1$, for each $V_2$ of interest, $t_2$ follows from (15) to within the uncertainty of choosing $g'(V_3')$. (This will normally be, at most, a factor of 2 for $V_2 \approx 1/2V_1$.) The envelope function $V_e(t)$ follows at once from (15) by setting $t_1$ and $\exp[-g(V_1')]$ to zero.

$$t = e^{-g(V_e')}/Ag'(V_3'). \tag{16}$$

If we note that $V_e'(t) = V_e(t) + bV_b$, differentiating (16) with respect to $V_b$ at fixed $t$ and assuming that $g'(V_3')$ is only weakly dependent on $V_b$, it follows at once that

$$\frac{dV_e}{dV_b} = -b.$$

Under these conditions, the vertical separation between envelope functions $V_e(t)$ corresponding to two different applied biases $V_b^1$ and $V_b^2$ (at the same temperature) will be $b(V_b^1 - V_b^2)$, even when these

envelope functions are not straight lines, but are given in general by (16).

Equation (15) may be rewritten in the characteristic form as

$$g(V_2') - g(V_1') = -\log_e [1 + Ae^{g(V_1')}g'(V_3')(t_2 - t_1)], \quad (17)$$

where we may write

$$g(V_2') - g(V_1') \approx g'(V_4')(V_2 - V_1). \quad (18)$$

Equation (18) follows from the mean-value theorem ($V_2 < V_4 < V_1$). Thus, while it is clear that the exact solution will in general not be precisely given by (17) and (18), such a solution will still show the same general features as the above for physically reasonable $g(V)$. It should be noted that, since $g'(V_3')$ and $g'(V_4')$ actually represent averaged quantities, and since $g(V_1')$ is simply an initial condition, the results given in (17) and (18) are rather insensitive to the detailed dependence of $g(V')$ on $V$.

One final point should be noted. For $V$ sufficiently small and $V_b = 0$, we may write

$$\frac{dV}{dt} \approx -f'(0)V. \quad (19)$$

The solution to (19) is the common exponential decay {$\exp(-t/t_o)$, $\tau_o = [f'(0)]^{-1}$} now with voltage $V$ an exponential rather than a logarithmic function of time $t$. For $t \gg \tau_{\text{relax}}$, clearly $V(t)$ will be of this form. But this is well beyond the region in time of the primary decay of the charge, and it is this primary decay governed approximately by (17) to (18), which is of crucial importance for evaluating charge storage devices.

## APPENDIX B

### *B.1 Decay through the insulator*

Section II indicated that the electric field $\mathcal{E}$ in the insulator drives the charge-decay current [see eq. (1)]. This electric field arises from the stored charge and the applied bias. The purpose of this appendix is to express $\mathcal{E}$ in terms of two measurable voltages, the flatband voltage, which measures the stored charge, and the applied-bias voltage.

Let $C_o$ be the capacitance of the oxide in series with the semiconductor, and let $C_i$ be the capacitance of the insulator. We assume $C_o$ (as well as $C_i$) is independent of voltage. (We thus ignore the bias dependence of the semiconductor capacitance.) Simple capacitative

division implies that, if $V_b$ is applied to the gate and there is no stored charge, then the voltage $\bar{V}_b$ dropped across $C_i$ is given by

$$\bar{V}_b = V_b \frac{1/C_i}{1/C_i + 1/C_o}. \tag{20}$$

On the other hand, if a charge $Q$ is stored and $V_b = 0$, then the voltage $\bar{V}_a$ across $C_i$ is given by

$$\bar{V}_a = Q(C_i + C_o)^{-1}. \tag{21}$$

See Fig. 1b. But $Q$ is related to the flatband voltage $V_{fb}$ by the relation

$$Q = C_i V_{fb}. \tag{22}$$

See Fig. 1d. Thus,

$$\bar{V}_a = V_{fb} \frac{1/C_o}{1/C_i + 1/C_o}. \tag{23}$$

The total voltage drop $V_T$ across $C_i$, which produces $\mathcal{E}$, will in general be a superposition of the two; thus,

$$V_T = \bar{V}_a = \bar{V}_b \frac{1/C_o}{1/C_i + 1/C_o} (V_{fb} + bV_b), \tag{24}$$

where

$$b = C_o/C_i. \tag{25}$$

In passing from eq. (1) to eq. (2) in the text, we have let $V_s(T)$ absorb the prefactor (24). Thus, $V_s(T)$ depends both on the physical processes of the decay and on the geometry of the device.

It is important to note that $b$ can be greater than as well as less than unity. In principle, $b$ can be computed from knowledge of $C_o$ and $C_i$. It is more easily obtained from the measured decay curves, a more reliable method. Its slight temperature variation is somewhat of a puzzle. If at higher temperatures additional capacitances $C_i'$ and $C_o'$ arose in series with $C_i$ and $C_o$, respectively, then $b$ would become $b'$ given by

$$b' = \frac{C_o}{C_i}\left(\frac{1 + C_i/C_i'}{1 + C_o/C_o'}\right). \tag{26}$$

One expects the semiconductor portion of the device $C_o$ to be more susceptible to degradation than the insulator side $C_i$. If $C_i' = \infty$ and $C_o' < \infty$, then $b' < b$, as observed. For lower temperatures, $b$ is close to its value predicted by (25). Fortunately, $b$ is determined within experimental error at a given temperature to be independent of $V_b$. Thus, at each temperature we can predict zero-bias behavior. We can

then predict room temperature, zero-bias behavior from higher temperature, and zero-bias behavior *without* knowledge of $b$. This is because only the factor $bV_b$ enters eq. (2), and $V_b = 0$ for zero bias.

### B.2 Decay through the oxide

If the primary discharging current is through the oxide (rather than through the insulator, as assumed throughout this paper), then the electric field that produces this current must be calculated from the voltage across the oxide. Using the notation introduced above, we note that the portion $\bar{V}_b'$ of the bias voltage $V_b$ dropped across the oxide is

$$\bar{V}_b' = V_b \frac{1/C_o}{1/C_i + 1/C_o}. \tag{27}$$

On the other hand, it is clear from Fig. 1b that the contribution $\bar{V}_a'$ of the voltage resulting from the stored charge is just $\bar{V}_a$ given by (23). Thus, the total voltage drop $V_T'$ across $C_o$ is given by

$$V_T' = \bar{V}_a' - \bar{V}_b' = \frac{1/C_o}{1/C_i + 1/C_o}(V - V_b). \tag{28}$$

Thus, unlike the case of $V_T$ (24), no geometrical factor analogous to $b$ enters.

## APPENDIX C

In this appendix, we derive two minor results used in the text.

(*i*) To show that $V_{fb}(t) < V_e(t)$, we need only compare the argument of the $\log_e$ in eq. (3) with $t_1 = 0$ and $V_1 = V_1^o$ with that of eq. (4) when written

$$V_e(t) = V_s \log_e \left[ \left( 0 + \frac{t}{\tau} e^{bV_b/V_s} \right)^{-1} \right]. \tag{29}$$

From (3), we have

$$V_{fb}(t) = V_s \log_e \left[ \left( e^{-V_1/V_s} + \frac{t}{\tau} e^{bV_s/V_s} \right)^{-1} \right]. \tag{30}$$

As the argument of the $\log_e$ in (30) is less than that in (29), it follows at once that

$$V_{fb}(t) < V_e(t). \tag{31}$$

[Additional decay mechanisms important for very short times (<1 minute), which are not included in (30), will, of course, reduce $V_{fb}$

even further.] It may be noted that $V_1$ and $t_1$ may be chosen so that $V_{fb}(t)$ given in (3) exceeds $V_s(t)$ and $V_{fb}$ approaches $V_s$ from above. Such unphysical behavior would result if one mistakenly chose a value of $t_1$ too large; that is, if, after completing the charging of the device ($t = 0$), one recorded $V_1$ at $t_1$ using a clock that read a time sufficiently larger than 0 at $t = 0$.

(*ii*) To show that $V_1 \approx V_1^o$ when $t_1 \ll \tau_{\text{relax}}$, we calculate $V_1$ from eq. (3), in which we set $V_1 = V_1^o$, $t_1 = 0$, and $V_b = 0$. Then it follows that

$$V_1 = V_s \log_e \left[ e^{V_1^o/V_s} \left( 1 + \frac{t_1}{\tau_{\text{relax}}} \right)^{-1} \right]. \tag{32}$$

If now $t_1 \ll \tau_{\text{relax}}$ (the usual case), then

$$V_1 \approx V_1^o - \frac{t_1}{\tau_{\text{relax}}} V_s. \tag{33}$$

Since $V_s < V_1^o$ and $t_1 \approx 10^{-2}\tau_{\text{relax}}$ at most (for room temperature and zero bias $10^{-7}$ is typical), we may use $V_1 \approx V_1^o$ to a very good degree of approximation. The actual, initial, flatband voltage may be larger than $V_1^o$, owing to other decay mechanisms which may be important for $t < 1$ minute. It is $V_1$ [or $V_1^o$ obtained from $V_1$ using eq. (3)], however, with which we are concerned.

## APPENDIX D

### *The "b-Factor"*

The disagreement between the predicted $b$-factor (25) of 2.0 and the much smaller measured $b$-factor (Table I) of 0.8 to 1.2, which we shall call $b'$, calls into question our assumption that the electric field in the insulator is spatially constant. This follows because, if the field is spatially constant (and if the current per carrier depends only on the electric field, as is physically reasonable), then the measured and predicted $b$-factors, $b'$ and $b$, would agree. Since, in fact, they do not agree, we conclude that the field is not spatially constant, a result, perhaps, of charge stored in the insulator. It is the purpose of this appendix to investigate some consequences of charge stored in the insulator (in addition to charge stored at the oxide-insulator interface) on the decay of the flatband voltage $V_{fb}$.

Let us begin by asking what information our flatband-voltage measurements tell us in light of the possibility of having charge storage in the insulator. Referring to our empirical result (2), we note that, at a fixed temperature, the rate of decay of $V_{fb}(t)$ is a function of

$[V_{fb}(t) + b'V_b]$. Thus, whatever electric field governs the decay of $V_{fb}(t)$, it too must depend on $V_{fb}$ and $V_b$ in the form $[V_{fb}(t) + b'V_b]$.

The next step is to express the various fields and voltages of interest in terms of the stored charge. Let $Q_s(t)$ be the charge stored at the oxide-insulator interface at time $t$, $\rho_i(x, t)$ the density of charge stored in the insulator, $\epsilon_i$ the dielectric constant of the insulator, and $\epsilon_o$ that of the oxide. It follows (see Fig. 1d) that the flatband voltage is given by

$$V_{fb}(t) = \frac{Q_s(t)}{C_i} + \int_0^{d_i} \frac{\rho_i(x, t)}{\epsilon_i} (d_i - x)dx \tag{34}$$

$$\equiv V_s(t) + V_m(t) \tag{35}$$

$$= \int_0^{d_i} \frac{\rho_T(x, t)}{\epsilon_i} (d_i - x)dx \tag{36}$$

$$\equiv V_T(t), \tag{37}$$

where

$$\rho_T = \rho_i + (Q_s/A)\delta(x). \tag{38}$$

Here $A$ is the cross-sectional area of the device, $d_i$ is the thickness of the insulator, $\delta(x)$ is the usual delta function, $V_s$ is the contribution of $Q_s$ to $V_{fb}$, and $V_m$ (the $m$ referring to moment of charge) is the contribution of the charge stored in the insulator to $V_{fb}$. (Throughout our discussion, we are assuming that *no* charge is stored in the oxide.) A straightforward but more involved calculation leads to the electric field in the insulator, $0 < x < d_i$, under general charging and applied-bias conditions:

$$E(x, t) = (1 + b)^{-1}\left[ \frac{Q_s(t)/d_i}{C_i} + b\frac{V_b}{d_i} + \int_0^x dx' \frac{\rho_i(x', t)}{\epsilon_i}\left(1 + b\frac{x'}{d_i}\right) - \int_x^{d_i} dx' \frac{\rho_i(x', t)}{\epsilon_i} b\left(1 - \frac{x'}{d_i}\right)\right] \tag{39}$$

$$= (1 + b)^{-1}\left[ b\frac{V_b}{d_i} + \int_0^x dx' \frac{\rho_T(x', t)}{\epsilon_i}\left(1 + b\frac{x'}{d_i}\right) - \int_x^{d_i} dx' \frac{\rho_T(x', t)}{\epsilon_i} b\left(1 - \frac{x'}{d_i}\right)\right]. \tag{40}$$

If we integrate (39) or (40) over the insulator, we obtain the average field times $d_i$, the voltage drop across the insulator $V_d(t)$ given by

$$V_d(t) = (1 + b)^{-1}[V_s(t) + V_m(t) + bV_b] = (1 + b)^{-1}[V_{fb}(t) + bV_b]. \tag{41}$$

Using these results, we can discuss what physical effects can and cannot lead to the observed behavior ($b' < b$).

Owing to the presence of charge in the insulator, the electric field at the oxide-insulator interface will be reduced from its value in the absence of such charge. If $E(0^+, t)$ controls the decay, perhaps this reduction can lead to a reduced $b$-factor. Unfortunately, this effect would result in ($b' > b$) rather than ($b' < b$). To see this, we note from (39) that

$$\begin{aligned} E(0^+, t) &= (1 + b)^{-1} d_i^{-1} [V_s(t) + bV_b - bV_m(t)] \\ &= \frac{b/b'}{(1 + b)d_i} \left[ \frac{b'}{b} V_s(t) - b'V_m(t) + b'V_b \right]. \end{aligned} \tag{42}$$

If this expression is to be a function of $[V_{fb}(t) + b'V_b]$, it is necessary that

$$V_{fb}(t) = (b'/b)V_s(t) - b'V_m(t). \tag{43}$$

To satisfy both (35) and (43), it is necessary to have

$$V_m(t) = \frac{1/b - 1/b'}{1/b + 1} V_{fb}(t) \tag{44}$$

and

$$V_s(t) = \frac{1 + 1/b'}{1/b + 1} V_{fb}(t). \tag{45}$$

However, if $b' < b$, then $1/b < 1/b'$, and the coefficient of $V_{fb}(t)$ in (44) is negative. This is rather disturbing since, as a result of the initial charging and subsequent decay of the charge through the insulator, one would expect $V_m(t)$, as defined in (35), to be positive. In addition, (44) implies that, as $V_{fb}$ decreases during charge decay, $V_m$ must increase algebraically. This can occur only if charge stored at the oxide-insulator interface becomes trapped in the insulator neutralizing the charge trapped there. This effect would be enhanced if more charge were neutralized near the interface ($x = 0$) than in the middle of the insulator. Although this may be the source of ($b' < b$), the origin seems physically unreasonable of the bulk trapped charge of sign opposite the stored charge which, although not neutralized by the charging current, is neutralized by the decay current. We conclude that, if ($b' < b$), then the field at the interface probably does *not* control the charge decay.

We note in passing that, since the voltage drop $V_d(t)$ across the insulator (41) is already a function of $[V_{fb}(t) + bV_b]$, it also cannot

be the average field in the insulator which is controlling the decay. This hardly requires further elaboration.

One effect that charge stored in the insulator has on the decay is to enhance the electric field near gate ($x = d_i$) relative to the field near the interface ($x = 0$). This enhancement may be sufficient to remove charge stored in the region $x_o < x < d_i$ of the insulator during the initial portion of the decay, leaving significant stored charge only in the region $0 < x < x_o$ of the insulator and at the interface. If, during the remainder of the decay, the decay rate is governed by the electric field at $x_o$, we then expect to observe ($b' < b$). Let us see how this comes about.

If $\rho_i(x, t) = 0$ for $x > x_o$, then the electric field at $x = x_o$ is, according to (39), given by

$$E(x_o, t) = \frac{b/b'}{(1 + b)d_i} \left\{ \frac{b'}{b} V_s(t) + b'V_b + \frac{b'}{b}\left[ (b + 1) \frac{Q_i(t)}{c_i} - bV_m(t) \right] \right\}, \quad (46)$$

where

$$Q_i(t) \equiv A \int_0^{d_i} dx' \rho_i(x', t) \quad (47)$$

$$= A \int_0^{x_o} dx' \rho_i(x', t), \quad (47a)$$

the size of the stored charge in the insulator. Alternatively, using (40) for the field, we obtain

$$E(x_o, t) = \frac{b/b'}{(1 + b)d_i} \left[ -b'V_T(t) + \frac{b'}{b} (1 + b) \frac{Q(t)}{c_i} + b'V_b \right], \quad (48)$$

where $Q(t)$ is the total stored charge defined by

$$Q(t) \equiv Q_s(t) + Q_i(t). \quad (49)$$

For $E(x_o, t)$ to be a function of $(V_{fb} + b'V_b)$, it is necessary that [using (37)]

$$V_T = -b'V_T + \frac{b'}{b} (1 + b) \frac{Q(t)}{C_i} \quad (50)$$

or that

$$V_T(t) = \frac{1 + 1/b}{1 + 1/b'} \frac{Q(t)}{C_i}. \quad (51)$$

Now ($b' < b$) implies that $C_i V_T(t) < Q(t)$, which means in turn that some charge is stored in the insulator, a consistent result.

Having seen how a measured $b$-factor ($b'$) can arise which is less than the computed $b$-factor ($b$), we must inquire as to whether the effect is sufficient to explain the measured results. For simplicity, let us assume that $\rho_i$ is uniform for $0 < x < x_o$. Then it follows that

$$V_T = \frac{Q_s}{C_i} + \frac{Q_t}{C_i}(1 - x_o/2d_i). \tag{52}$$

It then follows from (48), (51), and (52) that

$$\frac{Q/C_i}{V_T} = \frac{1 + Q_t/Q_s}{1 + (Q_t/Q_s)(1 - x_o/2d_i)} = \frac{1 + 1/b'}{1 + 1/b}. \tag{53}$$

If we consider the case at hand where $b = 2.0$ and $b'$ is essentially 1, then (53) implies that $x_o$ must be at least $0.50 \times d_i$, that is, that the stored charge in the insulator must extend at least 50 percent of the distance from the interface to the gate. If $Q_t/Q_s$ is 1.0, then $x_o = d_i$. Thus, $Q_t/Q_s$ must be at least 1.0 for, if it were smaller, then $x_o > d_i$, which is not possible. Therefore, if the charge stored in the insulator is uniform between $x = 0$ and $x = x_o$, then it is necessary to understand the observed $b'$ that $0.50 < x_o/d_i < 1$ and $1.0 < Q_t/Q_s$, the latter implying that more charge must be stored in the bulk at the insulator than at the interface.

In the preceding paragraph, we have assumed that the stored charge was uniformly distributed in the region $0 < x < x_o$. In fact, we expect the charge to be more dense near the interface ($x = 0$) where the field is lowest than near $x = x_o$. To satisfy (51), this would require larger values of $Q_t/Q_s$ and of $x_o$ than for the uniform case. We have also assumed that it is the electric field at $x_o$ that controls the decay. It is possible that $E(x, t)$ for $x < x_o$ in fact performs this function. This would further increase the values of $Q_t/Q_s$ and $x_o$ required to achieve ($b' < b$). One's lattitude here is rather limited, however, for, as we have seen, if it is $E(0^+, t)$ which controls the decay, then $Q_t/Q_s$ becomes negative. We offer the above, therefore, as possibilities only.

Another source of the ($b' < b$) effect is that the charge may be extraction-limited, that is, controlled by the field at $x = d_i$. If we put $x_o = d_i$ in eqs. (46) and (47a), then we again obtain (51) relating the total stored charge to its first moment. We noted above that we could understand the measured $b$-factor for $x_o = d_i$ and a uniform, stored, insulator charge if $Q_t/Q_s = 1.0$, a reasonable but perhaps somewhat

large value. However, if the current is extraction-limited, then we expect that the charge would be more dense near $x = d_i$ than in the bulk of the insulator. As a hypothetical example, suppose that the stored, insulator charge is uniform for $x_1 < x < d_i$ and zero for $0 < x < x_1$. Then from (37) we obtain

$$V_T = \frac{Q_s}{C_i} + \frac{Q_t}{C_i}\frac{1}{2}\left(1 - \frac{x_1}{d_i}\right). \tag{54}$$

It then follows from (49) and (51) that

$$\frac{Q/C_i}{V_T} = \frac{1 + Q_t/Q_s}{1 + (Q_t/Q_s)(1 - x_1/d_i)/2} = \frac{1 + 1/b'}{1 + 1/b}. \tag{55}$$

This relation provides a much greater possibility for obtaining ($b' < b$) than (53). For example, if $Q_t \gg Q_s$, then

$$b' \leqq (1 + 2/b)^{-1} < b/2. \tag{56}$$

Or, if $x_1 \approx d_i$, then

$$b' = b[(1 + b)Q_t/Q_s + 1]^{-1}, \tag{57}$$

from which $b' = 1$ would follow for $b = 2.0$ if $Q_t/Q_s = 0.333$, a very reasonable value. For $x_1 < d_i$, somewhat larger $Q_t/Q_s$ would be required to satisfy (55). However, in most cases $Q_t/Q_s$ would be smaller than that in the previous example (53), in which the stored charge was assumed near the oxide-insulator interface. We conclude that, while we have indicated the possibility of how ($b' < b$) can come about, further work is required to really pin down and calculate the measured $b$-factor $b'$.

## REFERENCES


1. F. A. Sewell, H. A. R. Wegener, and E. T. Lewis, "Charge Storage Model for Variable Threshold FET Memory Element," Appl. Phys. Lett. *14*, 1969, pp. 45–47.
2. J. T. Wallmark and J. H. Scott, "Switching and Storage Characteristics of MIS Memory Transistors," RCA Review, *30*, 1969, pp. 335–365.
3. E. C. Ross and J. T. Wallmark, "Theory of the Switching Behavior of MIS Memory Transistors," RCA Review, *30*, 1969, pp. 366–381.
4. D. Frohman-Bentchkowsky and M. Lenzlinger, "Charge Transport and Storage in Metal-Nitride-Oxide-Silicon (MNOS) Structures," J. Appl. Phys., *40*, 1969, pp. 3307–3319.
5. E. C. Ross, A. M. Goodman, and M. T. Duffy, "Operational Dependence of the Direct-Tunneling Mode MNOS Memory Transistor on the $SiO_2$ Layer Thickness," RCA Review, *31*, 1970, pp. 467–478.
6. C. M. Svensson and K. I. Lundström, "Theory of the Thin-Oxide MNOS Memory Transistor," Electronics Letters, *6*, 1970, pp. 645–647.
7. K. I. Lundström and C. M. Svensson, "Properties of MNOS Structures," IEEE Transactions on Electron Devices, *ED-19*, 1972, pp. 826–836.

8. M. A. C. S. Brown, J. E. Bounden, and G. F. Vanstone, "A Simple Equivalent Circuit for the MNOST Memory Element," Solid State Electronics, *15*, 1972, pp. 707–719.
9. M. H. White and J. R. Cricchi, "Characterization of Thin-Oxide MNOS Memory Transistors," IEEE Trans. on Electron Devices, *ED-19*, 1972, pp. 1280–1288.
10. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," B.S.T.J., *46*, No. 6 (July-August 1967), pp. 1288–1295.
11. D. Kahng, "Semipermanent Memory Using Capacitor Charge Storage and IGFET Read-out," B.S.T.J., *46*, No. 6 (July-August 1967), pp. 1296–1300.
12. D. Kahng and E. H. Nicollian, "Physics of Multilayer-Gate IGFET Memories," Applied Solid State Science, *3*, 1972, pp. 1–70.
13. D. Kahng, W. J. Sundburg, D. M. Boulin, and J. R. Ligenza, "Interfacial Dopants for Dual-Dielectric, Charge-Storage Cells," B.S.T.J., this issue, pp. 1723–1739.
14. H. H. Poole, Lond. Edin. Publ. Phil. Mag. *33*, 1916, p. 112; *34*, 1917, p. 195.
15. J. Frenkel, "On Pre-Breakdown Phenomena in Insulators and Electronic Semi-Conductors," Phys. Rev., *54*, 1938, pp. 647–648.
16. J. Frenkel, "On the Theory of Electric Breakdown of Dielectrics and Electronic Semiconductors," Tech. Phys. USSR, *5*, 1938, pp. 685–695.
17. R. M. Hill, "Poole-Frenkel Conduction in Amorphous Solids," Phil. Mag., *23*, 1971, pp. 59–86.
18. J. G. Simmons, "Conduction in Thin Dielectric Films," J. Phys., *D4*, 1971, pp. 613–657.
19. G. A. Brown, W. C. Robinette, and M. G. Carlson, "Electrical Characteristics of Silicon Nitride Films Prepared by Silane-Ammonia Reaction," J. Electrochem. Soc., *115*, 1968, pp. 948–955.
20. G. A. N. Connell, D. L. Camphausen, and W. Paul, "Theory of Poole-Frenkel Conduction in Low-Mobility Semiconductors," Phil. Mag., *26*, 1972, pp. 541–551.
21. K. K. Thornber and C. A. Mead, "Electronic Processes in $\alpha$-Sulfur," J. Phys. Chem. Solids, *26*, 1965, pp. 1489–1495.
22. K. K. Thornber and R. P. Feynman, "Velocity Acquired by an Electron in a Finite Electric Field in a Polar Crystal," Phys. Rev., *B1*, 1970, pp. 4099–4114; Errata, *4*, 1971, p. 674.
23. K. K. Thornber, "Linear and Nonlinear Electronic Transport in Electron-Phonon Systems: Self-Consistent Approach within the Path Integral Formalism," Phys. Rev. *B3*, 1971, pp. 1929–1941; Errata, *4*, 1971, p. 675.
24. J. P. Sitarik, "DC Conduction in Anodic Oxides," unpublished work.
25. R. M. Hill, "Poole-Frenkel Conduction in Amorphous Solids," Phil. Mag, *23*, 1971, pp. 59–86.
26. H. Adachi, Y. Shibata, and S. Dno, "Electronic Conduction through Evaporated Silicon Oxide Films," J. Phys., *D4*, 1971, pp. 988–994.
27. R. B. Hall, "The Poole-Frenkel Effect," Thin Solid Films, *8*, 1971, pp. 263–271.
28. R. H. Walden, "A Method for the Determination of High-Field Conduction Laws in Insulating Films in the Presence of Charge Trapping," J. Appl. Phys., *43*, 1972, pp. 1178–1186.
29. A. V. Ferris-Prabhu, "Charge Transfer in Layered Insulators," Solid-State Electronics, *16*, 1973, pp. 1086–1087.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 9, November 1974
*Printed in U.S.A.*

# Input Amplifiers for Optical PCM Receivers

By J. E. GOELL

(Manuscript received March 1, 1974)

*This paper describes the noise performance of input amplifiers for optical pulse-code-modulation repeaters. The noise is treated in terms of an effective noise generator in parallel with the photocurrent induced in the detector and the effective noise, in turn, is related to error performance. The analysis applies to both conventional and integrating front ends. Both field effect and bipolar transistor amplifiers are treated. For the latter, an optimum bias current that minimizes the effect of thermal noise is derived. Finally, predicted and measured performance are compared for silicon field-effect transistor input amplifiers at 6.3 Mb/s and 50 Mb/s, and for bipolar transistor and GaAs field-effect transistor input amplifiers at 274 Mb/s.*

## I. INTRODUCTION

The factors which limit the performance of an optical receiver are optical quantum noise, leakage noise of the detector, thermal noise introduced by the detector load resistor, and various forms of noise introduced by the input amplifier. If an avalanche detector is used, the leakage noise has two components—one which is gain independent and the other which is gain dependent. In addition, the gain process introduces a signal-dependent noise. In high-speed pulse-code-modulation (PCM) receivers, input-amplifier noise plays an important role in the determination of system performance. If leakage current is negligible, it can be shown[1,2] that without avalanche gain the required signal power to achieve a given error probability varies as the square root of the thermal noise power and with optimum avalanche gain with the sixth root of the thermal noise power. Furthermore, the optimal avalanche gain varies as the cube root of the thermal noise power.*

* These results assume an excess noise coefficient of 0.5.

In applications employing a detector with a capacitive impedance such as nuclear particle counters[3] and television camera input amplifiers,[4] an approach has been employed in which the input circuit integrates the signal and the signal is equalized after amplification. Personick[2] has analyzed the performance of PCM repeaters with integrating front ends and Goell has experimentally verified some of his results at 6.3 Mb/s.[5]

The noise performance of an amplifier is often stated in terms of noise figure. This approach is attractive when thermal noise can be reduced by matching the source impedance to the optimum noise impedance of the amplifier. For optical receivers and other systems with a capacitive source the noise of the source is often a negligible portion of the total thermal noise, and the noise contributed by transformer losses would more than negate the advantage gained by transformation. For this case, noise figure, which is inversely proportional to source noise power, is a poor figure of merit and it is better to treat the noise directly.

The approach chosen here is to describe the noise of the amplifier by an equivalent input noise current generator which produces the same noise at the output as the internal sources of the amplifier. The intermediate step of finding a short-circuit input noise current generator and open-circuit input noise voltage generator is dispensed with here because it does not contribute to physical understanding when the noise figure concept is not employed.

If the linear portion of the receiver is modeled by a cascade of an amplifier, an equalizer to give a flat frequency response in the band of interest, and a filter to set the noise bandwidth and control the pulse response, then the equivalent current is a particularly convenient way to express the noise. This approach is conceptually simple, the effect of equalization is implicitly included, and filtering has the same effect on the equivalent noise as on the signal.

The error performance of a PCM receiver can be related to the ratio of the peak signal to rms noise ratio at the regenerator. It will be shown that this ratio is equal to the ratio of the average received signal power to an effective noise current which can be readily derived from the equivalent input noise current.

In the previously mentioned work by Personick the amplifier noise was modeled by the series voltage and shunt current noise generator. Results were given only for field effect transistor (FET) input amplifiers. In this paper the noise of both field effect and bipolar transistor amplifiers is analyzed. The effect of spreading resistance and load

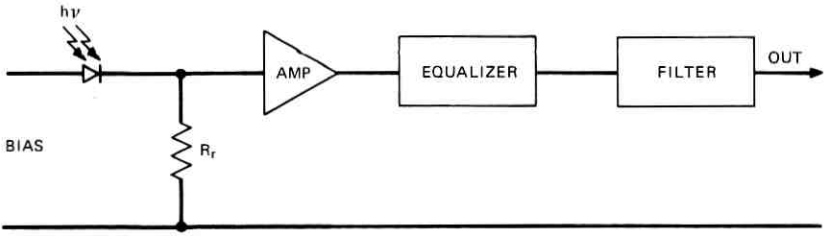



Fig. 1—Typical input circuit of an optical repeater.

noise is included and the application of the method to cases in which the noise of more than one stage is significant is described. The approach is used to compare the performance of silicon bipolar and field effect transistors and GaAs field effect transistors as a function of bit rate. Theoretical predictions are compared with measured results for silicon FET input amplifiers at 6.3 Mb/s and 50 Mb/s[6] and for bipolar transistor and GaAs FET input amplifiers at 274 Mb/s.

## II. CALCULATION OF EQUIVALENT NOISE

A model of the initial stages of a typical optical receiver is shown in Fig. 1. The resistor $R_r$ is provided to return the detector bias current to its source. Equalization is provided to compensate for the frequency dependence of the amplifier gain, and filtering is provided to limit the noise bandwidth and shape the signal. For the cases to be described here, it is convenient to think in terms of an equalizer to give a flat frequency response followed by a filter which describes the frequency dependence of the transfer characteristic of the receiver, although in practice they can be combined. In some applications, such as where matched filtering is to be employed, another approach might be used.

The detector circuit without avalanche gain can be modeled as shown in Fig. 2. Here $i_p$ is the induced photocurrent, $C_j$ the detector

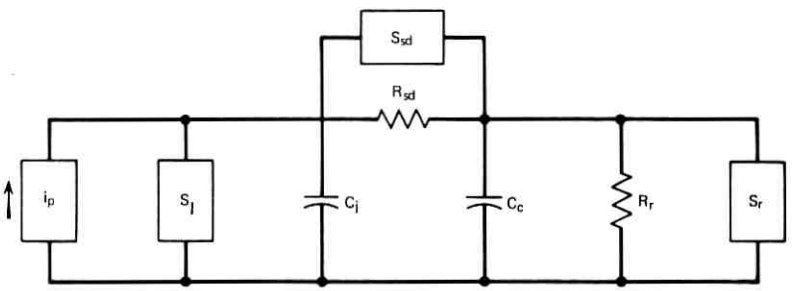



Fig. 2—Detector circuit.

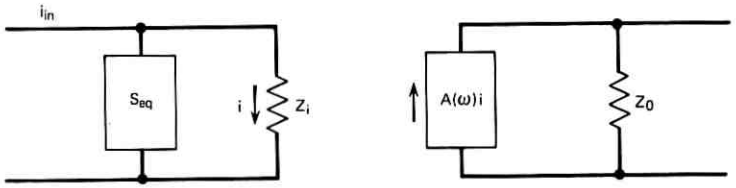



Fig. 3—Equivalent noise circuit.

junction capacitance, $R_{sd}$ the diode spreading resistance and $S_{sd}$ the noise current spectral density associated with it, $C_c$ the capacitance of the interconnection circuit, $S_r$ the noise current spectral density of the dc return resistor, and $S_l$ the spectral density of the gain-independent diode-leakage noise current.* With avalanche gain, $i_p$ is the photocurrent multiplied by the mean gain, $\bar{g}$.

The performance of any linear amplifier can be described by the equivalent circuit of Fig. 3. Here, the impedance $Z_i$ is the parallel combination of the input impedance of the circuit and the output impedance of the preceding circuit, and $A(\omega)$ is the current amplification of the stage. $S_{eq}$ is the frequency-dependent equivalent noise current spectral density of the noise current generator which when applied to the input gives the same output noise spectral density as the internal noise generators.

An equivalent circuit which applies to both bipolar and field-effect transistors is shown in Fig. 4. Spreading resistance has been ignored for the present, but will be analyzed later as a separate stage.

The equivalent noise current spectral density is given by

$$S_{eq} = S_{eq1} + S_{eq2}, \tag{1}$$

where $S_{eq1}$ is the contribution to $S_{eq}$ of $S_1$, the emitter (source) noise current spectral density, and $S_{eq2}$ is the contribution to $S_{eq}$ of $S_2$, the collector (drain) noise current spectral density. The calculation of $S_{eq1}$ and $S_{eq2}$ for common emitter (source), collector (drain), and base (gate) stages is described in the appendix. A comparison of amplifier configuration is also given.

At low frequencies, that is, when the transit time of carriers through the device is short compared to the period of the signal, for a field-effect transistor at pinch-off

$$S_1 = 4kT\theta(\omega c_g)^2/g_m + 2eI_g \tag{2}$$

* The effect of the gain-dependent leakage noise has been described by Personick.[2]

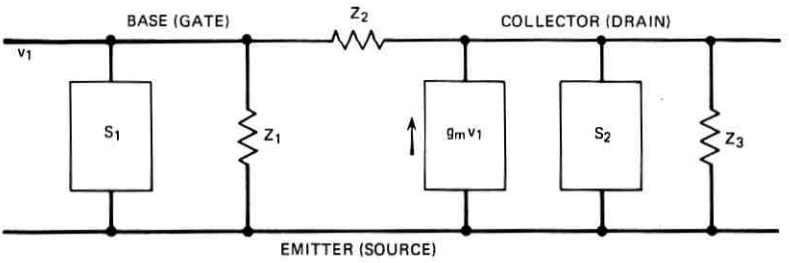



Fig. 4—Equivalent circuit for both bipolar and field effect transistors.

and

$$S_2 = 4kT\Gamma g_m, \tag{3}$$

where $T$ is the temperature, $g_m$ is the transconductance, $I_g$ is the gate leakage current, $c_g$ is the gate capacitance, and k is Boltzmann's constant.[7] For junction field-effect transistors, $\Gamma = \frac{2}{3}$ and $\theta$ varies from $\frac{1}{10}$ to $\frac{4}{15}$.[8] With currently available devices, drain noise has been found to predominate with an untuned input circuit.

At very low frequencies account must also be taken at $1/f$ noise. The $1/f$ noise region for junction field-effect transistors extends to a few hundred kilohertz, while for metal-oxide semiconductor field effect transistors it can extend to above 10 MHz. For GaAsFETs intervalley scattering adds an additional component of noise. For this case $\Gamma = 1.1$.[9]

For bipolar transistors with short base transit time, the input and output noise generator spectral densities are given by

$$S_1 = 2eI_b, \tag{4}$$

$$S_2 = 2eI_c, \tag{5}$$

where $I_b$ is the base current and $I_c$ is the collector current.[7] The above applies for the case of short base transit time and the base current much greater than the reverse base saturation current.

### III. EFFECTIVE NOISE IN PCM RECEIVERS

For PCM systems, performance is measured in terms of required power to achieve a specified error probability and the error probability is determined from the ratio of the peak signal to the rms noise at the regenerator input. In this section the relationship between the peak-signal-to-rms-noise ratio, the average received signal, and the equivalent noise will be described.

We shall assume that the detected photocurrent, an undistorted version of the transmitted signal, is given by

$$\sum_{k=-\infty}^{\infty} T_b b_k I_{av} h_p(t - kT_b),$$

where

$$\int_{-\infty}^{\infty} h_p(t - kT_b)dt = 1,$$

$b_k = 0, 1$ depending on the signal statistics, $T_b$ is the bit interval, and $I_{av}T_b$ is the detected charge if a single pulse were received. It is assumed that the transmitted pulses are distinct.

The output signal is given by

$$\sum_{k=-\infty}^{\infty} b_k h_m h_o(t - kT_b),$$

where $h_o(t)$ has been normalized to have unit peak amplitude and $h_m$ is the peak output signal current. It will be assumed that intersymbol interference is negligible, that is, that at the sampling instant, $t_s$,

$$h_o(t_s - kT_b) = 0$$
$$k = \pm 1, \pm 2, \cdots.$$

An inconsequential time displacement has been ignored. Then the transfer function is given by

$$A(\omega) = \frac{h_m H_o(\omega)}{T_b I_{av} H_p(\omega)}, \tag{6}$$

where $H_p(\omega)$ and $H_o(\omega)$ are the Fourier transforms of $h_p(t)$ and $h_o(t)$, respectively. The mean-square noise after amplification and filtering is given by

$$\overline{n^2} = \frac{1}{2\pi} \int_0^{\infty} |A(\omega)|^2 S_{eq}(\omega)d\omega.$$

Substituting the mean-square noise into eq. (6) gives the peak-signal-to-rms-noise ratio

$$\frac{h_m}{\sqrt{\overline{n^2}}} = \frac{I_{av}}{\sqrt{\overline{i_{eff}^2}}}, \tag{7}$$

where

$$\overline{i_{eff}^2} = \frac{1}{2\pi T_b^2} \int_0^{\infty} \left|\frac{H_o(\omega)}{H_p(\omega)}\right|^2 S_{eq}d\omega. \tag{8}$$

Thus the quantity $\sqrt{\overline{i_{\text{eff}}^2}}$ is an effective noise current which relates the thermal noise to the average induced photocurrent.

For the case where most of the noise originates in the first amplifier stage we will see that under conditions which apply in several cases of interest the equivalent noise can be expanded in the Taylor series

$$S_{\text{eq}}(\omega) = \sum_j \alpha_j(\omega)\mathcal{g}_j.$$

It is then convenient to put the effective noise in the form

$$\overline{i_{\text{eff}}^2} = f_b \sum_j \alpha_j (2\pi f_b)^j \mathcal{g}_j, \tag{9}$$

where

$$\mathcal{g}_j = \frac{f_b^{1-j}}{(2\pi)^{j+1}} \int_0^\infty \left|\frac{H_o(\omega)}{H_p(\omega)}\right|^2 \omega^j d\omega \tag{10}$$

and $f_b = 1/T_b$. The normalization ratio was chosen so that the $\mathcal{g}$'s depend only on the shape of the pulse relative to the bit interval, not on the bit interval.

It can be shown[2,5] that for binary PCM with an avalanche detector the power required to achieve a specified error probability, $P_e$, is given by

$$p = \frac{h\nu Q}{2\eta}\left[\frac{Q\overline{g^2}f_b}{\bar{g}^2} + \frac{2}{\bar{g}e}\sqrt{\overline{i_{\text{eff}}^2}}\right], \tag{11}$$

where

$\eta$ = detector quantum efficiency,
$e$ = electronic charge,
$\nu$ = optical frequency,
$\beta$ = detector quantum efficiency,
$\bar{g}$ = mean detector gain,
$\overline{g^2}$ = mean-squared detector gain,

and

$$G = \sqrt{2}\ \text{erfc}^{-1}(2P_e).$$

The function erfc is the error function complement. For a gainless detector when thermal noise predominates, the first term in the bracket of eq. (11) can be neglected.

A value of $\bar{g}$ exists[1] which minimizes $p$. The ratio of $\overline{g^2}/\bar{g}^2$ can be approximated* by $\bar{g}^x$. Then the value of $\bar{g}$ which minimizes $p$ is given by

$$g_{\text{opt}} = \left( \frac{2\sqrt{\overline{i_{\text{eff}}^2}}}{xeQf_b} \right)^{1/(x+1)} \tag{12}$$

and the required power at optimum gain is given by

$$P_{\text{opt}} = \frac{h\nu Q^{(x+2)/(x+1)}}{\eta e} \left( \frac{xef_b}{2} \right)^{1/(x+1)} \left( 1 + \frac{1}{x} \right) \overline{i_{\text{eff}}^2}^{x/(2x+2)}. \tag{13}$$

## IV. APPLICATIONS

In this section, the effective noise for both field effect and bipolar transistors will be described. Emphasis will be placed on conditions encountered in receivers operating in the 1–Mb/s to 1–Gb/s range. It will be assumed that most of the noise originates in or before the first amplifier stage. Corrections for the load, subsequent stage, and spreading resistance noises will be included.

Both detectors and transistors have distributed junction capacitance and series resistance followed by a case capacitance. However, at the frequencies to be covered here the equivalent circuit can be approximated by capacitors across the detector source and transistor base (gate) separated by the total spreading resistance of both devices. The spreading resistance can be considered as a separate stage with a transfer function $A(\omega)$ given by

$$A(\omega) = \frac{1}{1 + j\omega C_d R_s}$$

and noise current spectral density

$$S_s = 4kT\omega^2 C_d^2 R_s,$$

where $C_d$ is capacitance across the detector source.

The gain-independent leakage current $I_d$ can be accounted for by a noise spectral density

$$S_{eqd} = 2eI_d$$

in parallel with the detector current generator. Noise of spectral density $S_r$ is contributed by the return resistor, $R_r$, which will be assumed to be on the amplifier side of $R_s$. This assumption is valid as long as the contribution of $R_s$ to the noise is small.

---

* Silicon detectors with $x$ between 0.3 and 0.5 are available today commercially.

The total equivalent noise current spectral density can be written as

$$S_{eq} = S_s + S_{eq\,d} + \frac{S_{eq1} + S_r + S_{eq2}}{|A(\omega)|^2}$$

$$= 4kT\omega^2 C_d^2 R_s + 2eI_d + \frac{4kT}{|A(\omega)|^2}\left[\frac{\theta(\omega C_g)^2}{g_m} + \frac{1}{R_r} + \frac{\Gamma g_m + \frac{1}{R_L}}{|g_m Z_g|^2}\right],$$

where

$$\frac{1}{Z_g} = \frac{1}{R_r} + j\omega\left(C_g + \frac{C_d}{1 + j\omega C_d R_s}\right),$$

$$C_g = C_{gs} + C_{gd},$$

and $R_L$ is the amplifier load resistance. Thus,

$$S_{eq} \approx 4kT\left[\omega^2 C_d^2 R_s + \frac{1 + \omega^2 C_d^2 R_s^2}{R_r} + \frac{\theta(\omega C_g)^2}{g_m}(1 + \omega^2 C_d^2 R_s^2)\right.$$

$$+ \frac{\Gamma g_m + \frac{1}{R_L}}{g_m^2}\left\{\left(\frac{1}{R_r} - \omega^2 C_g C_d R_s\right)^2\right.$$

$$\left.\left. + \omega^2\left(C_t + C_d\frac{R_s}{R_r}\right)^2\right\}\right] + 2eI_d, \quad (14)$$

where

$$C_t = C_d + C_{gs} + C_{gd}.$$

In practice $R_r$ can be made extremely large so

$$\frac{R_s}{R_r} \ll 1$$

and, from eq. (9),

$$\overline{i_{eff}^2} = 2eI_d f_b g_o + 4kTf_b\left[\left\{\frac{1}{R_r} + \frac{\Gamma g_m + \frac{1}{R_L}}{g_m^2 R_r^2}\right\} g_o\right.$$

$$+ (2\pi f_b)^2\left\{g_g + \frac{\Gamma g_d + \frac{1}{R_L}}{g_m^2} C_t^2\right\} g_2$$

$$\left. + (2\pi f_b)^4\left\{\frac{\Gamma g_d + \frac{1}{R_L}}{g_m^2} C_g^2 + g_g C_d^2 R_s^2\right\} g_4\right]. \quad (15)$$

From this relation, it is clear that, from the standpoint of thermal noise, $R_r$ should be as large as possible, even if signal integration takes place.

For junction field-effect transistors such as the 2N3823, the gate noise is negligible. Assuming $R_r$ is large and the detector leakage current is negligible gives the simple relation

$$\overline{i_{\text{eff}}^2} = \frac{8\text{k}Tf_b(2\pi f_b C_t)^2 g_2}{3g_m}. \tag{16}$$

Thus, for this case which is typical for junction FETs operating at a bit rate below 25 Mb/s, the quantity

$$\frac{\sqrt{g_m}}{C_t}$$

is a figure of merit for the required power without avalanche gain. Under the above assumption, the effective noise current increases with the $\frac{3}{2}$ power of the bit rate.

We now turn our attention to the bipolar transistor. Unlike the FET, for a bipolar transistor the optimum bias is dependent on frequency. For a bipolar transistor,

$$S_{\text{eq1}} = 2eI_b,$$

$$S_{\text{eq2}} = \frac{2eI_C + \dfrac{4\text{k}T}{R_L}}{|g_m Z_b|^2},$$

where $Z_b$, the impedance across the base, is given by

$$\frac{1}{Z_b} = \frac{1}{R_b} + j\omega\left(C_b + \frac{C_d}{1 + j\omega C_d R_s}\right)$$

and

$$R_b = \frac{kT}{eI_b}.$$

Assuming that the dc current gain equals the ac current gain, the equivalent noise current spectral density is

$$\begin{aligned} S_{\text{eq}} &= S_s + S_d + \frac{S_{\text{eq1}} + S_{\text{eq2}}}{|A(\omega)|^2} \\ &= 2eI_b\left(1 + \frac{1}{\beta}\right) + \frac{4\text{k}T}{\beta^2 R_L} + 2eI_d + \omega^2\left[4\text{k}TC_d^2R_s + 2eI_bC_d^2R_s^2\right. \\ &\quad \left. + \left(\frac{2eI_b}{\beta} + \frac{4\text{k}T}{\beta^2 R_L}\right)\left\{C_t^2\left(\frac{\text{k}T}{eI_b}\right)^2 + C_d^2R_s^2 + 2C_d^2R_s\frac{\text{k}T}{eI_b}\right\}\right] \\ &\quad + \omega^4\left(\frac{2eI_b}{\beta} + \frac{4\text{k}T}{\beta^2 R_L}\right)C_b^2C_d^2R_s^2\left(\frac{\text{k}T}{eI_b}\right)^2. \end{aligned}$$

Then

$$\overline{i_{\text{eff}}^2} = 2f_b e\left[\left\{\left(1+\frac{1}{\beta}\right)I_b + \frac{2\text{k}T}{e\beta^2 R_L} + I_d\right\}\mathcal{g}_o + (2\pi f_b)^2\left\{\frac{2\text{k}TC_d^2R_s}{e} + C_d^2R_s^2I_b + \left(\frac{I_b}{\beta}+\frac{2\text{k}T}{\beta^2 eR_L}\right)\left[C_t^2\left(\frac{\text{k}T}{eI_b}\right)^2 + C_d^2R_s^2 + 2C_d^2R_s\left(\frac{\text{k}T}{eI_b}\right)\right]\right\}\mathcal{g}_2 + (2\pi f_b)^4\left\{\left(\frac{I_b}{\beta}+\frac{2\text{k}T}{\beta^2 eR_L}\right)C_d^2C_b^2R_s^2\left(\frac{\text{k}T}{eI_b}\right)^2\right\}\mathcal{g}_4\right].$$

At values of base bias which will shortly be shown to lead to minimum effective noise at frequencies up to about 1 GHz,

$$\frac{R_s}{R_b} \ll \tfrac{1}{2},$$

$$(2\pi f_b)^2 C_d^2 C_b^2 R_s^2 \mathcal{g}_4 \ll C_t^2 R_b^2 \mathcal{g}_2,$$

and then

$$\overline{i_{\text{eff}}^2} = 2f_b e\left[\left\{\left(1+\frac{1}{\beta}\right)I_b + \frac{2\text{k}T}{e\beta^2 R_L} + I_d\right\}\mathcal{g}_o + (2\pi f_b)^2\left\{\frac{2\text{k}TC_d^2R_s}{e} + \left(\frac{I_b}{\beta}+\frac{2\text{k}T}{\beta^2 eR_L}\right)C_t^2\left(\frac{\text{k}T}{eI_b}\right)\right\}\mathcal{g}_2\right]. \quad (17)$$

The base capacitance is the sum of the diffusion, depletion, and stray capacitances. The diffusion and depletion capacitances increase with base bias and the stray capacitance is bias independent. We will approximate $C_t$ by

$$C_t \approx C_\alpha + C_\beta I_b.$$

The optimum bias current is then found by differentiating $\overline{i_{\text{eff}}^2}$; however, a third-order equation results. Assuming most of the noise originates from the first stage ($R_L \to \infty$), the optimum bias current is

$$I_{bo} = \frac{\text{k}T}{e}\omega_b C_\alpha \sqrt{\frac{\mathcal{g}_2}{(1+\beta)\mathcal{g}_o}} \sqrt{\frac{1}{1+\dfrac{\omega_b^2 C_\beta^2\left(\dfrac{\text{k}T}{e}\right)^2\mathcal{g}_2}{(\beta+1)\mathcal{g}_o}}}. \quad (18)$$

The second radical is close to unity under most practical conditions (e.g., up to 1 Gb/s with $\beta = 100$, $C_\beta I_b = 1$ pF) and thus $I_{bo}$ is set by the zero bias capacitance. Substituting $I_{bo}$ into $\overline{i_{\text{eff}}^2}$, again assuming all of the noise originates in the transistor, gives

$$\overline{i_{\text{eff}}^2} = \overline{i_{\text{eff}\,o}^2}\left\{1 + \frac{C_\beta I_{bo}}{C_\alpha}\left(1+\frac{C_\beta I_{bo}}{2C_\alpha}\right)\right\}, \quad (19)$$

where

$$\overline{i_{\text{eff}\,o}^2} = \frac{8\pi k T f_b^2 C_\alpha}{\sqrt{\beta}} \sqrt{g_o g_2 \left(1 + \frac{1}{\beta}\right)}. \tag{20}$$

The term $C_\beta I_{bo}$ is the diffusion capacitance at optimum bias.

We can now go back and determine correction terms for the noise contributed by the series and load resistance terms at $I_{bo}$. The ratio of the effective series noise current to $\overline{i_{\text{eff}\,o}^2}$ is given by

$$\frac{\overline{i_{\text{eff}\,s}^2}}{\overline{i_{\text{eff}\,o}^2}} \approx 2\pi f_b C_d R_s \frac{C_d}{C_\alpha} \sqrt{\frac{g_2}{g_o}\beta} \tag{21}$$

and the ratio of the effective load noise current to $\overline{i_{\text{eff}\,o}^2}$ by

$$\frac{\overline{i_{\text{eff}\,L}^2}}{\overline{i_{\text{eff}\,o}^2}} \approx \frac{4kT}{\beta e R_L I_{bo}}. \tag{22}$$

The input time constant at optimum bias is given approximately by

$$R_b C_t = \frac{C_t}{C_\alpha} \sqrt{\frac{(1+\beta) g_o}{g_2}} \frac{T_b}{2\pi}. \tag{23}$$

Since $g_o > g_2$ and $\beta \gg 1$ for practical conditions, the time constant will greatly exceed the bit interval. Finally, the current gain at optimum bias is given by

$$G(\omega) = \frac{\beta}{1 + j \dfrac{\omega C_t}{\omega_b C_\alpha} \sqrt{\dfrac{(1+\beta) g_o}{g_2}}}.$$

Thus the transfer function is a function of $\omega/\omega_b$ and, for a fixed pulse shape and $\beta$, the frequency dependence of the current gain, $A(\omega)$, scales with bit rate and its magnitude is independent of bit rate.

## V. EXPERIMENTAL RESULTS

The preceding analysis will now be compared with experimental results that have been obtained by Goell at 6.3 Mb/s and 274 Mb/s and by Runge at 50 Mb/s.

The 6.3- and 50-Mb/s repeaters employed 50-percent-duty-cycle rectangular return-to-zero (RZ) optical pulses. For 274 Mb/s, the optical pulses were nonreturn-to-zero (NRZ). For the 6.3-Mb/s repeater the baseband pulses were RZ and for the other two repeaters the baseband signal was NRZ. In all of the cases the baseband pulses were close to the shape to be expected with a raised-cosine spectrum,

Table I — Signal formats and pulse shape integrals

| | Transmitted Pulse | Equalized Pulse | $g_0$ | $g_2$ |
|---|---|---|---|---|
| Case I | 50 percent duty cycle | RZ | 0.30 | 0.13 |
| Case II | 50 percent duty cycle | NRZ | 0.40 | 0.036 |
| Case III | NRZ | NRZ | 0.55 | 0.085 |

but with some intersymbol interference at the highest bit rate. For the purpose of this paper we shall assume a raised-cosine spectrum. Since the error rate is a rapid function of the signal level, the effect of intersymbol interference on error rate can be accounted for by a reduction of received average power equal to the reduction of the worst-case pulse-pattern signal to threshold level. The integrals $g_0$ and $g_2$ are summarized for each of the signaling formats in Table I.

For the 6.3-Mb/s and 50-Mb/s repeaters, SiFET front ends were employed. In the former case, $g_m$ was 0.006 mho and the total circuit capacitance 8 pF and for the latter these parameters were 0.006 mho and 6.7 pF, respectively. For the 274-Mb/s repeater two front ends were tested; the first with a common-emitter followed by a common-collector stage, the second with a common-source GaAsFET input stage followed by an emitter follower. In both cases the input capacitance was 4 pF. The bipolar transistor used was an FMT4000.

Figure 5 shows the base current vs collector current for the FMT4000. The current gain for this device is 160 for base currents to well below 1 $\mu$A.

Both the calculated and measured effective noise current is shown in Fig. 6 for the FMT4000. The experimental values were inferred from the power which gave the signal-to-noise ratio experimentally determined to give a $10^{-9}$ error probability. Experimental curves are given for the input transistor and for the complete receiver. At very low bias currents the noise originating after the first stage becomes significant. Near optimum bias, which was calculated to be 5.5 $\mu$A, the required signal is increased by about 0.5 dB by noise originating after the first stage.

The measured optimum for the first stage is about 4 $\mu$A indicating the base noise is somewhat higher than assumed in the theory. The error-rate performance was found to be optimum at about 7 $\mu$A with about $2\frac{1}{2}$ dB more power than predicted from the calculated first-stage

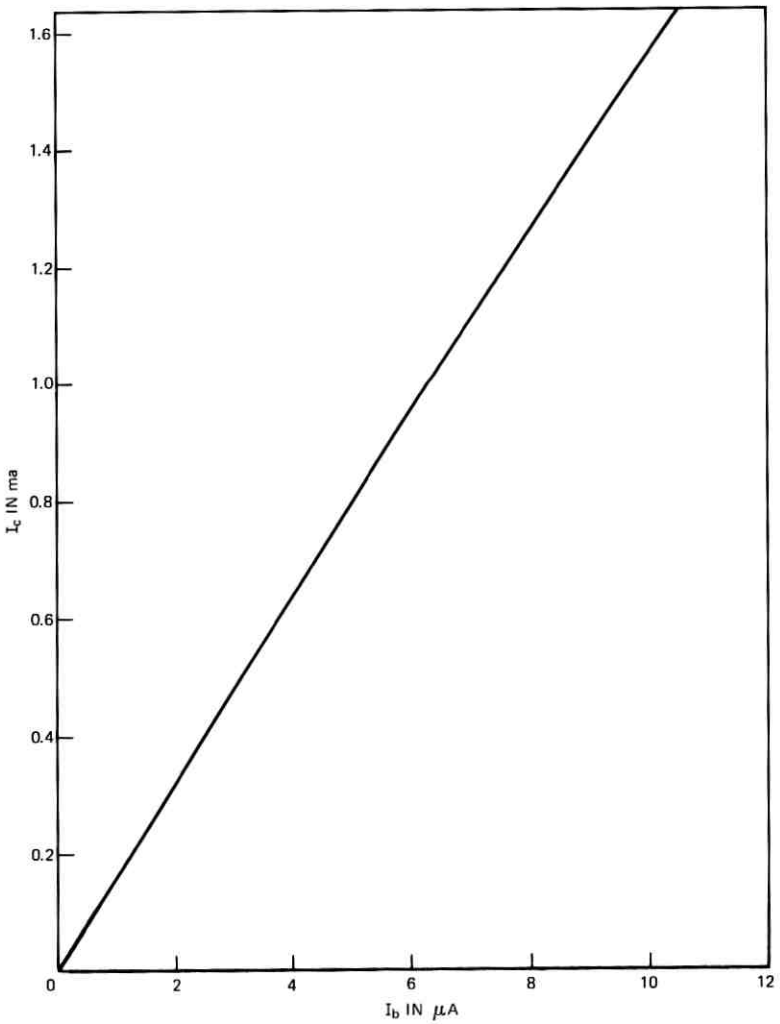



Fig. 5—$I_c$ vs $I_b$ for an FMT4000 bipolar transistor with 10 V collector bias.

and measured subsequent-stage noise. The pulse response for pseudo-random data indicated that about $1\frac{1}{2}$ to 2 dB of this discrepancy were due to intersymbol interference.

The measured and calculated effective noise currents for all of these bit rates are summarized in Table II. For the GaAsFET case an FMT901 transistor was used with a $g_m$ of 0.016 mho. The leakage was

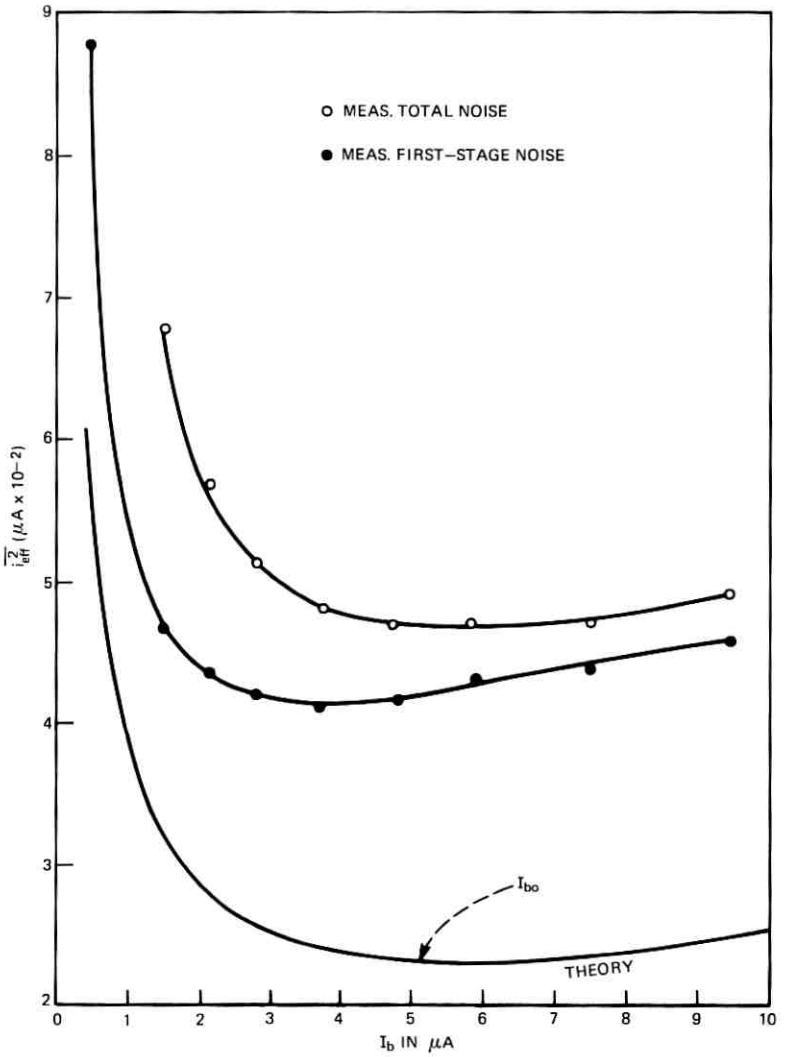



Fig. 6—Effective noise current vs base bias current for an FMT4000 bipolar transistor.

under 1 $\mu$A, which is negligible for 274 Mb/s. Only about half the total noise originated in the GaAsFET. This accounts for a significant part of the error. The noise of the later stages could have been reduced by the inclusion of an extra stage in the front end. However, the performance would still have been inferior to that of a bipolar transis-

Table II — Comparative performance

| Bit Rate (Mb/s) | Transmitted Pulse Shape | Equalized Pulse Shape | Device | $g_m$ (mmho) | $\beta$ | $C_t$ | Measured Total $\sqrt{\overline{i_{eff}^2}}$ (nA) | Measured First Stage $\sqrt{\overline{i_{eff}^2}}$ (nA) | Calculation $\sqrt{\overline{i_{eff}^2}}$ (nA) |
|---|---|---|---|---|---|---|---|---|---|
| 6.3 | 50% RZ | RZ | SiFET | 6 | | 8.0 | 0.46 | 0.40 | 0.39 |
| 50 | 50% RZ | NRZ | SiFET | 6.4 | | 6.7 | 4.2 | * | 3.68 |
| 274 | NRZ | NRZ | BP | | 160 | 4 | 49 | 42 | 23 |
| 274 | NRZ | NRZ | GaAsFET | 16 | | 4 | 79 | 55 | 36 |

* Not available.

tor. The theoretical and experimental effective noise currents shown in Table II for the 6.3-Mb/s and 50-Mb/s cases are in close accord with theory.

## VI. COMPARISON OF DEVICES

Several considerations enter into the selection of an input stage for an optical receiver. Among these factors are sensitivity, dynamic range, power consumption, temperature stability, and cost. Each of these factors, with the exception of the last which is beyond the scope of this paper, will now be discussed.

Equations (15) and (20) indicate that the effective noise for bipolar transistors and field-effect transistors without leakage have a different dependence on frequency. Since the rate of noise increase with frequency is lowest for bipolar transistors one would expect them to be best at high frequencies. At lower frequencies FETs could be expected to be superior. GaAsFETs have appreciable leakage currents. With the FMT901 transistor the leakage varies between 0.1 and 100 $\mu$A, though it is typically below 10 $\mu$A. Thus the noise performance degrades at low frequencies. In addition, $1/f$ noise may be a problem. To date little is known about this source of noise in GaAsFETs and the possibility of its being significant at bit rates on the order of 10 Mb/s cannot as yet be ruled out.

Figure 7 illustrates the dependence of the noise of SiFETs, GaAsFETs, and bipolar transistors as a function of frequency. The leakage current noise, shown separately, must be added to the device noise to get the total circuit noise. This is especially important for GaAsFETs where the leakage of the device can be the limiting factor.

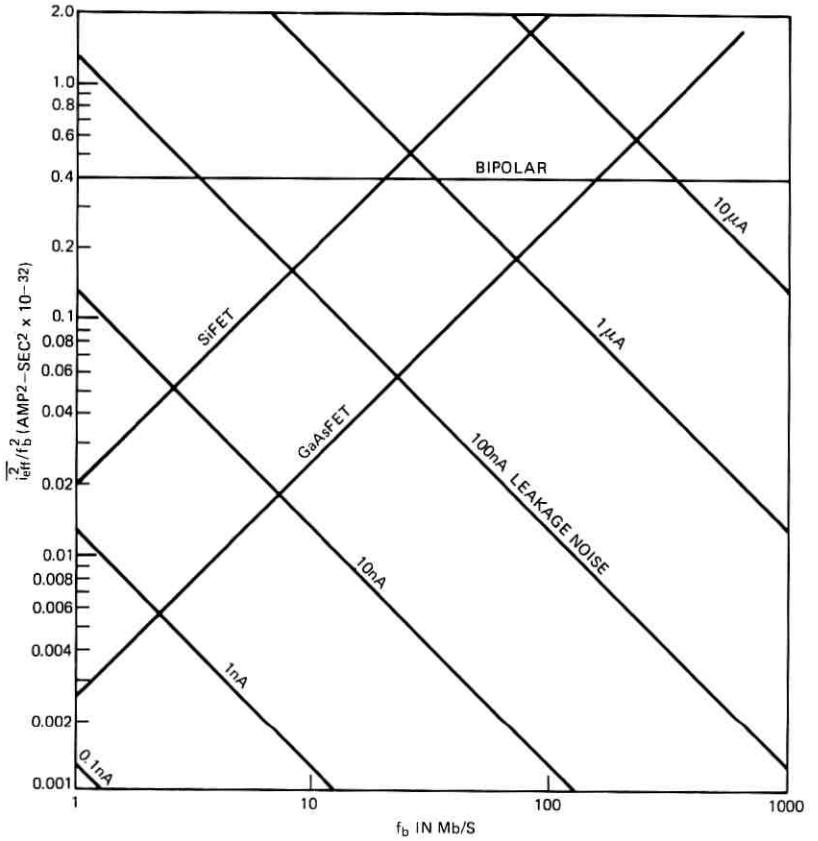



Fig. 7—Effective noise current vs bit rate.

The second case of Table I (50-percent-duty-cycle optical pulses, NRZ baseband) is chosen to illustrate the behavior to be expected. It is assumed that the device parameters of Table III apply. The FET curves terminate at the frequency where $g_m Z_1 = 1$. From the curves for this case bipolar transistors are superior to GaAsFETs above 150 Mb/s and to SiFETs above 20 Mb/s. If leakage were negligible, the GaAsFET would be superior to the SiFET. However, since for currently available devices the leakage runs from about 0.1 to 100 $\mu$A it must be taken into account in the determination of the relative merit of the GaAsFET. For example, with 0.2 $\mu$A of leakage the GaAsFET is superior to the SiFET above about 5 Mb/s. Unless the leakage is below about 1 $\mu$A it is never superior to the bipolar transistor.

Table III — Device parameters for Fig. 7

| Device | $C_t$ (pF) | $g_m$ (mho) | $\beta$ |
|---|---|---|---|
| Bipolar Transistors | 4 | | 160 |
| SiFET | 8 | 0.005 | |
| GaAsFET | 4 | 0.016 | |

The previous cases apply to readily achievable results with commercial devices. However, these devices are not optimized for optical receiver applications. Figure 8 shows the optimum bias as a function of bit rate for each of the cases of Table I. The optimum bias current is extremely small and for many transistors the beta will start to drop before optimum is reached. By reducing the area of devices the range

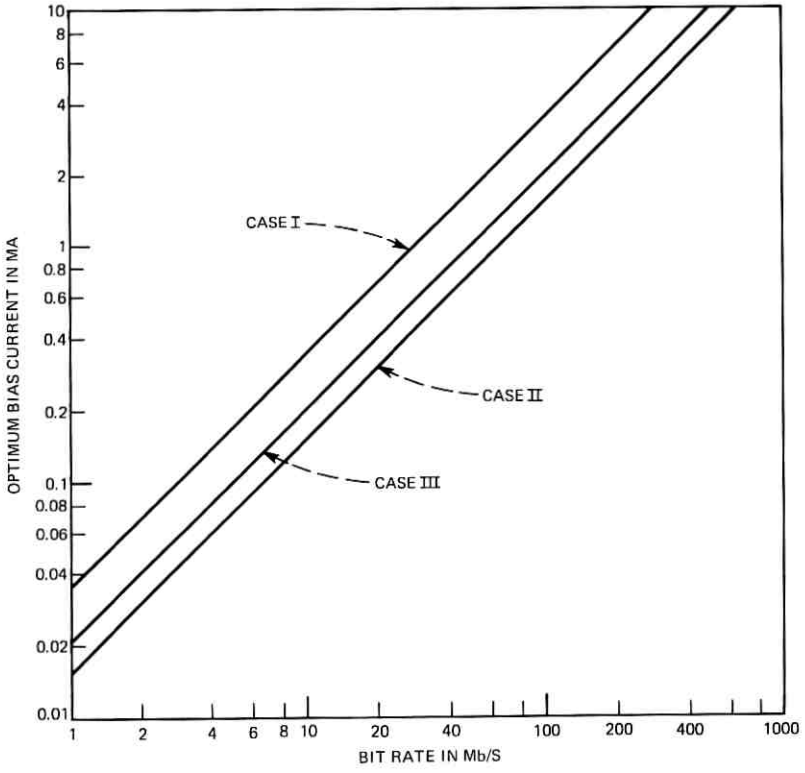



Fig. 8—Optimum bias vs bit rate.

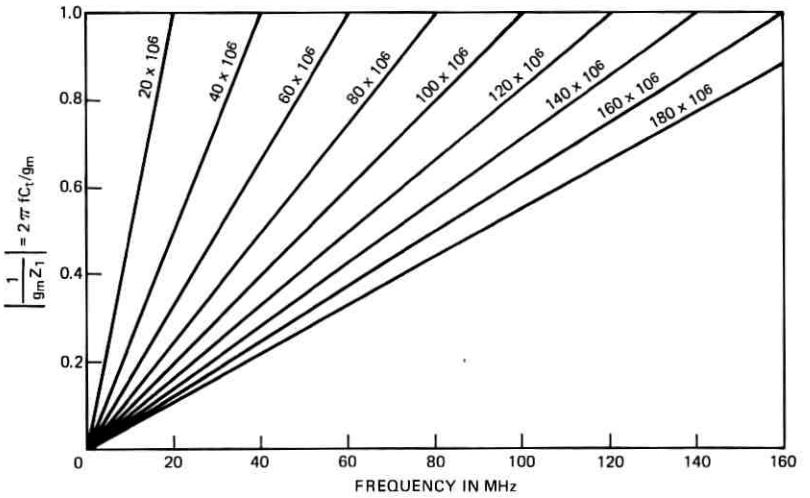



Fig. 9—Reciprocal gain vs frequency with gain bandwidth product ($g_m/2\pi C_t$) as a parameter.

of operation can be extended. Furthermore, it will reduce the capacitance. If the stray and detector capacitances can also be reduced an additional improvement in performance can be achieved.

The upper frequency at which an FET can be used is set by the gain bandwidth product of the device in the circuit. Figure 9 shows the gain versus frequency with $g_m/2\pi C_t$ as a parameter. With an integrating second stage the second-stage noise becomes insignificant if the reciprocal gain, $1/g_m Z_1$, at the highest frequency of interest is below 0.5 while for the noise from a nonintegrating second stage to be insignificant the reciprocal gain must be below about 0.1 at the highest frequency of interest.

For an FET, $g_m$ and $C_g$ are both proportional to gate length.* Increasing gate length increases drain noise ($g_m/C_g^2 \propto 1/\text{length}$). On the other hand, decreasing gate length reduces the total capacitance more slowly than $g_m$ due to the detector and stray capacitance and thus increases the noise contribution of the succeeding stages. Thus an optimum value exists† which can be determined from eq. (15) or eq. (14) if the load noise is frequency independent.

---

* Gate length refers to the dimension perpendicular to current flow and width to the direction of current flow. The former dimension is readily changed. The latter can press the technology.

† G. L. Miller has recently pointed out that $C_g = C_d$ is the optimum condition when the first-stage drain noise predominates.

With improving technology it is expected that the gate width can be significantly reduced. Typical commercial silicon devices with 5-$\mu$m gate widths have gain bandwidth products of 132 MHz; experimental devices with 0.5-$\mu$m gate widths have been reported with a gain bandwidth product near 1 GHz.[10]

We will now direct our attention to the question of dynamic range. Without avalanche gain, for an ac-coupled FET the voltage swing at the detector for the required minimum signal increases with the square root of the reciprocal bit rate. This effect has not been fully analyzed and data are not presently available indicating the bit rate at which amplitude distortion prevents proper equalization. However, at 6.3 Mb/s, problems were not encountered with the signal 10 dB above that required for $10^{-9}$ error rate.

For a bipolar transistor the signal current can become comparable with the bias current leading to gain variations with word pattern for an integrating front end. For example, with a $-31$-dBm signal ($10^{-9}$ error rate at 274 Mb/s) without avalanche gain the signal current is about 0.4 $\mu$A average. Since the optimum bias is near 5 $\mu$A this is appreciable. The ratio $\sqrt{\overline{i_{\text{eff}}^2}}/I_{bo}$ is proportional to $1/\sqrt{C_\alpha}$ and independent of frequency. Thus the ratio of the required minimum signal current to bias current is frequency independent, but can be expected to increase as $C_\alpha$ decreases.

If $C_\alpha$ could be greatly reduced so that

$$\sqrt{\overline{i_{\text{eff}\,o}^2}} \gg I_{bo}$$

then the bias could be turned off and integration would not take place. The noise would only be present when the signal was on and would be given by

$$S_{\text{eq}}^2 = 2eI_{\text{signal}}.$$

This expression is identical to the one which applies to ideal photomultipliers and avalanche detectors with an infinite electron/hole ionization coefficient and large gain; the required signal is 3 dB above the quantum limit.

Present bipolar transistors have a higher transconductance-to-device-current ratio than an FET. Thus, they consume less power. With avalanche gain where the required optical power is relatively insensitive to amplifier noise (varies at $\frac{1}{6}$ root), supply power may often be the deciding factor. It has recently been suggested that FETs have the same limiting transconductance-to-current ratio as bipolar

transistors,[11] $e/kT$. However, as yet this limit is not approached under practical considerations.

Finally, bipolar transistors operated well above the reverse saturation current have better temperature stability than FETs. In applications employing dc coupling this could be an important consideration.

## APPENDIX

The equivalent input noise current spectral density of a cascade of circuits is calculated as follows. First, the circuits are partitioned in any convenient manner. Next, the output impedance of each stage is found going from the input to the output of the cascade since the output impedance is a function of the input loading. Next, the contribution of the equivalent noise current of each stage is referred to the input by dividing it by the product of the short-circuit current gains of all of the preceding stages. Finally, the individual contributions are summed in the square sense, that is,

$$S_{eq} = S_{eq}^1 + \sum_{i=2}^{n} S_{eq}^i \prod_{j=1}^{i-1} \frac{1}{|A_j^2(\omega)|}, \tag{24}$$

where $A_j(\omega)$ is the short-circuit current gain of the $j$th stage and $S_{eq}^i$ is the value of $S_{eq}$ for the $i$th stage.

The current gain, output impedance, and contribution to $S_{eq}$ of $S_1$ and $S_2$ ($S_{eq1}$ and $S_{eq2}$) for the common emitter (source), base (gate), and collector (drain) configurations are summarized in Table IV in terms of the common emitter (source) parameters. The primed and subscripted impedances are the parallel combination of the impedance

Table IV — Equivalent parameters

| Configuration (Common Terminal) | Equivalent Noise Due to $S_1$–$S_{eq1}$ | Equivalent Noise Due to $S_2$–$S_{eq2}$ | Current gain $A(\omega)$ | Output Impedance |
|---|---|---|---|---|
| Emitter (source) | $S_1$ | $\dfrac{S_2}{g_m Z_1' \left(\dfrac{Z_2 - \frac{1}{g_m}}{Z_1' + Z_2}\right)}$ | $-g_m Z_1 \left(\dfrac{Z_2 - \frac{1}{g_m}}{Z_1' + Z_2}\right)$ | $Z_3\left[\dfrac{Z_1' + Z_2}{Z_1' + Z_2 + Z_3 + g_m Z_1' Z_3}\right]$ |
| Base (gate) | $S_1$ | $\dfrac{Z_3 S_2}{Z_1' + g_m Z_1' Z_3}$ | $\dfrac{Z_1' + g_m Z_1' Z_3}{Z_1' + Z_3 + g_m Z_1' Z_3}$ | $Z_2 \dfrac{Z_1' + Z_3 + g_m Z_1' Z_3}{Z_1' + Z_3 + Z_3 + g_m Z_1' Z_3}$ |
| Collector (drain) | $\dfrac{Z_1(g_m Z_2' - 1) S_1}{Z_2'(g_m Z_1 + 1)}$ | $\dfrac{(Z_1 + Z_2) S_2}{Z_2'(1 + g_m Z_1)}$ | $\dfrac{Z_2(1 + g_m Z_1)}{Z_1 + Z_2'}$ | $Z_3 \dfrac{Z_1 + Z_2'}{Z_1 + Z_2' + Z_3 + g_m Z_1 Z_3}$ |

with the same subscript and the output impedance of the preceding stage.

When

$$Z_2 > Z_1', Z_3, Z_1' + Z_3$$

and

$$Z_2 \gg 1/g_m$$

the common emitter (source) and base (gate) configurations give appreciable current gain. Under this condition the approximate relations of Table V apply for $S_{eq1}$, $S_{eq2}$, $A$, and $Z_o$ emitter (source) and base (gate) stages; and for $S_{eq1}$, $S_{eq2}$, and $A$ for a common collector (drain) stage. Table V also gives approximate parameters for output impedance for a common collector (drain) stage for the cases

$$g_m Z_1 \gg 1, \quad Z_2' \gg Z_1$$

and

$$Z_2' \ll g_m Z_1 Z_3, \quad Z_2' \ll Z_1.$$

The first case is typical of bipolar and the second of field-effect transistors. The $\|$ symbol represents the parallel combination of impedances.

Insight into the choice of circuit configuration can be gained by examining the results given in Table V when the preceding assumptions apply. If the gain of a single stage is large enough that the noise of subsequent stages is negligible, then all three configurations give identical results for $S_{eq1}$ and $S_{eq2}$. For this case, the choice of configuration would be made on the basis of output impedance. Since the detector impedance, except at extremely high frequencies, is large, a common base (gate) stage can usually be ruled out and the selection

**Table V — Approximate equivalent parameters**

| Configuration (Common terminal) | $S_{eq1}$ | $S_{eq2}$ | $A(\omega)$ | $Z_0$ |
|---|---|---|---|---|
| Emitter (Source) | $S_1$ | $\frac{S_2}{g_m(Z_s \| Z_1)}$ | $-g_m(Z_s \| Z_1)$ | $Z_3$ |
| Base (Gate) | $S_1$ | $\frac{S_2}{g_m(Z_s \| Z_1)}$ | 1 | $g_m Z_1' Z_3$ |
| Collector (Drain) | $S_1$ | $\frac{S_2}{g_m(Z_s \| Z_1)}$ | $g_m(Z_s \| Z_1)$ | $Z_3 \| \frac{Z_2'}{g_m Z_1}$ (Case I)<br>$\frac{1}{g_m}$ (Case II) |

between the common emitter (source) and common collector (drain) stage made on the basis of desired output impedance. Parenthetically, it is interesting to note that the output voltage for the two cases is identical—the first gives voltage gain but the second presents a higher impedance to the source.

For multistage amplifiers a cascode configuration [common emitter (source) stage followed by a common base (gate) stage] is often chosen for the first two stages. A cascode amplifier has a wider bandwidth than a common emitter (source) stage and thus can simplify equalization. However, it does not lead to the lowest total noise since $A \approx 1$ for a common base (gate) stage and therefore the full value of $S_{eq1}$ for the following stage contributes to the equivalent noise spectral density.

The approach which gives the lowest equivalent noise is a cascade of common emitter (source) stages—possibly followed by a common collector (drain) stage to match impedance—because common emitter (source) stages exhibit the current gain of common collector (drain) stages, but present a higher output impedance to the following stage, thereby minimizing the second-stage $S_{eq2}$ contribution to $S_{eq}$.

## REFERENCES


1. H. Melchior and W. T. Lynch, "Signal and Noise Response of High Speed Germanium Avalanche Photodiodes," IEEE Trans. Elec. Dev., *ED-13*, No. 12 (December 1966), pp. 829–838.
2. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems, I," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 843–875.
3. A. B. Gillespie, *Signals, Noise, and Resolution in Nuclear Counter Amplifiers*, New York: McGraw-Hill, 1953.
4. O. H. Schade, Sr., "A Solid-State Low-Noise Preamplifier and Picture-Tube Drive Amplifier for a 60 MHz Video System," RCA Review, *29*, No. 1 (March 1968), p. 3.
5. J. E. Goell, "An Optical Repeater With High-Impedance Input Amplifier," B.S.T.J., *53*, No. 4 (April 1974), pp. 629–643.
6. P. Runge, private communication.
7. A. van der Ziel, *Noise: Sources, Characterization, Measurement*, Englewood Cliffs, N. J.: Prentice-Hall, 1970.
8. Das B. MuKunda, "FET Noise Sources and Their Effects on Amplifier Performance at Low Frequencies," IEEE Trans. Elec. Dev., *ED-19*, No. 3 (March 1972), pp. 338–348.
9. W. Baechtold, "Noise Behavior of GaAs Field-Effect Transistor with Short Gate Lengths," IEEE Trans. Elec. Dev., *ED-19*, No. 5 (May 1972), pp. 674–680.
10. W. Baechtold, K. Dactwyler, T. Forster, T. O. Mohr, W. Walter, and P. Wolf, "Si and GaAs 0.5 μm-Gate Schottky-Barrier Field Effect Transistors," Elec. Letters, *9*, No. 10 (May 17, 1973), pp. 232–234.
11. F. O. Johnson, "The Insulated Gate Field-Effect Transistor—A Bipolar Transistor in Disguise," RCA Review, *34*, No. 3 (March 1973), pp. 80–94.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 9, November 1974
*Printed in U.S.A.*

# Reduction of Multimode Pulse Dispersion by Intentional Mode Coupling

By D. MARCUSE

(Manuscript received May 3, 1974)

*Guidelines for the design of multimode, step-index fibers with intentional fluctuations of the refractive index of the core are given, with the aim of reducing multimode pulse dispersion. It appears possible to engineer a fiber with carefully designed refractive index fluctuations, the azimuthal variation of which is governed by the function cos $\phi$ and the z dependence of which has a spatial Fourier spectrum with a sharp cutoff frequency. By limiting the location of the index fluctuations to a region below a certain radius $r_{max}$, coupling to modes with large azimuthal mode numbers can be avoided and power loss via coupling to radiation modes can be held to a minimum.*

## I. INTRODUCTION

Optical fibers supporting many guided modes suffer from multimode dispersion. A pulse launched into a multimode fiber excites many modes, each traveling at a different group velocity. At the far end of the fiber the pulse is spread out in time because of the different group delays of each mode. This multimode dispersion effect is usually more serious than the single-mode dispersion caused by the dispersive effect of the dielectric material of the waveguide core and by the inherently dispersive nature of mode guidance. Discussions of multimode dispersion in the absence of mode coupling can be found in Refs. 1, 2, and 3.

S. D. Personick discovered that multimode dispersion in fibers can be reduced by intentional (or unintentional) mode coupling. If the power carried in the fiber transfers back and forth between slow and fast modes, averaging takes place, so that the pulse no longer breaks up into a sequence of pulses but is forced to travel at an average group delay with a concomitant reduction in pulse spreading. Although the spread of a pulse carried by uncoupled modes is proportional to the

length of the fiber, it only becomes proportional to the square root of its length if the pulses are coupled among each other.[3–5]

However, reduction of multimode pulse dispersion by means of mode coupling must be bought at a price. Any mechanism that causes coupling among the guided modes also tends to couple guided modes to the continuum of radiation modes. Power coupled into radiation modes radiates away causing losses. Radiation loss can be reduced by careful control of the coupling process.[5] It is possible, at least in principle, to provide strong coupling among the guided modes but only very little coupling to radiation modes. The loss penalty can thus be controlled and kept to small amounts.[3,5]

Coupling between two fiber modes is caused by a specific spatial frequency of the Fourier spectrum of the coupling function. Two modes couple via a spatial frequency that is equal to the difference of the propagation constants of the two modes. Control of the loss penalty for multimode dispersion is thus possible by shaping the Fourier spectrum of the coupling function. In general, it is desirable to achieve a spectrum that provides a sufficient number of spatial frequencies below a critical frequency and a sharp cutoff of the spectrum at the critical spatial frequency.[3,5]

In this paper we discuss means of mode coupling by employing intentional fluctuations of the refractive index of a fiber whose unperturbed core has a constant index of refraction (step-index fibers). It is necessary to shape the core-index fluctuations so that only modes with adjacent azimuthal mode numbers $\nu$ couple to each other. Additional control of the coupling process must be provided by a sharp cutoff of the coupling spectrum that can be achieved by careful design of the $z$ dependence ($z$ is the axial direction) of the index fluctuations. Finally, radiation losses can be minimized by limiting the index fluctuations to a region near the fiber axis.

The paper begins with a discussion of the requirements on the Fourier spectrum imposed by the desire to minimize the loss penalty. Next we provide explicit expressions for the power-coupling coefficients and estimate the amount of index fluctuation that is necessary to achieve a desired reduction in multimode dispersion.

This discussion is intended as a guide to the fiber designer, pointing out the possibilities available for reduction of multimode dispersion and explaining the difficulties that must be overcome.

## II. SHAPING THE SPATIAL FOURIER SPECTRUM

We consider two modes with propagation constants $\beta_i$ and $\beta_j$. Interaction between these modes is described by a coupling coefficient that

depends on the distortion of the core boundary or on refractive-index irregularities. The azimuthal symmetry of the irregularity provides selection rules for the coupling process. The $z$ dependence of the irregularity enters the coupling process via its spatial Fourier spectrum. A sinusoidal component of the Fourier spectrum of the form

$$F(\theta) \cos \theta z \tag{1}$$

couples the two modes only if the relation[3,6]

$$|\beta_i - \beta_j| = \theta \tag{2}$$

is satisfied. (This requirement stems from first-order perturbation theory and is valid provided that the coupling is weak.)

It is not hard to envision a sufficient number of spatial frequencies that couple all modes among each other. However, as pointed out in Section I, coupling to radiation modes causes power loss by radiation from the fiber core.[3,6] It is thus essential to avoid coupling between guided and radiation modes. To see whether this is possible, we must study the spacing (in $\beta$ space) between the guided modes. I have computed the propagation constants of all the guided modes for a step-index fiber with

$$V = (n_1^2 - n_2^2)^{\frac{1}{2}} ka = 40, \tag{3}$$

where $n_1$ = refractive index of fiber core, $n_2$ = index of cladding, $k$ = free-space propagation constant, $a$ = fiber core radius. The propagation constants were obtained as solutions of the simplified eigenvalue equation of the optical fiber.[1,3] When the propagation constants are listed in order of their numerical values, regardless of mode number, they appear approximately evenly spaced. This behavior of the propagation constants of fiber modes contrasts with the behavior of the modes of a dielectric slab. Here we find that the spacings between modes increase monotonically so that the spatial frequencies, required to couple nearest neighbors, also increase.[3] By providing a cutoff to the Fourier spectrum of spatial frequencies contained in the coupling function, we can provide coupling among lower-order modes of the slab and uncouple either the last guided mode or a few of the higher-order guided modes, depending on the value of the spatial cutoff frequency (see Fig. 1). Once the highest-order (or a few high-order) guided modes are uncoupled from the rest, there is no danger of incurring radiation loss caused by the coupling process. Things are not quite that simple in the round optical fiber because the modes are not naturally arranged with ever-increasing spacings between neighbors.

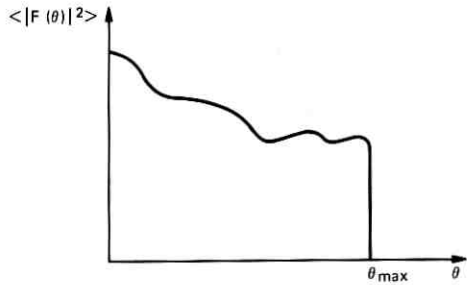



Fig. 1—Schematic drawing of the desired spatial Fourier spectrum of the $z$ dependence of the index fluctuations. The shape of the function is unimportant except for its abrupt cutoff.

A good approximation to the actual solution of the eigenvalue equation is obtained if we approximate the Bessel function by the formula[7]

$$J_\nu(x) = \sqrt{\frac{2}{\pi}} \frac{\cos\left[(x^2 - \nu^2)^{\frac{1}{2}} - \nu \arccos\left(\frac{\nu}{x}\right) - \frac{\pi}{4}\right]}{(x^2 - \nu^2)^{\frac{1}{4}}}. \tag{4}$$

In weakly guiding fibers the transverse electric field component can be represented as [1,3]

$$E_y = AJ_\nu(\kappa r) \cos(\nu\phi) e^{-i\beta z}, \tag{5}$$

with

$$\kappa = (n_1^2 k^2 - \beta^2)^{\frac{1}{2}}. \tag{6}$$

To a good approximation, we may assume that $E_y = 0$ at the core radius $r = a$. This approximation is better for modes far from their cutoff value but it gives a reasonable indication of the propagation constants for practically all modes. An approximate eigenvalue equation of the guided modes thus follows from (4),

$$[(\kappa a)^2 - \nu^2]^{\frac{1}{2}} - \nu \arccos\left(\frac{\nu}{\kappa a}\right) - \frac{\pi}{4} = (2m - 1)\frac{\pi}{2} \tag{7}$$

for $m = 1, 2, 3, \cdots$. By regrouping this equation we obtain a form that is useful for iterative solutions,

$$\kappa a = \left\{\nu^2 + \left[(m - \tfrac{1}{4})\pi + \nu \arccos\left(\frac{\nu}{\kappa a}\right)\right]^2\right\}^{\frac{1}{2}}. \tag{8}$$

Using (6) and (7) we derive the following approximate expression for

the spacing between guided modes,

$$\Delta\beta = -\frac{\kappa^2}{\beta}\left\{\frac{\Delta\nu}{\nu} + \frac{\pi\Delta m - \frac{\Delta\nu}{\nu}(m - \frac{1}{4})\pi}{[(\kappa a)^2 - \nu^2]^{\frac{1}{2}}}\right\}. \tag{9}$$

$\Delta\beta$ is the spacing between modes that are separated by an amount $\Delta\nu$ of the azimuthal mode number $\nu$ and by a change $\Delta m$ of the radial mode number $m$.

If we place no restriction on the allowed values of $\Delta\nu$ or $\Delta m$, it is impossible to shape the spatial Fourier spectrum of the coupling function so that coupling to radiation modes is avoided. It is thus necessary to introduce definite "selection rules" for the coupling among the guided modes. We shall see later that it is possible to shape the refractive-index distribution or the deformation of the core-cladding boundary such that only modes with

$$\Delta\nu = \pm 1 \tag{10}$$

can couple to each other. We shall thus assume that the selection rule (10) is enforced and continue our discussion on this assumption. The allowed values for $\Delta m$ remain arbitrary. However, it is true that the spatial frequencies for coupling between modes (that is, the value of $\Delta\beta$) are larger for larger values of $\Delta m$. Since it is our aim to introduce a cutoff frequency into the spatial Fourier spectrum so that modes with large spatial frequency separation will be uncoupled, we restrict the discussion also to a limited range of values for $\Delta m$ and consider only the case

$$\Delta m = 0 \quad \text{or} \quad \pm 1. \tag{11}$$

Using (10) and taking $\Delta m = 0$ we obtain from (9)

$$|\Delta\beta| = \frac{\kappa^2}{\nu\beta}\left|\frac{(m - \frac{1}{4})\pi}{[(\kappa a^2) - \nu^2]^{\frac{1}{2}}} - 1\right| \quad \text{for} \quad \begin{matrix}\Delta m = 0 \\ \Delta\nu = \pm 1.\end{matrix} \tag{12}$$

For $\Delta m = \pm 1$ we obtain from (10) and (9)

$$|\Delta\beta| = \frac{\kappa^2}{\nu\beta}\left|\frac{[(m - \frac{1}{4}) + \nu]\pi}{[(\kappa a)^2 - \nu^2]^{\frac{1}{2}}} - 1\right| \quad \text{for} \quad \begin{matrix}\Delta m = -\Delta\nu \\ \Delta\nu = \pm 1\end{matrix}. \tag{13}$$

The case $\Delta m = +\Delta\nu$ has been excluded since it leads to larger spatial frequencies than those obtained from (13).

Figures 2 and 3 illustrate the boundaries of various regions in mode-number space $\nu$, $m$. Both figures were drawn for $V = 40$ [see eq. (3)], $n_1 = 1.515$, and $n_2 = 1.5$ so that $n_1/n_2 = 1.01$. The solid line delineates

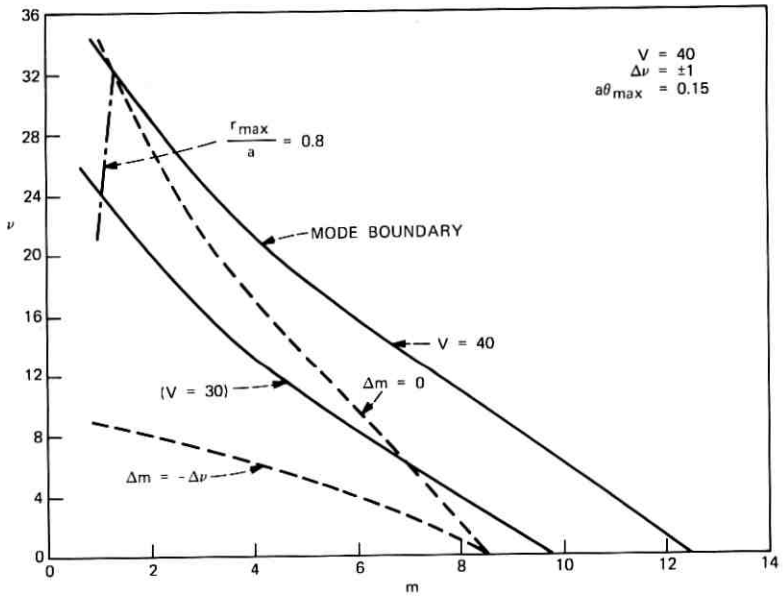



Fig. 2—Various regions in mode-number space $\nu$, $m$. All curves belong to $V = 40$, $n_2 = 1.5$, and $n_1/n_2 = 1.01$ with the exception of the solid line labeled ($V = 30$). The solid lines indicate the boundaries of the range of guided modes for the respective $V$ values. The broken line labeled $\Delta m = 0$ delineates the area below which coupling of modes with the selection rule $\Delta\nu = \pm 1$ and $\Delta m = 0$ is possible. The broken line labeled $\Delta m = -\Delta\nu$ limits the range of coupling with the selection rule $\Delta\nu = \pm 1$, $\Delta m = -\Delta\nu$. The dash-dotted line is the limit of the coupling range caused by the location of $r = r_{max} = 0.8a$. The spatial Fourier spectrum cuts off at $\theta_{max}a = 0.15$.

the boundary of guided modes; it is obtained by plotting those values of $\nu$ and $m$ that result in $\kappa a = V$ [see eq. (14)]. All guided modes are located to the left and below the solid line. The broken lines* represent lines of constant spatial frequency. They result from plotting the combinations of $\nu$ and $m$ values that result in $\Delta\beta = \theta_{max}$. The line labeled $\Delta m = 0$ was computed from (12) and the line $\Delta m = -\Delta\nu$ was obtained from (13). The broken lines delineate the boundaries for mode coupling with the spatial Fourier spectrum of the coupling function of Fig. 1, the cutoff frequency of which is $\theta = \theta_{max}$. Modes below and to the left of the broken lines couple to their nearest neighbors via the selection rule $\Delta\nu = \pm 1$ and $\Delta m = 0$ or $\Delta m = -\Delta\nu$. Modes located to the right and above the broken lines cannot couple to each other

* The meaning of the dash-dotted lines and the solid line labeled ($V = 30$) will be explained later.

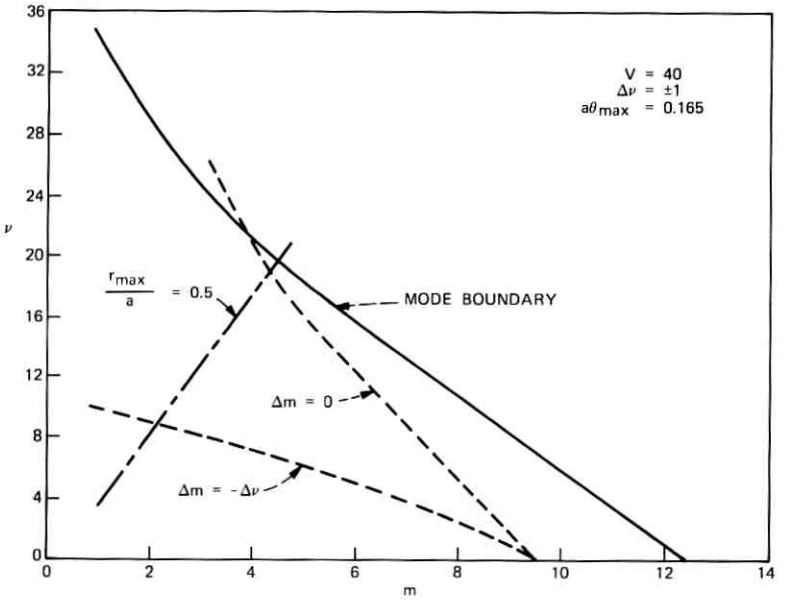



Fig. 3—Same as Fig. 2 except that $\theta_{max}a = 0.165$, $r_{max} = 0.5a$.

since no spatial frequencies achieving this coupling are available. Figures 2 and 3 differ only in the choice of the cutoff frequency $\theta_{max}$. There are no transitions with $\Delta\nu = \pm 1$ that couple via smaller spatial frequencies than the ones indicated in the figures. It is thus apparent that it is possible to provide coupling among most guided modes by means of the Fourier spectrum shown in Fig. 1. However, modes to the right and above the uppermost broken line remain uncoupled. It is necessary to uncouple a few higher-order modes in order to avoid coupling into the continuum of radiation modes. The conditions shown in Fig. 2 achieve this goal almost completely. Only the mode $m = 1$, $\nu = 34$, lying on the boundary of the guided-mode region, is coupled to radiation modes as well as the other guided modes. Power is thus able to flow out of the guided-mode region causing radiation losses via this one guided mode. This power loss could be avoided by decreasing $\theta_{max}$.

The conditions prevailing in Fig. 3 would result in a high loss penalty since all modes with $m < 4$, $\nu > 21$ on the boundary of the guided-mode region couple to guided as well as radiation modes. However, we shall show later that it is possible to prevent mode coupling for modes exceeding a certain maximum $\nu$ value that can be chosen by a suitable

design of the intentional index fluctuations. It is thus possible to achieve a low loss penalty even for the conditions shown in Fig. 3 provided the coupling mechanism is carefully designed to avoid coupling for modes with $\nu > 20$.

Modes located between the two broken lines in Figs. 2 and 3 can couple only to their neighbors above and below in the mode-number plane. However, since the members with low $\nu$ values below the line $\Delta m = -\Delta\nu$ are able to couple to their neighbors to the left and to the right, all modes below the uppermost broken line are actually coupled together.

We can give an approximate rule for calculating the spatial cutoff frequency appearing in Figs. 2 and 3. We begin by specifying the maximum $\nu$ value on the mode boundary for which mode coupling should cease. As mentioned earlier, the design specification for this value $\nu_{\max}$ will be given in the section on mode coupling. Next we need to know the corresponding value of $m$ on or near the mode boundary—the solid line in Figs. 2 and 3. We obtain it from the cutoff condition $\kappa a = V$ and (8),

$$m_{\max} = \frac{1}{4} - \frac{\nu_{\max}}{\pi} \arccos\left(\frac{\nu_{\max}}{V}\right) + \frac{1}{\pi}(V^2 - \nu_{\max}^2)^{\frac{1}{2}}. \tag{14}$$

Substitution of $\nu_{\max}$ and $m_{\max}$ into (12) using $\beta = n_2 k$ yields the desired value for $\Delta\beta = \theta_{\max}$. For $V = 40$ and $\nu_{\max} = 20$ we obtain from (14) $m_{\max} = 4.61$ and from (12) $\theta_{\max} a = 0.17$ in agreement with Fig. 3. Of course, it does not make physical sense to use a noninteger $m_{\max}$, but it is advisable to use this value in (12) in order to obtain a more accurate value of $\theta_{\max}$. Incidentally, (14) defines the mode boundary if we use it for all possible values $\nu = \nu_{\max}$.

## III. POWER COUPLING COEFFICIENTS

Mode coupling in multimode dielectric optical waveguides is most conveniently described by a coupled-mode theory. The power coupling coefficients are defined as follows:[3,5]

$$h_{\nu n,\mu m} = \langle |K_{\nu n,\mu m}|^2 \rangle. \tag{15}$$

The symbol $\langle\ \rangle$ indicates an ensemble average. The coefficient $K_{\nu n,\mu m}$ stems from the coupled amplitude equations and is defined as[8]

$$K_{\nu n,\mu m} = \frac{\omega \epsilon_0}{4iP} \int_0^{2\pi} d\phi \int_0^{\infty} r dr (n^2 - n_0^2) \mathcal{E}_{\nu n}^* \cdot \mathcal{E}_{\mu m}. \tag{16}$$

The angular frequency of the radiation is $\omega$, $\epsilon_0$ is the dielectric permit-

tivity of vacuum, and $P$ designates the power normalization constant of the modes, the electric field vectors of which are indicated by script letters. The refractive-index distribution of the actual guide with index fluctuations is $n$ while $n_0$ indicates the index distribution of a perfect guide from which the actual guide deviates only slightly. Our interest is focused on introducing intentional index fluctuations for the purpose of mode coupling. We are thus free to choose $n$ to achieve our goal. It was pointed out earlier that coupling to radiation modes is unavoidable unless certain selection rules are imposed on the coupling process. Our discussion in the last section was based on the selection rule $\Delta\nu = \pm 1$. To achieve this selection rule we must require that the refractive-index distribution be of the following general form:

$$n^2 - n_0^2 = 2n_1\Delta n g(r) f(z) \cos\phi. \tag{17}$$

We know from earlier work that the $\phi$-dependence of this index distribution leads to the desired selection rule.[9]

We found in the preceding section that it is also desirable to avoid coupling among modes with large $\nu$ values. It follows from the properties of Bessel functions that the field intensity of the transverse field components is very weak for radii that obey the relation

$$\kappa r < \nu.$$

This result is easily interpreted in terms of ray optics. Modes with large values of $\nu$ are represented by skew rays that spiral around the fiber axis. These rays avoid the vicinity of the waveguide axis and stay nearer to the core boundary for larger values of $\nu$. The radius defined by $\kappa r = \nu$ represents the turning point below which a ray with a given value of $\nu$ does not penetrate. The coupling formula (16) shows that mode coupling depends on the field strength at the point where the index irregularity is located. By providing refractive-index variations that do not extend beyond a radius $r_{max}$, it is possible to limit mode coupling to modes whose $\nu$ values remain below a maximum value near the mode boundary that is defined as

$$\nu_{max} = V\frac{r_{max}}{a}. \tag{18}$$

$\kappa a$ has been replaced by its maximum value $V$. The values of $\nu_{max}$ and the corresponding values for $m_{max}$ and $\theta_{max}$ defined by (14) and (12) determine the position at which the dotted curves labeled $\Delta m = 0$ cross the solid curves in Figs. 2 and 3 that define the guided-mode boundary in mode-number space. Coupling of guided modes to radia-

tion modes can be avoided by limiting the $\nu$ values of those modes that are coupled by index fluctuations. If the intentional index fluctuations do not extend beyond the radius $r_{max}$, coupling is restricted to those modes that remain below a boundary defined by the equation

$$\nu = \kappa r_{max}. \tag{19}$$

If we combine this equation with formula (8) we obtain the function $m = m(\nu)$ that defines the boundary in mode-number space beyond which mode coupling ceases, because the index fluctuations are restricted to radii $r \leqq r_{max}$,

$$m = \frac{1}{4} + \frac{\nu}{\pi}\left\{\left(\frac{a^2}{r_{max}^2} - 1\right)^{\frac{1}{2}} - \arccos\frac{r_{max}}{a}\right\}. \tag{20}$$

This boundary is shown as a dash-dotted line in Figs. 2 and 3. We now have the means of providing coupling among all guided modes that remain inside the nearly triangular areas that are bounded by the dotted lines labeled $\Delta m = 0$ and the dash-dotted lines in Figs. 2 and 3. If these areas remain below the guided-mode boundary (the solid line labeled $V = 40$ in Fig. 2), coupling to radiation modes can be avoided.

After this digression into the fundamental properties of mode coupling we proceed with the derivation of specific coupling formulas. Substitution of the field vectors, given by (5) and in more detail by Ref. 3, into (16) we obtain with the help of (17)

$$K_{\nu n,\mu m} = \frac{k\gamma_{\nu n}\gamma_{\mu m} f(z)\int_0^a rg(r)J_\nu(\kappa_{\nu n}r)J_\mu(\kappa_{\mu m}r)dr}{2in_1V^2\left|J_{\nu-1}(\kappa_{\nu n}a)J_{\nu+1}(\kappa_{\nu n}a)J_{\mu-1}(\kappa_{\mu m}a)J_{\mu+1}(\kappa_{\mu m}a)\right|^{\frac{1}{2}}}\delta_{\nu\pm1,\mu}. \tag{21}$$

The parameter $\gamma_{\nu n}$ is defined as

$$\gamma_{\nu n} = (\beta_{\nu n}^2 - n_2^2k^2)^{\frac{1}{2}} \tag{22}$$

and $\delta_{\nu\mu}$ is Kronecker's delta symbol.

Of the many possible choices for the function $g(r)$ we use only two examples that may be of particular practical interest. First we use

$$g(r) = W\delta(r - r_{max}). \tag{23}$$

$W$ is the very narrow width of the ring of index fluctuations. We substitute this equation into (21) but proceed immediately to the power coupling coefficient (15). The Fourier spectrum is defined as

$$F(\theta) = \lim_{L\to\infty}\frac{1}{\sqrt{L}}\int_0^L f(z)e^{-i\theta z}dz \tag{24}$$

and the power coupling coefficient for a narrow, ring-shaped refractive-index fluctuation is given as

$$h_{\nu n,\mu m} = \left\{ \frac{\Delta n W k r_{\max} \gamma_{\nu n} \gamma_{\mu m} J_\nu(\kappa_{\nu n} r_{\max}) J_\mu(\kappa_{\mu m} r_{\max})}{V^2 J_{\nu-1}(\kappa_{\nu n} a) J_{\mu-1}(\kappa_{\mu m} a)} \right\}^2 \langle |F|^2 \rangle \delta_{\nu \pm 1,\mu}. \quad (25)$$

Because the relation

$$J_\nu(\kappa_{\nu n} a) \approx 0 \quad (26)$$

is approximately valid we have used

$$J_{\nu+1}(\kappa_{\nu n} a) = -J_{\nu-1}(\kappa_{\nu n} a). \quad (27)$$

As a second example we consider the function

$$g(r) = \begin{cases} 1 & \text{for } r < r_{\max} \\ 0 & \text{for } r > r_{\max}. \end{cases} \quad (28)$$

Limiting the index fluctuations to a wide range, $0 < r < r_{\max}$. In this case the power coupling coefficient assumes the form

$$h_{\nu n,\mu m} = \left\{ \frac{\Delta n k r_{\max} \gamma_{\nu n} \gamma_{\mu m} [\kappa_{\mu m} J_\nu(\kappa_{\nu n} r_{\max}) J_{\mu-1}(\kappa_{\mu m} r_{\max}) - \kappa_{\nu n} J_{\nu-1}(\kappa_{\nu n} r_{\max}) J_\mu(\kappa_{\mu m} r_{\max})]}{V^2 [\kappa_{\nu n}^2 - \kappa_{\mu m}^2] J_{\nu-1}(\kappa_{\mu n} a) J_{\nu-1}(\kappa_{\mu m} a)} \right\}^2 \cdot \langle |F|^2 \rangle \delta_{\nu \pm 1,\mu}. \quad (29)$$

The argument of the power spectrum $\langle |F^2| \rangle$ in (25) and (29) is $\beta_{\nu n} - \beta_{\mu m}$.

## IV. PULSE WIDTH REDUCTION BY INTENTIONAL MODE COUPLING

Random coupling of the modes of a multimode fiber causes the many pulses, traveling on different guided modes at different group velocities, to be coupled together so that an equilibrium pulse establishes itself traveling at an average group velocity. Its ensemble average has a gaussian shape[3,5] the width of which is given by the formula[10]

$$T = 4(\rho_2 L)^{\frac{1}{2}}, \quad (30)$$

with[11]

$$\rho_2 = \sum_{i=2}^{N} \frac{\left[ \sum_{\nu=1}^{N} B_\nu^{(1)} \left( \frac{1}{v_\nu} - \frac{1}{v_a} \right) B_\nu^{(i)} \right]^2}{\rho_0^{(i)} - \rho_0^{(1)}}. \quad (31)$$

Equation (30) shows that the width of the pulse grows proportionally to the square root of the length $L$ of the fiber. The parameter $\rho_2$ is the second perturbation of the first eigenvalue that results from an algebraic eigenvalue problem, $B_\nu^{(i)}$ are the components of the $i$th eigen-

vector, $\rho_0^{(i)}$ is the $i$th eigenvalue of zero order, and $v_\nu$ and $v_a$ are the group velocity of mode $\nu$ and the average group velocity.

The evaluation of this expression is difficult. If only a few guided modes exist, computer solutions of the eigenvalues and eigenvectors can be obtained. If many modes are guided, it is possible to convert the algebraic eigenvalue problem to a partial differential equation, provided that it can be assumed that only nearest neighbors couple among each other.[3,12] For modes close to the edge of the coupling range, indicated by the dotted lines in Figs. 2 and 3, the assumption of nearest-neighbor coupling is well justified since the spatial Fourier spectrum of the coupling function lacks the higher frequencies required to couple a mode to a neighbor farther away. However, the lower-order modes are crowded more closely so that there are spatial frequencies available for coupling to modes other than the nearest neighbors. This increase in coupling strength is partially compensated for by the fact that modes above the dotted line labeled $\Delta m = -\Delta\nu$ in Figs. 2 and 3 can couple only vertically. In order to be able to give an order-of-magnitude estimate of the index fluctuations required to achieve a certain reduction of the pulse width, we shall assume that nearest-neighbor coupling can be assumed and provide an expression for the pulse width reduction that may be regarded as a crude approximation.

We use the theory presented in Ref. 3. The only modification necessary to the formulas presented in Section 5.6 of Ref. 3 consists in the realization that our coupling scheme avoids radiation losses, so that the highest-order members of the group of coupled modes are not depleted contrary to the assumption in Ref. 3. The eigenvectors and eigenvalues defined on p. 235 of Ref. 3 can be used, except that the first eigenvector is now constant, independent of the mode number. All eigenvectors are mutually orthogonal and properly normalized. Using the procedure explained in Ref. 3 we obtain the result

$$R = \frac{T}{\Delta\tau} = \frac{0.225V}{(Lh)^{\frac{1}{2}}}. \tag{32}$$

$T$ is the full width at the $1/e$ points of the gaussian-shaped pulse, $\Delta\tau$ is the width of the pulse train that would exist in the absence of coupling, $V$ is the normalized frequency parameter defined by (3), $L$ is the length of the fiber, and $h$ the power coupling coefficient. It was assumed that only nearest neighbors couple to each other and the strength of these coupling coefficients was assumed to be identical for all the modes.

$R$ is the improvement factor that indicates the reduction of the length of the pulse relative to its uncoupled length. It is thus desirable to achieve as small a value of $R$ as possible. Values of $R > 1$ do not describe a physically meaningful situation. The formula may result in values $R > 1$. This indicates that the coupling is not strong enough to achieve an equilibrium pulse in the length of fiber available.

We are discussing numerical values for the two types of couplings described by (25) and (29). Values have been computed for

$$\bar{h}_{\nu n;\nu+1,m} = \left[ \frac{\gamma_{\nu n} a \gamma_{\nu+1,m} a J_\nu(\kappa_{\nu n} r_{\max}) J_{\nu+1}(\kappa_{\nu+1,m} r_{\max})}{J_{\nu-1}(\kappa_{\nu n} a) J_\nu(\kappa_{\nu+1,m} a)} \right]^2 \tag{33}$$

and

$$\bar{H}_{\nu n;\nu+1,m} = \left\{ \frac{\gamma_{\nu n} a \gamma_{\nu+1,m} a [\kappa_{\nu+1,m} J_\nu(\kappa_{\nu n} r_{\max}) J_\nu(\kappa_{\nu+1,m} r_{\max}) - \kappa_{\nu n} J_{\nu-1}(\kappa_{\nu n} r_{\max}) J_{\nu+1}(\kappa_{\nu+1,m} r_{\max})]}{a(\kappa_{\nu n}^2 - \kappa_{\nu+1,m}^2) J_{\nu-1}(\kappa_{\nu n} a) J_\nu(\kappa_{\nu+1,m} a)} \right\}^2 \tag{34}$$

for all the modes of a fiber with $V = 40$, $n_1 = 1.515$, and $n_1/n_2 = 1.01$. A few sample values are listed in Tables I and II. The values listed

**Table I — Sample values of normalized coupling coefficients for a narrow ring of refractive-index fluctuations located at $r = r_{\max}$ with $r_{\max}/a = 0.8$**

| $\nu$ | $n$ | $\bar{h}_{\nu n;\nu+1,n}$ | $\bar{h}_{\nu n;\nu+1,n-1}$ |
|---|---|---|---|
| 1 | 1 | $2.212\ 10^6$ | 0. |
| 1 | 2 | $4.608\ 10^6$ | $3.987\ 10^6$ |
| 1 | 3 | $1.023\ 10^6$ | $2.879\ 10^6$ |
| 1 | 4 | $1.287\ 10^3$ | $7.733\ 10^4$ |
| 1 | 5 | $4.780\ 10^5$ | $5.678\ 10^3$ |
| 1 | 6 | $3.844\ 10^6$ | $1.931\ 10^6$ |
| 1 | 7 | $2.726\ 10^6$ | $4.107\ 10^5$ |
| 1 | 8 | $7.018\ 10^4$ | $8.497\ 10^5$ |
| 1 | 9 | $2.262\ 10^4$ | $5.610\ 10^3$ |
| 1 | 10 | $2.431\ 10^6$ | $5.888\ 10^5$ |
| 1 | 11 | $4.236\ 10^6$ | $4.181\ 10^6$ |
| 1 | 12 | $3.964\ 10^5$ | $2.115\ 10^6$ |
| 10 | 1 | $4.945\ 10^7$ | 0. |
| 10 | 2 | $7.071\ 10^3$ | $1.981\ 10^5$ |
| 10 | 3 | $1.597\ 10^7$ | $8.981\ 10^5$ |
| 10 | 4 | $1.517\ 10^7$ | $1.841\ 10^7$ |
| 10 | 5 | $1.321\ 10^5$ | $2.600\ 10^6$ |
| 10 | 6 | $1.467\ 10^6$ | $1.451\ 10^5$ |
| 10 | 7 | $3.179\ 10^7$ | $9.235\ 10^6$ |
| 10 | 8 | $4.896\ 10^7$ | $4.112\ 10^7$ |
| 20 | 1 | $5.807\ 10^7$ | 0. |
| 20 | 2 | $3.229\ 10^8$ | $1.227\ 10^8$ |
| 20 | 3 | $4.017\ 10^7$ | $1.360\ 10^8$ |
| 20 | 4 | $4.954\ 10^7$ | $2.556\ 10^7$ |

Table II — Sample values of normalized coupling coefficients for a wide band of refractive-index fluctuations extending from $r = 0$ to $r = r_{max}$ with $r_{max}/a = 0.8$

| $\nu$ | $n$ | $\bar{H}_{\nu n;\nu+1,n}$ | $\bar{H}_{\nu n;\nu+1,n-1}$ |
|---|---|---|---|
| 1 | 1 | $6.078\ 10^4$ | 0. |
| 1 | 2 | $1.141\ 10^5$ | $1.349\ 10^4$ |
| 1 | 3 | $8.734\ 10^4$ | $6.439\ 10^4$ |
| 1 | 4 | $4.845\ 10^4$ | $8.157\ 10^4$ |
| 1 | 5 | $4.439\ 10^4$ | $5.351\ 10^4$ |
| 1 | 6 | $6.481\ 10^4$ | $3.728\ 10^4$ |
| 1 | 7 | $7.628\ 10^4$ | $4.844\ 10^4$ |
| 1 | 8 | $6.336\ 10^4$ | $6.806\ 10^4$ |
| 1 | 9 | $5.318\ 10^4$ | $6.709\ 10^4$ |
| 1 | 10 | $6.474\ 10^4$ | $5.431\ 10^4$ |
| 1 | 11 | $8.767\ 10^4$ | $5.794\ 10^4$ |
| 1 | 12 | $1.136\ 10^5$ | $8.628\ 10^4$ |
| 10 | 1 | $1.817\ 10^5$ | 0. |
| 10 | 2 | $5.642\ 10^4$ | $8.361\ 10^5$ |
| 10 | 3 | $1.215\ 10^5$ | $2.730\ 10^5$ |
| 10 | 4 | $1.652\ 10^5$ | $3.173\ 10^5$ |
| 10 | 5 | $1.377\ 10^5$ | $4.373\ 10^5$ |
| 10 | 6 | $1.627\ 10^5$ | $4.063\ 10^5$ |
| 10 | 7 | $3.429\ 10^5$ | $4.248\ 10^5$ |
| 10 | 8 | $1.299\ 10^6$ | $9.413\ 10^5$ |
| 20 | 1 | $7.904\ 10^4$ | 0. |
| 20 | 2 | $5.450\ 10^5$ | $1.264\ 10^6$ |
| 20 | 3 | $6.717\ 10^5$ | $4.260\ 10^6$ |
| 20 | 4 | $2.118\ 10^6$ | $7.286\ 10^6$ |

in Table I fluctuate because the narrow band of refractive-index variations may occasionally find itself located near a node of one of the two field functions so that the coupling coefficient can even vanish for a certain pair of modes. For this reason it is advisable to provide at least two bands of the kind (23) at different radii. The coupling coefficients for the second case, listed in Table II, corresponding to a wide band of refractive-index fluctuations, show far less variations. For want of a better procedure we use average values of the coupling coefficients in (32). It may be expected that the low values of the coupling coefficients determine the rate at which power is exchanged among the modes. On the other hand, we know that more than just nearest neighbors couple for low values of $\nu$ and $n$. In order to achieve an order-of-magnitude estimate we use the arithmetic mean of the entries in the first block of data in the tables for $\nu = 1$ and obtain for the average value of the coupling coefficient $h$ from Table I and (25)

$$h = 2 \times 10^6 \left( \frac{\Delta n k r_{max}}{a V^2} \right)^2 \frac{W^2}{a^2} \langle |F|^2 \rangle. \tag{35}$$

Likewise, we find from Table II and (29)

$$h = 7 \times 10^4 \left( \frac{\Delta n k r_{\max}}{a V^2} \right)^2 \langle |F|^2 \rangle. \tag{36}$$

It is clear that (35) and (36) hold only for the special case $V = 40$, $n_1 = 1.515$, and $n_1/n_2 = 1.01$. These values correspond to a fiber with radius $a = 30\ \mu$m if the free-space wavelength is assumed to be $\lambda_0 = 1\ \mu$m.

We are now asking for a pulse width reduction of $R = 0.1$ in a fiber whose length is $L = 1$ km. From (32) and (36) we obtain with $r_{\max} = 0.8a$

$$(\Delta n)^2 \langle |F|^2 \rangle = 4 \times 10^{-7} a. \tag{37}$$

The value obtained from (32) and (35) is identical with (37) if

$$W/a = 0.19. \tag{38}$$

It is apparent from (37) that it is the product of $(\Delta n)^2$ with the amplitude of the spatial power spectrum [of the $z$ dependence of the index fluctuations $f(z)$] that determines the effectiveness of the index fluctuations for reducing pulse dispersion. We relate these quantities to the rms variation of the refractive index as follows. For slight index differences we have

$$n^2 - n_0^2 = 2n_1(n - n_0). \tag{39}$$

Thus we obtain from (17), using $g(r) = 1$ according to (28),

$$\langle (n - n_0)^2 \rangle = \tfrac{1}{2} (\Delta n)^2 \langle f^2(z) \rangle. \tag{40}$$

The average $\langle\ \rangle$ includes in this case also averaging over $\cos^2 \phi$. The variance $\langle f^2 \rangle$ is related to the power spectrum by the equation

$$\langle f^2 \rangle = \frac{1}{\pi} \int_0^\infty \langle |F(\theta)^2| \rangle d\theta = \frac{\theta_{\max}}{\pi} \langle |\bar{F}|^2 \rangle. \tag{41}$$

On the right-hand side of this equation we introduced the average value of the amplitudes of the spatial power spectrum $\langle |\bar{F}^2| \rangle$ and the spatial cutoff frequency $\theta_{\max}$. If we interpret $\langle |F^2| \rangle$ in (37) as the average amplitude appearing in (41) we obtain from (37) through (41)

$$[\langle (n - n_0)^2 \rangle]^{\frac{1}{2}} = 2.52 \times 10^{-4} (\theta_{\max} a)^{\frac{1}{2}}. \tag{42}$$

With $\theta_{\max} a = 0.15$ of Fig. 2 we obtain finally

$$[\langle (n - n_0)^2 \rangle]^{\frac{1}{2}} = 10^{-4}. \tag{43}$$

For our specific example, an rms deviation of this magnitude is required to achieve a relative pulse width improvement of $R = 0.1$ (a ten-times-shorter pulse of coupled modes compared to operating without mode coupling). The spatial Fourier spectrum is assumed to be flat from zero spatial frequencies to a cutoff spatial frequency of

$$\theta_{\max} = \frac{0.15}{a} = 50 \text{ cm}^{-1}. \tag{44}$$

The shortest period appearing in the Fourier spectrum is thus

$$\Lambda = \frac{2\pi}{\theta_{\max}} = 0.13 \text{ cm}. \tag{45}$$

The index fluctuation of the narrow ring defined by (23) is of the same order of magnitude as (43) if the relation (38) holds. However, one narrow ring causes gaps in the coupling process for those modes whose nulls coincide with the position of the ring. It is thus advisable to use at least two rings.

## V. SUGGESTIONS FOR THE DESIGN OF INDEX FLUCTUATIONS

The spatial Fourier spectrum with a sharp cutoff frequency shown in Fig. 1 can be generated by passing noise through a low-pass filter. Another method of producing the desired index fluctuations consists in superimposing a number of sinusoidal variations. From (17), (28), and (39) we have

$$n - n_0 = \Delta n f(z) \cos \phi. \tag{46}$$

If $f(z)$ is a superposition of sine waves with random phase we have

$$n - n_0 = \Delta n \left[ \sum_{\nu=1}^{M} \sin (\Omega_\nu z + \psi_\nu) \right] \cos \phi. \tag{47}$$

The $\psi_\nu$ are random phase functions of the individual sine-wave components. As a practical matter, it is probably easiest to generate these random phases by letting the phase stay constant over a distance $D$ at which point it makes a random jump to another constant value. We assume for the purpose of our analysis that the phases are uncorrelated among each other.

It can be shown that the power spectrum of the function $\sin (\Omega_\nu z + \psi_\nu)$ may be approximated by (see appendix)

$$\langle |F_\nu(\theta)|^2 \rangle = \frac{\sin^2 (\theta - \Omega_\nu) \frac{D}{2}}{(\theta - \Omega_\nu)^2 D}. \tag{48}$$

The total power spectrum of the function $f(z)$ is thus

$$\langle |F(\theta)|^2 \rangle = \sum_{\nu=1}^{M} \frac{\sin^2 (\theta - \Omega_\nu) \frac{D}{2}}{(\theta - \Omega_\nu)^2 D}. \tag{49}$$

$D$ is the correlation length of the phase functions. In the case that the phases stay constant and jump randomly to a new value after a distance $D$, this distance is identical to the correlation length. The full width of the spectrum (48) is given by

$$\Delta\theta = \frac{4\pi}{D}. \tag{50}$$

$\Delta\theta$ is the distance between the first two zeros of the $(\sin x)/x$ function on either side of its main maximum. Since we need to fill a spectral region of width $\theta_{max}$ we need a total number of

$$M = \frac{\theta_{max} D}{4\pi} \tag{51}$$

sinusoidal components. If we use $D/a = 1{,}000$ we need with $\theta_{max} a = 0.15$, $M = 12$ sinusoidal components in (47).

The necessary amplitudes $\Delta n$ of the sinusoidal index variations are obtained from (37). Since the spectral components in (49) overlap only slightly, we may use

$$\langle |F|^2 \rangle = \frac{D}{4} \tag{52}$$

at the peak of each sinusoidal contribution. Using the numerical value of (37) we thus have

$$\Delta n = 1.3 \times 10^{-3} \left( \frac{a}{D} \right)^{\frac{1}{2}}. \tag{53}$$

With $D/a = 1{,}000$ we would thus have $\Delta n = 4 \times 10^{-5}$.

It is clear from our treatment that the numbers given here are only valid to an order of magnitude because of the many approximations that were made for their derivation.

The implementation of this prescription for the desired index fluctuations to the design of a fiber is not a trivial matter.

If the refractive-index increase of the core material is achieved by a doping process, it may be possible to program the doping procedure to result in the desired fluctuations. Some processes add the dopant in the gaseous phase to the material of the fiber preform. In this case it may be possible to control the flow of dopant at a predetermined rate that

is synthesized as the superposition of sine waves shown in (47). If the fiber preform is treated in this way, it would be necessary to compress the periods of the sine waves in such a way that the desired periods result after the drawing process. As mentioned earlier, the desired Fourier spectrum can also be derived from filtered electrical noise signals.

## VI. DISCUSSION

We have studied the possibility of reducing multimode pulse dispersion by means of introducing refractive-index fluctuations into the core of the fiber. Several requirements must be imposed. The coupling process must be able to couple all modes of as large a mode group as possible among each other without coupling to the modes of the continuous spectrum. To achieve this, it is necessary to limit coupling to the modes inside of the area of guided modes in the mode-number plane. Modes near the outer edge of this range should remain uncoupled to avoid radiation losses. This goal can be achieved only by designing the coupling mechanism so that a definite selection rule is imposed. For simplicity we considered index fluctuation with an azimuthal dependence of the form $\cos \phi$. This azimuthal index distribution ensures that a mode with azimuthal mode number $\nu$ couples only to modes with $\nu + 1$ and $\nu - 1$. Once this selection rule is imposed, it is possible to limit coupling to modes inside of an area in mode-number space not including its outer edge. The boundary delineating the area of coupled modes in the mode-number plane tends to cross the boundary enclosing all guided modes at large values of $\nu$. Thus the danger exists that power outflow to radiation modes occurs via modes with large $\nu$ values. This remaining problem can be avoided by limiting the intentional index fluctuations to a region inside the fiber core that remains below a certain radius $r < r_{max} < a$.

The spectrum of spatial frequencies of the $z$-dependent part of the coupling process must extend from (essentially) zero to a cutoff value $\theta_{max}$. A prescription for finding $\theta_{max}$ was given in the section on the spatial Fourier spectrum. In principle, this spectrum can be synthesized as a superposition of sinusoidal components with random phase. This problem is discussed in the section on design suggestions for index fluctuations.

Our procedure leaves one question unanswered: What happens to those modes that remain intentionally uncoupled, don't they broaden the impulse response of the fiber? These modes may be naturally more heavily attenuated than the coupled modes because of their proximity

to the cutoff region in mode-number space. However, if the naturally occurring losses are insufficient to suppress these modes, steps must be taken to filter them out. This filtering is possible either by constricting the fiber at the receiver or by using spatial filtering of the radiation escaping from the end of the fiber prior to entering the detector. A reduction of the fiber diameter reduces the $V$ value of eq. (3). Figure 2 shows a solid line labeled ($V = 30$). In the constricted fiber this line becomes the new mode boundary. It is clear that the constriction removes all the uncoupled modes in addition to some of the coupled ones. An additional amount of loss must be tolerated as payment for the pulse cleaning operation. The total number of guided modes supported by the fiber is given as[1,3]

$$N_t = \frac{V^2}{2}. \tag{54}$$

As the $V$ value is reduced from 40 to 30 the mode number drops from 800 to 450. We lose roughly half the power in the attempt to clean up the pulse distortion that results from the free-running, uncoupled pulses. However, this 3-dB loss penalty is independent of fiber length and may be well worth paying in return for a considerable reduction of pulse dispersion. The conditions of Fig. 3 are far less favorable, here about 75 percent of the power would be lost.

Practical implementation of these ideas will tax the ingenuity of the fiber designer. Introducing desired index fluctuations with the required sharp cutoff of their spatial Fourier spectrum is an exacting requirement. The cutoff frequency $\theta_{max}$ must be kept to within a few percent. This means that the spectrum must terminate with a sharp slope. If the spectrum is synthesized as a superposition of sine waves we must require that

$$\frac{\Delta\theta}{\theta_{max}} = \frac{4\pi}{D\theta_{max}} = \frac{1}{M} < 0.01. \tag{55}$$

The number 0.01 was chosen arbitrarily, but it is certainly of the correct order of magnitude, $\Delta\theta$ was obtained from (50) and $M$ from (51). This estimate shows that approximately 100 sinusoidal components are required to ensure that the spectrum has a sufficiently steep slope at its cutoff point. Of course, this requirement could be somewhat relaxed by allowing the sinusoidal components deeper inside the spectrum to be broader than those near its edge. These considerations may enter into the compromise that the designer wishes to achieve.

The mode spacing (in $\beta$ space) changes as the fiber is bent. The modes of a dielectric slab waveguide have smaller mode spacing in

circularly bent sections of the guide. In principle, this effect could alter our conclusions regarding the possibility of uncoupling high-order modes near cutoff by limiting the width of the spatial Fourier spectrum of the coupling function. However, for the numerical example used in our discussions, the change in mode spacing due to waveguide curvature is unimportant for bent slabs whose radii remain about the limit

$$R_c > 1{,}000a.$$

For $a = 30\ \mu\text{m}$, $R_c$ would have to be kept larger than 3 cm. It is not known if the numbers for the round fiber would be identical to the slab model, but one may expect that waveguide curvature is unimportant if the radius of curvature remains larger than a few centimeters.

## APPENDIX

We provide a simple derivation of eq. (28). Let

$$f(z) = \sin [\Omega z + \psi(z)]. \tag{56}$$

The phase function $\psi(z)$ is assumed to be constant over a distance $D$ but jumps randomly to a new constant value at the end of each interval. The Fourier transform of $f(z)$ is

$$\begin{aligned} F(\theta) &= \frac{1}{\sqrt{L}} \sum_{n=0}^{m-1} \int_{nD}^{(n+1)D} \sin [\Omega z + \psi_n] e^{-i\theta z} dz \\ &= \frac{1}{\sqrt{L}} \sum_{n=0}^{m-1} e^{i\psi_n} e^{i(\Omega-\theta)(n+\frac{1}{2})D} \frac{\sin\left[(\Omega - \theta)\frac{D}{2}\right]}{i(\Omega - \theta)}. \end{aligned} \tag{57}$$

A term with $\Omega + \theta$ has been neglected since its contribution for large values of $D$ is negligible. It is also assumed that $L$ is an integral multiple of $D$,

$$L = mD. \tag{58}$$

When we form the absolute-square value of $F$ and take the ensemble average, all cross terms in the resulting double summation vanish because of our assumption of random values for $\psi_n$. The phase terms in the remaining terms of the summation reduce to unity by the process of taking the absolute-square value. With no summation index left under the summation sign, the sum results simply in a factor $m$. Using (58) we finally obtain

$$\langle |F(\theta)|^2 \rangle = \frac{\sin^2\left[(\theta - \Omega)\frac{D}{2}\right]}{(\theta - \Omega)^2 D}. \tag{59}$$

## REFERENCES


1. D. Gloge, "Weakly Guiding Fibers," Appl. Opt., *10*, No. 10 (October 1971), pp. 2252–2258.
2. D. Gloge, "Dispersion in Weakly Guiding Fibers," Appl. Opt., *10*, No. 11 (November 1971), pp. 2442–2445.
3. D. Marcuse, *Theory of Dielectric Optical Waveguides*, New York: Academic Press, 1974.
4. S. D. Personick, "Time Dispersion in Dielectric Waveguides," B.S.T.J., *50*, No. 3 (March 1971), pp. 843–859.
5. D. Marcuse, "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1199–1232.
6. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand-Reinhold Co., 1972.
7. M. Abromovitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series, *55*, National Bureau of Standards, Washington, D. C., 1965.
8. Ref. 3, eq. (4.7-5), p. 168.
9. Ref. 3, Section 3.6.
10. Ref. 3, eq. (5.5-25), p. 206.
11. Ref. 3, eq. (5.5-45) and (5.5-55), pp. 209–210.
12. D. Gloge, "Optical Power Flow in Multimode Fibers," B.S.T.J., *51*, No. 8 (October 1972), pp. 1767–1783.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 9, November 1974
*Printed in U.S.A.*

# Outage of the L4 System and the Geomagnetic Disturbances of 4 August 1972

By C. W. ANDERSON III, L. J. LANZEROTTI, and C. G. MACLENNAN

(Manuscript received April 24, 1974)

*An outage of the Plano, Illinois, to Cascade, Iowa, link of the L4 coaxial cable occurred at about 2240 UT on 4 August 1972 during a large geomagnetic storm. The available geomagnetic data measured in North America, as well as data received from two satellite instruments, are analyzed. These data show that, at the time of the L4 outage, the boundary of the magnetosphere was pushed to unusually low altitudes by a greatly enhanced solar wind. As a result, large, rapid changes of the earth's magnetic field strength were observed over North America. It is demonstrated that the field changes at about 2241 to 2242 UT were of such magnitude as to induce earth currents of sufficient strength to produce the L4 outage by causing a high-current shutdown of the system link. The geomagnetic disturbances that produced the shutdown were not of the auroral-electrojet type normally associated with disruptions of power systems.*

## I. INTRODUCTION

The solar and geomagnetic disturbances resulting from solar active region 11976 were truly outstanding in many regards. The principal solar region of the activity was in the highest solar activity class on an absolute scale; this major solar activity occurred during the declining phase of the 11-year solar cycle.[1] The solar disturbances, propagating outward into interplanetary space, produced the largest galactic cosmic ray decrease on record.[2] The geomagnetic storms (with accompanying ionospheric and auroral disturbances) resulting from the interaction of the propagating solar disturbances with the earth's magnetosphere were the most severe recorded in well over a decade.

The interaction of the greatly enhanced solar wind[3] with the earth's magnetic field on 4 August 1972 produced extreme compressions and distortions of the magnetosphere. It was during the period of the most

severe magnetospheric distortion that an outage of the L4 coaxial cable carrier system occurred over the link from Plano, Illinois, to Cascade, Iowa. This paper, utilizing most of the available North American geomagnetic data, describes the geophysical occurrences and conditions at the time of the L4 outage. These data, plus magnetic field data from two satellites, demonstrate that the L4 outage was associated with the extreme compressions and distortions of the magnetosphere and not simply with greatly enhanced auroral currents, as are often assumed in discussions and models of magnetic storm-induced power-system disruptions.[4] The North American geomagnetic data are used as input to a model for the calculation of currents induced in the earth by the geomagnetic disturbances. It is shown that these earth currents were sufficient to produce a shutdown of the L4 system.

## II. GEOMAGNETIC OCCURRENCES

The subsequent discussions of the available geomagnetic data during the August 1972 storm period must be placed in the perspective of normal magnetospheric conditions. Figure 1 is a view of the earth's magnetosphere in the equatorial plane. The distance from the earth's surface to the magnetospheric boundary is approximately $9R_E$ ($1R_E \approx 6.5 \times 10^3$ km) for normal solar wind density conditions of $\approx 5$ protons $cc^{-1}$. Figure 1 also shows the circular, earth-synchronous satellite orbit at an altitude of $\approx 5.5R_E$, which is occupied by all common-carrier communications satellites. For later reference, the location of the NASA synchronous altitude Applications Technology Satellite 5 (ATS 5) at 2200 UT is indicated. Also indicated, for the same UT, is the apogee location of the near-equatorial, elliptical-orbit satellite Explorer 45.[5]

The approximate geomagnetic latitude and longitude of Plano, Illinois (50.8°N, 336.8°E) are shown on the earth in Fig. 1 for 2200 UT. Also indicated on the earth are the locations of several North American magnetic observatories: Meanook (61.81°N, 301.07°E), Churchill (68.69°N, 322.63°E), and Ottawa (56.80°N, 351.52°E). All geomagnetic stations used in this study are listed in Table I, together with their coordinates.

Acquisition of sufficient geomagnetic data for a definitive analysis of the geomagnetic effects of the storm in North America was made difficult by the inability of most magnetic observatory instruments to record the very large and rapid variations that occurred during the hour 22 UT on 4 August. Because of the instruments' limitations, it is likely that many of the actual variations were larger than those

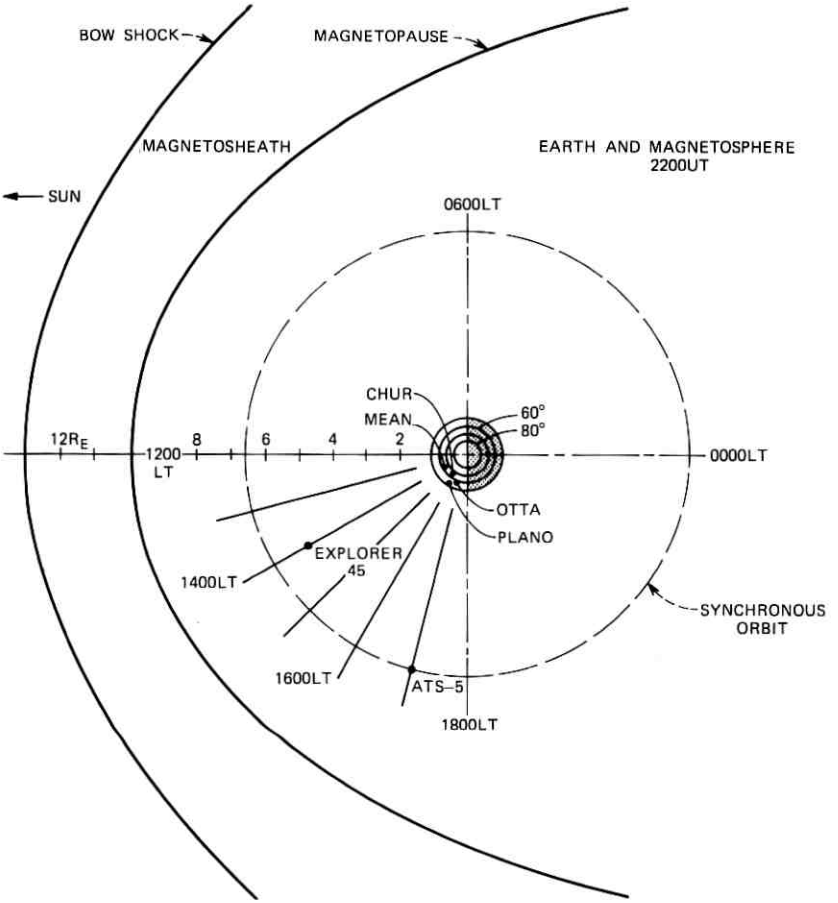



Fig. 1—View of earth and magnetospheric configuration in equatorial plane under normal solar wind flow conditions.

used in this paper.[6] Indeed, for the analysis of the 20-minute interval (2230 to 2250 UT, 4 August), around the time of the L4 outage, data at one-minute time intervals could be scaled from the continental U. S. and Alaska standard observatory chart records for only the observatories at College, Sitka, Tucson, and Fredericksburg. Scalings of $2\frac{1}{2}$ minutes were obtained from observatories at Castle Rock and Dallas. Scalings of 1 minute were obtained from a special National Oceanic and Atmospheric Administration station near Boulder. Fortunately, the Earth Physics Branch of the Department of Energy, Mines,

Table I

| Geomagnetic Station | Geomagnetic Coordinates | |
|---|---|---|
| | Lat(°N) | Long(°E) |
| Baker Lake | 73.74 | 315.31 |
| Boulder | 48.85 | 316.44 |
| Cambridge Bay | 77.7 | 300.3 |
| Castle Rock | 43.48 | 298.62 |
| College | 64.66 | 256.51 |
| Dallas | 42.96 | 327.75 |
| Fredericksburg | 49.55 | 349.84 |
| Fort Churchill | 68.69 | 322.63 |
| Meanook | 61.81 | 301.07 |
| Ottawa | 56.80 | 351.52 |
| Sitka | 60.00 | 275.34 |
| St. Johns | 58.50 | 21.24 |
| Tucson | 40.03 | 311.41 |
| Victoria | 54.08 | 293.04 |

and Resources in Ottawa was operating a number of digital-recording magnetometers in Canada during 1972.[6] These data have greatly facilitated our analysis of the magnetic disturbances. Most of the available U. S. magnetometer data were obtained from the World Data Center A and were scaled for us by the Geophysical Institute of the University of Alaska.

Plotted in Fig. 2 are the magnetic variations measured in the north-south ($H$ or $X$) and east-west ($D$ or $Y$) orthogonal directions at the three Canadian observatories, Meanook, Churchill, and Ottawa, during the time interval 2200 to 2300 UT on 4 August. At the bottom of both sets of data is plotted the magnetic field trace of the $Z$-component recorded by a magnetometer on the ATS 5 satellite.[7] The $Z$-component of the magnetic field at the ATS 5 location is measured parallel to the earth's spin axis.

The data plotted in Fig. 2 show particularly large field changes occurring during the several-minute time interval following 2240 UT. For example, at about 2242 UT, the field change $\Delta H/\Delta t$ measured at Meanook was $\approx 1800\ \gamma$/min. ($1\gamma = 10^{-5}$ gauss). The field data from the ATS 5 satellite are particularly interesting in that, at $\approx 2225$ UT and $\approx 2240$ UT, the field direction reversed. At $\approx 2241$ UT, the field intensity was $\approx -400\gamma$; i.e., the measured field pointed in a direction opposite to that of the normal dipole field. At $\approx 2242$ UT, the field direction at ATS 5 suddenly became normal; at this time, $\Delta Z/\Delta t$ at ATS 5 was $\approx 575\ \gamma$/min.

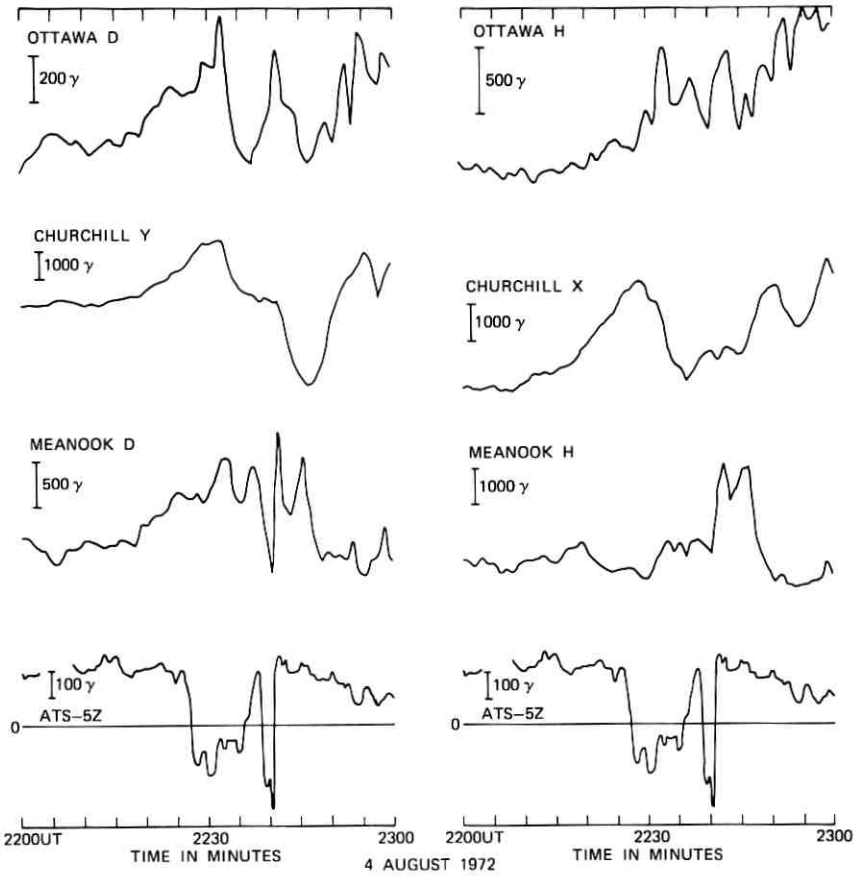



Fig. 2—Magnetic field variations observed at three Canadian observatories (Meanook, Churchill, and Ottawa) in north-south and east-west directions during the hour 2200 to 2300 UT on 4 August 1972. Plotted beneath each set of data are the variations in the field observed near the equator on the ATS-5 spacecraft.

Reversals of the earth's field direction, similar to those shown in Fig. 2, have been observed in the past by magnetometers on the ATS 1[8] and ATS 5[9] satellites. These occurrences have been attributed to a movement of the magnetopause to a location *inside* the orbit of the synchronous satellite, which then measures the magnetic fields in the earth's magnetosheath region. The complex particle and field changes that occur in the magnetosphere during such a "boundary-crossing" event have been discussed in a series of papers devoted to extensive study of one such event.[8,10–13]

The distortions of the magnetospheric boundary as evidenced by the boundary-crossing event observed on ATS 5 on 4 August 1972 were accompanied by the large geomagnetic field changes observed on the ground in North America. It is interesting to note that the study of several boundary crossing events on ATS 5[9] suggested that the largest effects were observed on the ground when the magnetosphere distortion occurred in the local afternoon side of the magnetosphere rather than in the local morning side of the magnetosphere. For the 4 August event, no local morning data are available from a synchronous satellite. However, the magnetosphere distortions that were observed were in the local afternoon sector.

The one-minute values of the magnetic field changes [$H(X)$ and $D(Y)$ components] that could be determined from North American observatory data were used to make contour maps of the field changes at one-minute intervals in the U. S. and Canada. These contours for 2238 to 2243 UT are plotted in Figs. 3a and 3b in a geomagnetic coordinate system. Because of the spread in distance between observatories and the impossibility of obtaining a reliable magnetic field reading from the chart records of several observatories, the locations of some contour lines in Figs. 3a and 3b are, at best, extrapolations. For lack of a better justified procedure, the contours have been constructed on the basis of linear interpolation between observatory field values. It should be noted that, for more conventional magnetic disturbances arising from auroral electrojet current systems, the fall-off in magnetic disturbances from higher to lower latitudes is nonlinear; i.e., the disturbance level falls off more rapidly at the lower latitudes. The changes in the location and intensity of the magnetic disturbances from the relatively undisturbed period at $\approx$ 2238 UT until the large disturbance at Meanook at $\approx$ 2242 UT is quite evident in the contours of Figs. 3a and 3b.

There is evidence in the contours of Figs. 3a and 3b of perhaps some progression of the geomagnetic disturbance from the higher to the lower latitudes. For example, in the interval 2240 to 2241 UT the disturbance change is largest at College, while in the interval 2241 to 2242 UT the disturbance change becomes largest at Meanook. If this progression of the disturbance is associated with a distortion and compression of the magnetospheric boundary to smaller radii (and therefore lower latitudes), the contours suggest that the boundary compression may have reached field lines that intersect with latitudes as low as that of Meanook. That this was apparently the case is discussed below.

Plotted in Fig. 4 as contours (linear interpolations) on a geographical map of North America are the magnitudes of the total horizontal field changes and the angles [from the $D(Y)$ direction] of the changes measured between 2241 and 2242 UT, at about the time of the sudden changes in the apparent ATS 5 location from the magnetosheath to the magnetosphere (see Fig. 1). At this time, the field intensity at Plano is interpolated to be $\approx 700\gamma$ and the field change is aligned in an approximately NE–SW direction. These are the important geomagnetic parameters that will be used in the next section to calculate the expected induced earth currents at Plano.

Before calculating the earth currents however, it is of interest to further examine the magnetospheric environment at $\approx$ 2241 UT. At about this time, the Explorer 45 satellite was located at its apogee position and was about two hours closer to local noon than ATS 5 (see Fig. 1). Magnetic field data and charged-particle data from Explorer 45 indicate that, at $\approx$ 2242 UT, when the magnetosphere boundary expanded outward beyond ATS 5, a movement of the boundary inward, inside Explorer 45, was recorded.[14] Inspection of the magnetograms from College, near local noon, at $\approx$ 2241 UT indicate that the field increased in magnitude and that the field changes were predominantly in the $H$ direction (see Fig. 2). This observation also suggests a compression of the magnetosphere boundary near local noon.

Hence, summarizing and synthesizing as well as possible the available magnetospheric data, the boundary of the magnetosphere in the local afternoon sector at $\approx$ 2240 and $\approx$ 2242 UT could be pictured as in Fig. 5. Not only is the boundary greatly compressed from the normal location (see Fig. 1), but it is also highly distorted as evidenced by the simultaneous observations of the magnetopause at altitudes higher than ATS 5 but lower than Explorer 45 at $\approx$ 2242 UT. This observation by Explorer 45 provides evidence of the most compressed position of the magnetopause yet recorded. The extreme compression and distortion of the magnetosphere recorded in the sector above the western hemisphere at $\approx$ 2242 UT was undoubtedly the primary cause for the large magnetic field changes recorded in North America at the time.

### III. TELLURIC DISTURBANCES

To put into perspective the analysis used to calculate the induced earth currents at Plano, it is of interest to review the basic relationships between earth resistivity structure, geomagnetic disturbances,

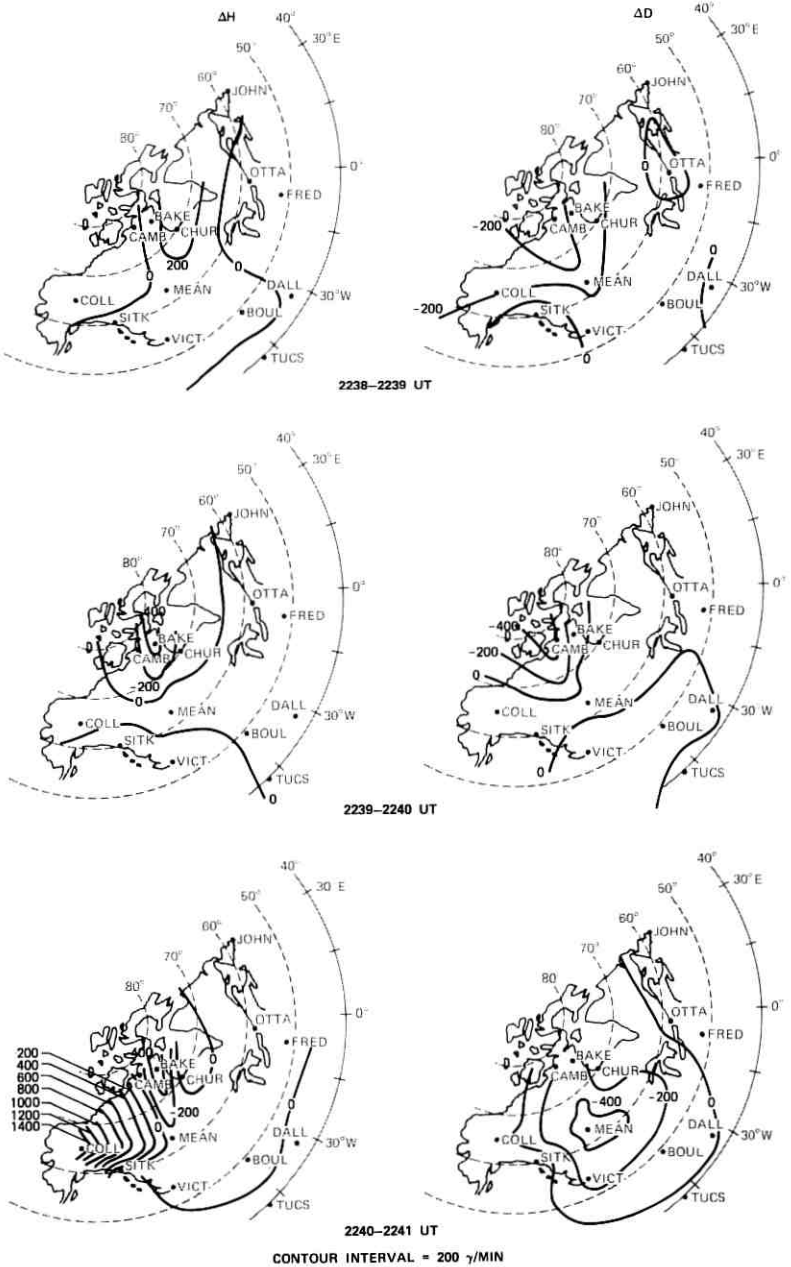



Fig. 3a—Minute-by-minute rate of change of magnetic field in the horizontal plane over North America from 2238 to 2241 UT.

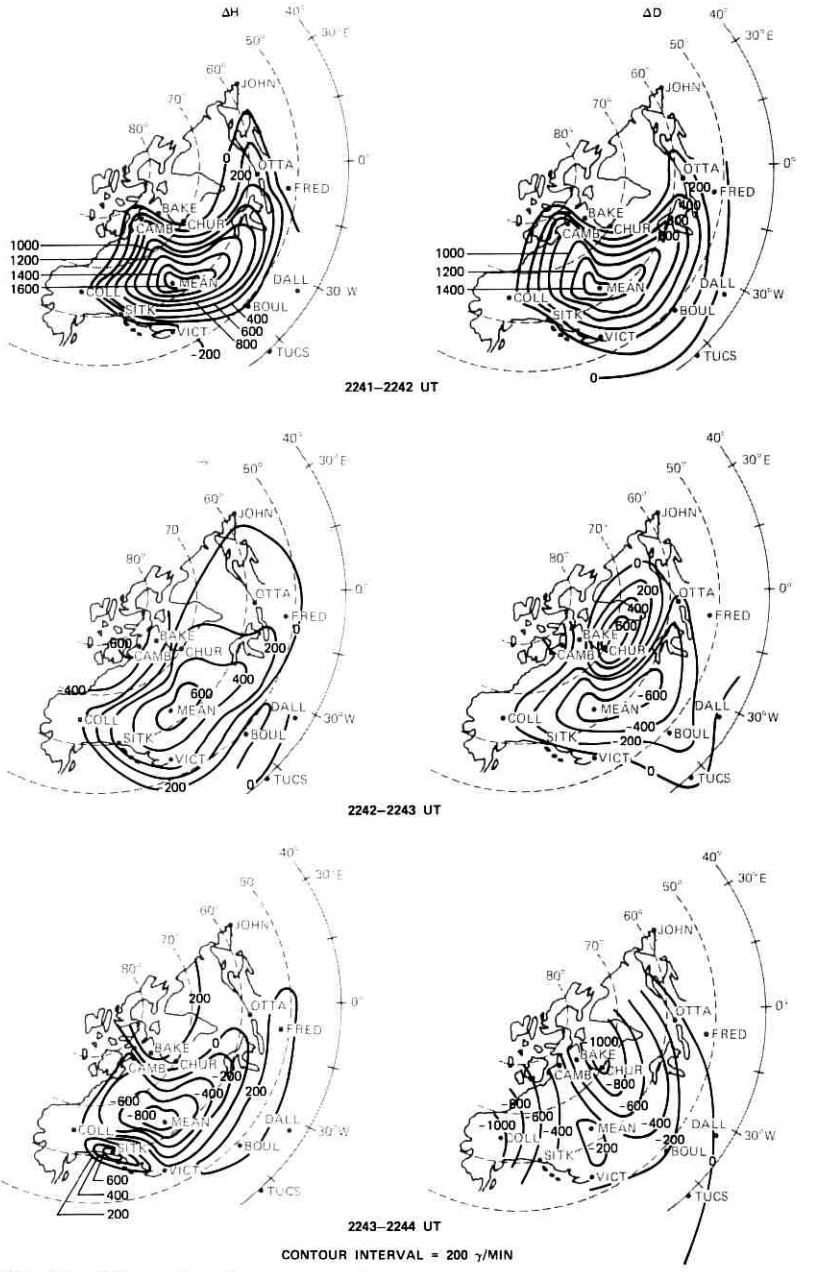



Fig. 3b—Minute-by-minute rate of change of magnetic field in the horizontal plane over North America from 2241 to 2244 UT.

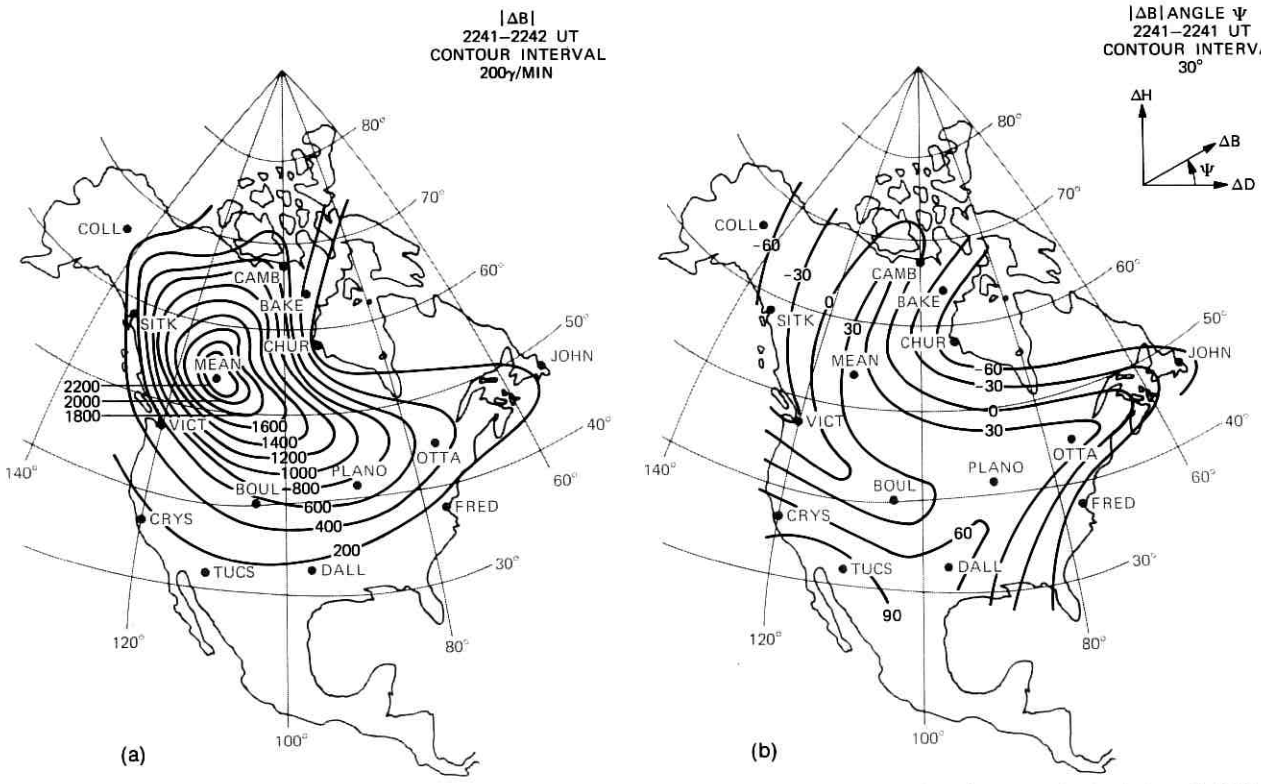



Fig. 4—Rate of change of magnetic field intensity and direction over North America for the one-minute interval 2241 to 2242 UT.

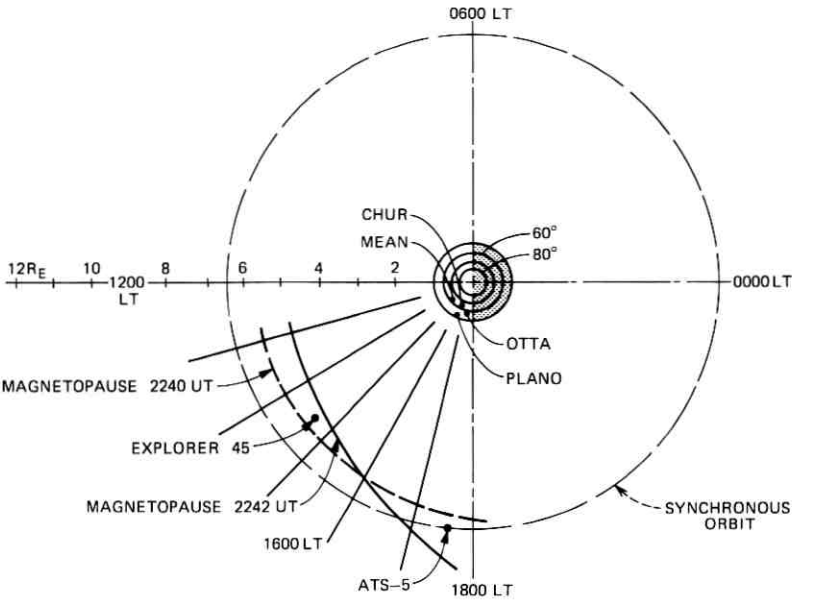



Fig. 5—Equatorial plane view of earth and magnetospheric boundary in afternoon sector at 2240 UT and 2242 UT.

and telluric disturbances; i.e., induced electric fields at the earth's surface.

The basic theory of tellurics is contained in a boundary value problem involving Maxwell's equations and the resultant electromagnetic wave equation. An external exciting source is assumed. The phase and amplitude relationships between the orthogonal components of the horizontal electric and magnetic fields observed at the surface of the earth are measures of the electrical properties of the earth.

Electromagnetic induction by a uniform horizontal magnetic field $B$ [with components $H$(NS) and $D$(EW)] in a uniform semi-infinite earth produces orthogonal horizontal electric fields, $E$, which satisfy the following relationships:[15]

$$\text{mod } (E/B) \propto 1/T^{\frac{1}{2}} \tag{1a}$$

$$\arg (E/B) = \pi/4, \tag{1b}$$

where $T$ is the period of the magnetic field $B$. Hence, the ratio of the amplitude of the electric field to the magnetic field is proportional to the inverse square root of the period $T$ and the phase difference is always $\pi/4$.

It is necessary to measure three mutually orthogonal components in order to completely describe the vector magnetic field. It is customary in land-based magnetic variation surveys at midlatitudes to measure the horizontal $H$ (geomagnetic NS) varying, horizontal $D$ (geomagnetic EW) varying, and vertical $Z$ components of the magnetic field. The sense of the variations is positive north for $H$, positive east for $D$, and positive down for $Z$. The fields are referenced from the earth's surface and magnetic north. The induction process, therefore, is such that a positive $H$ variation will produce an east-to-west flowing telluric current (a negative surface electric field). [At high latitudes it is customary to use the geographic NS ($X$) and geographic EW ($Y$) components.]

Telluric current observations, together with the magnetic field measurements, usually show that the amplitude ratio $E/B$ is proportional to the inverse square root of the period. Thus, the relationship in eq. (1a) suggests the use of a uniform conductivity earth model. However, observations of the phase differences between $E$ and $B$ seldom agree with the predictions of the uniform earth model (1b). Phase differences as small as 20° for magnetic field variations with periods of 20 to 30 minutes and as large as 45° for much shorter periods have been reported.[15] These large discrepancies in phase differences between the theory and the observations underscore the fact that a uniform earth model is seldom applicable to the real earth at an observing location.

Wait[16] has developed an analytical formulation of the electromagnetic fields of an infinite line source above a horizontally stratified earth. The earth model can be extended to include any number of layers. The line source can serve as an equivalent current system to the real current systems (e.g., ionospheric or magnetospheric current systems), which can produce natural electromagnetic fields at the earth's surface. The basic relationship between the electric and magnetic fields at the surface of the earth is given as[15]

$$\frac{-\mu_o E_y}{H} = Z_1, \tag{2}$$

where $\mu_o$ is the free space permeability and $Z_1$ is defined as the surface impedance. For a three-layer earth model, the surface impedance is given as[16]

$$Z_1 = \frac{i\mu_o\omega}{u_1} Q, \tag{3}$$

where $u_1$ is the propagation constant for the first layer, $\mu_o$ is the free space permeability, $\omega$ is the angular frequency, and $Q$ is a correction term that accounts for the resistivities and thicknesses of the three layers. The mks system is used throughout.

The surface impedance defines the relationships between the tangential electric and magnetic fields at the earth's surface. The surface impedance also completely represents the electrical properties of the earth, when displacement currents are neglected. Estimates of the surface impedance can be made from knowledge of the geology in a specific area.

The apparent resistivity of the three-layer earth model is related to the surface impedance by the expression[14]

$$\rho_a(\omega) = \frac{1}{\omega\mu_o}|Z_1(\omega)|^2. \tag{4}$$

The apparent resistivity $\rho_a$ (in ohm-meters) is highly dependent on the angular frequency $\omega$ of the source and the conductivity structure of the three-layer earth. It is important to note that many published values of earth resistivity are, in fact, apparent resistivities that are valid only in a limited frequency range. For example, the use of earth resistivity values determined at 60 Hz will not normally be valid at the frequencies at which telluric currents are important (see Fig. 7).

The numerical evaluation of eq. (2) is substantially simplified if the electromagnetic fields are considered plane waves. Since no physical sources exist in nature that produce plane waves, it is necessary to determine the range of frequencies and geographic conditions in which the fields of a line source can be successfully approximated by mathematical plane waves.

Peeples[17] has shown that the plane wave approximation will be valid for wave periods less than 500 seconds for earth models that have a highly conducting sedimentary first layer, if the line current is located at a height of at least 100 km. Morrison[18] has shown that, as the resistivity of the first layer increases, the period range decreases for which a plane wave analysis is valid. The plane wave approximation has been found valid in the frequency range ($\approx$0.01 to 0.1 Hz) in which telluric current effects are important to long-haul communication systems. Hence, the surface impedance and apparent resistivity for a plane wave can be readily computed from eqs. (3) and (4) once the parameters of the three-layer earth are established. The surface electric field can be computed for any value of the magnetic induction field by inserting the appropriate surface impedance value into eq. (2).

## IV. PLANO EARTH RESISTIVITY MODEL

In this section, a three-layer earth resistivity model is developed. The model is derived from the modified Cantwell-McDonald earth resistivity model[19] and detailed information of the geology[20] in the vicinity of Plano, Illinois. This model is assumed to be representative of the entire L4 route from Plano, Illinois, to Cascade, Iowa.

Geophysical studies in the vicinity of Plano indicate that the granitic basement rocks are at a depth of about 1.2 km. Above the basement lies a relatively thick sequence of flat-lying sedimentary rocks consisting of Cambrian and Ordovician sandstones, shales, and dolomites. Unconsolidated glacial deposits overlie the bedrock with thicknesses of 18 to 36 meters.

The three-layer earth model in Fig. 6 was arrived at by merging the upper crustal model with the modified Cantwell-McDonald earth resistivity model. Because of the relatively low frequencies encountered in telluric studies, it was felt that the layer of unconsolidated glacial deposits could be ignored without affecting the results.

The top layer of sedimentary rocks was given a resistivity of 100 Ωm and a thickness of 1.22 km; the second layer of granitic basement rock was given a resistivity of 5000 Ωm and a thickness of 300 km; the bottom layer, representing the upper mantle, was given a resistivity of 10 Ωm and an infinite thickness. The surface impedance was computed using the earth parameters from Fig. 6 in eq. (3) after modification of the latter for the plane wave approximation.

The more familiar apparent resistivity, which is related to the surface impedance through eq. (4), was determined to show the frequency

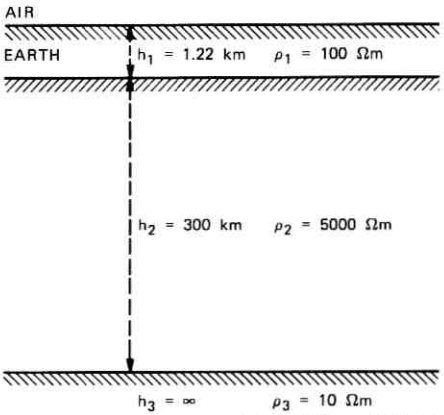



Fig. 6—Three-layer resistivity model for Plano, Illinois, vicinity.

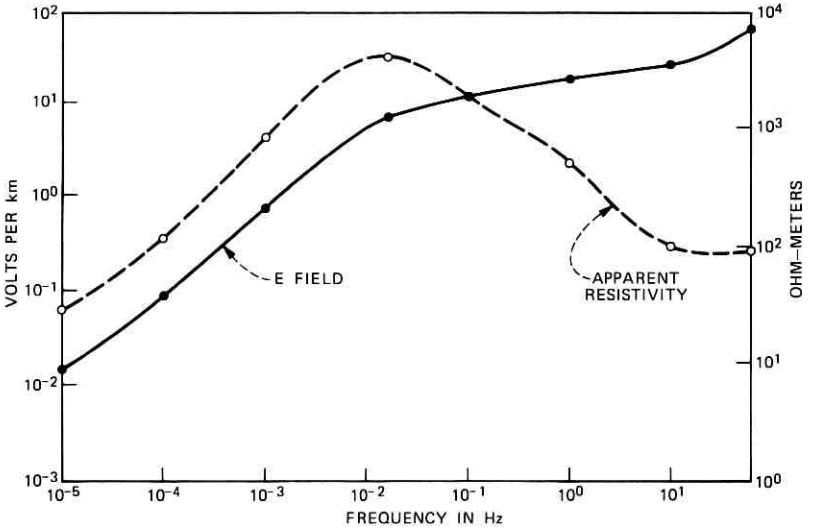



Fig. 7—Variation with frequency of apparent resistivity and surface electric field for the three-layer model of Fig. 6.

response of the earth model. Figure 7 illustrates the variation in apparent resistivity as a function of frequency. The decrease in the apparent resistivity at the higher frequencies arises from the influence of the top layer. A maximum apparent resistivity is shown at about $10^{-2}$ Hz; this reflects the influence of the highly resistive second layer, the basement rock. The drop in apparent resistivity below $\approx 10^{-2}$ Hz arises from the influence of the less resistive upper mantle as the frequency decreases and the depth of penetration of the magnetic field increases. Hence, in this layered earth model, there are certain frequencies for which the apparent resistivities are much higher than others. Earth resistivity models in general are nonunique, and extrapolations from them should be done with caution.

The surface impedance based on the above earth model was used to compute the variation in surface electric field as a function of the source frequency. Figure 7 shows the variation in surface electric field when the orthogonal horizontal magnetic induction field is 700 gammas. The induced electric field is highly dependent on frequency. The flattening of the induced electric field at about $10^{-2}$ Hz corresponds to the maximum in the apparent resistivity at this frequency. A uniform earth model would not show this flattening, but only a continuous decrease in electric field magnitude with decreasing frequency.

## V. THE L4 SYSTEM AND EARTH POTENTIAL OUTAGES

The L4 system consists of coaxial cables powered in pairs by power-feed stations. Normally, the maximum distance between power-feed stations is 242 km (150 miles), with a nominal repeater spacing of 2 miles. The power system is grounded at one end, and the output voltages of the four dc-to-dc converters (each rated to deliver 1800 V and a nominal line current of 520 mA) are balanced so that the voltage to ground at the "floating ground" end is zero. Figure 8 shows a typical L4 system powering section in the presence of an earth potential. In the presence of slowly varying voltages, the floating ground has a threshold of 370 V, above which it becomes automatically grounded. The automatic grounding feature serves as protection to the system. The floating ground must be restored to its normal condition manually.

Direct-current earth potentials produce changes in the L4 system line current that can cause transmission impairment and/or converter shutdown. The line current is unaffected by earth potentials until the 370 V threshold is exceeded. After grounding occurs at the floating point end, the earth potential appears in series with the metallic power-feed loops of both lines. The earth potential will increase the voltage on the line of the same polarity, causing an increase in that line current. The potential will decrease the voltage on the line with the opposing polarity, causing a decrease in that line current. Since earth potentials produced by geomagnetic variations frequently reverse polarity on rather short time scales, both lines will experience high and low currents. The magnitudes of the line currents are dependent on the following factors:

(*i*) Magnitudes of the earth potentials.
(*ii*) Degree of balance of the power system.
(*iii*) Dynamic resistance of the line and the converters.

System designers have analyzed an ideal model of the L4 system to estimate the line current variations as a function of earth potential. The results of this analysis with respect to converter shutdown on a standard 242-km power section are:[21,22]

(*i*) At $\approx 6.5$ V/km, the line with the aiding earth potential will experience a high current shutdown.
(*ii*) At $\approx 8.9$ V/km, the line with the opposing earth potential will experience a low current shutdown of both converters feeding the line (see Fig. 8).

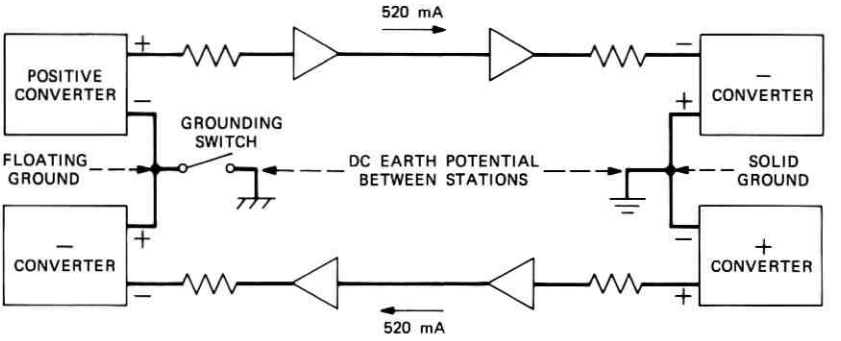



Fig. 8—L4 power-feed section in the presence of an earth potential.

The high current shutdown at $\approx 6.5$ V/km is considered the most serious effect of earth potentials and, as is discussed below, is most likely the cause of the Plano-Cascade L4 shutdown of 4 August 1972.

The surface electric field induced along the Bell System L4 route from Plano to Cascade at approximately 2242 UT, 4 August 1972, can be determined from eq. (2). The magnetic field variation $|\Delta B/\Delta t|$ can be determined from Fig. 4 and the surface impedance can be determined from eq. (3) using the Plano earth resistivity model. The estimated geomagnetic disturbance of $\approx 700$ $\gamma$/min (see Fig. 4) will induce a perpendicular surface electric field of $\approx 7.4$ V/km (see Fig. 7).

The estimated direction of the geomagnetic disturbance vector was about 40° north of east (see Fig. 4), making an angle of approximately 70° with the Plano L4 route. For this angle, about 94 percent of the induced electric field, i.e., $\approx 7$ V/km, would directly affect the L4 power-feed section. Bell Laboratories engineers have established a working shutdown value range for earth potentials of 6.5 V/km $\pm 20$ percent.[22] Comparison of this range to the $\approx 7$ V/km value deduced for 4 August from magnetic field fluctuations shows that the surface electric fields that were most probably induced along the Plano to Cascade route at $\approx$ 2242 UT 4 August 1972 were probably sufficient to cause the L4 shutdown that took place.

### *The Uniqueness of the Plano L4 Outage*

A complete explanation of why the Plano-Cascade section of the transcontinental L4 route experienced a shutdown and why other sections did not is not available at the present time. To establish some

argument for uniqueness, it is worthwhile to outline some factors that might contribute to a telluric current shutdown. The major factors can be divided into geophysical conditions and system susceptibility.

Previous sections of this paper have dealt with the importance of the rate of change of the inducing field $|\Delta B/\Delta t|$ and of the surface impedance in determining the induced telluric currents. To achieve a maximum telluric current or earth surface potential along an L4 route, the horizontal magnetic field variation must be perpendicular to the direction of the route.

The contour maps shown in Fig. 4 illustrating the area distribution of $|\Delta B/\Delta t|$ and its direction angle in the interval 2241 to 2242 UT across North America are a valuable tool for determining the magnitude of the geomagnetic disturbances that actually affect an L4 power-feed section. As mentioned above, an estimated 94 percent of $|\Delta B/\Delta t|$ at 2242 UT affected the Plano-to-Cascade power-feed section. The L4 power sections immediately to the east of Plano and immediately to the west of Cascade both have a more east-west direction than the Plano-to-Cascade section. This means that only an estimated 60 percent of $|\Delta B/\Delta t|$, producing an electric field of $\approx 4.7$ V/km, will affect these sections. This estimated electric field value is below the lower-limit system shutdown value of 5.2 V/km. These comments must be qualified by the fact that, as noted earlier, the available magnetometer data were certainly not optimum to determine field changes and directions with high accuracy.

The frequency response of the earth's surface impedance along the cable route is also a very important factor in producing maximum telluric currents. As shown in Fig. 7, the frequency response of the Plano earth model was a maximum at $\approx 10^{-2}$ Hz, which is the frequency of the largest magnetic variations at 2242 UT. Analytical results from various earth resistivity models have shown the frequency response of the surface impedance to be fairly sensitive to changes in layer thickness.[21] If the earth resistivity structure profile were known along the entire transcontinental L4 route, it would aid in the determination of which sections of the route are most susceptible to telluric currents. This information is not presently available, and probably will not be for quite some time. Thus, this discussion indicates that a combination of geophysical factors is needed to produce maximum telluric currents along a specified cable route. All these factors are very difficult to estimate at any particular time from the distribution of geomagnetic and telluric observatories presently existing in North America.

From the point of view of system susceptibility, each power-feed section of the L4 route will most likely have a different earth potential shutdown value. This condition exists because of variations in the power system balance, because of variations of dynamic resistance in the individual lines and convertors, and because of the length of any particular power-feed section. For instance, the Plano-to-Cascade L4 power-feed section, at $\approx 248$ km, is the longest in the Bell System. The L4 sections immediately to the east and to the west are $\approx 213$ and $\approx 230$ km long, respectively. The decreased lengths in these sections increase the critical shutdown voltages to $\approx 7.4$ and $\approx 6.8$ V/km, respectively.

## VI. SUMMARY

The preceding discussions, using most of the existing North American data available for the time period around the L4 system outage of 4 August 1972, have indicated that the geomagnetic disturbances were apparently sufficient to produce the outage on the Plano, Illinois, to Cascade, Iowa, route. The geomagnetic disturbances were produced primarily by large distortions of the earth's magnetosphere, whose boundary was observed to be pushed inward to an altitude of about four to five earth radii over North America at the time of the L4 outage. Therefore, the geomagnetic disturbances that produced the outage did not arise from enhanced auroral current systems that are normally associated with power system problems during magnetic storms.

It is difficult to establish the uniqueness of the L4 problem along the Plano-Cascade route during this large storm. A large part of this problem arises because of the insufficiency of the geophysical data, data concerning the magnitudes and spatial distributions of the magnetic fields and earth currents during the magnetic storm as well as data on the geological structure under the L4 route.

In addition to problems of coaxial system outages from large geomagnetic disturbances, it is likely that smaller magnetic storms may occasionally induce sufficient currents on some routes in certain locales to produce transmission impairments. Experimental work to study this problem is presently under way.

## VII. ACKNOWLEDGMENTS

We thank our colleagues in the geophysical community for their discussions and help in supplying data for this study. In particular,

we acknowledge profitable discussions with H. Fukunishi, D. J. Williams, and L. J. Cahill, and the assistance of J. N. Barfield, G. Jeans, and W. Paulishak of NOAA, Boulder, Colorado, in obtaining magnetograms. D. B. Ledley of NASA/Goddard kindly supplied the ATS-5 data. We thank S.-I. Akasofu and M. Buhler of the Geophysical Institute, University of Alaska, for digitizing the U. S. magnetograms that were readable. In particular, we thank G. J. Van Beek, P. Serson, and J. Walker of the Division of Geomagnetism, Earth Physics Branch, Department of Energy, Mines, and Resources in Ottawa for their generosity in supplying data from their array of digital magnetometer stations in Canada. We thank also the Illinois Geological Survey for their description of the Plano area geology. A special note of thanks goes to E. W. Geer and K. D. Tentarelli, Bell Laboratories, for their many helpful discussions on the operation of the L4 system.

## REFERENCES

1. J. G. Roederer and A. H. Shapley, "Overall Summary of August 1972 Phenomena and Data," *Collected Data Reports on August 1972 Solar-Terrestrial Events*, Part I, H. E. Coffey, ed., Boulder, Colorado: U. S. Dept. Commerce, NOAA, July 1973, pp. 1–2.
2. M. Pomerantz and S. P. Duggal, "Record-breaking Cosmic Ray Storm stemming from Solar Activity in August 1972," Nature, *241*, February 2, 1973, pp. 331–332.
3. T. A. Croft, "Traveling Regions of High Solar Wind Density Observed in Early August 1972," J. Geophys. Res., *78*, June 1, 1973, pp. 3159–3166.
4. V. D. Albertson and J. A. Van Baelen, "Electric and Magnetic Fields at the Earth's Surface due to Auroral Currents," IEEE, *PAS-89*, 1969.
5. G. W. Longanecker and R. A. Hoffman, "S$^3$-A Spacecraft and Experiment Description," J. Geophys. Res., *78*, August 1, 1973, pp. 4711–4718.
6. E. I. Loomer and G. J. van Beek, "The Development of Current Vortices in the Polar Cap Preceding the Large Magnetic Storms on August 4–5 and August 9, 1972," Collected Data Reports on August 1972 Solar-Terrestrial Events, Part III, H. E. Coffey, ed., Boulder, Colorado: U. S. Dept. Commerce, NOAA, July 1973, pp. 722–726.
7. D. B. Ledley, private communication.
8. W. D. Cummings and P. J. Coleman, Jr., "Magnetic Fields in the Magnetopause and Vicinity at Synchronous Altitude," J. Geophys. Res., *73*, September 1, 1968, pp. 5699–5718.
9. T. L. Skillman and M. Sugiura, "Magnetopause Crossings of the Geostationary Satellite ATS 5 at 6.6 $R_E$," J. Geophys. Res., *76*, January 1, 1971, pp. 44–50.
10. J. W. Freeman, Jr., C. S. Warren, and J. J. Maguire, "Plasma Flow Directions at the Magnetopause on January 13 and 14, 1967," J. Geophys. Res., *73*, September 1, 1968, pp. 5719–5732.
11. L. J. Lanzerotti, W. L. Brown, and C. S. Roberts, "Energetic Electrons at 6.6 $R_E$ during the January 13–14, 1967, Geomagnetic Storm," J. Geophys. Res., *73*, September 1, 1968, pp. 5751–5760.
12. T. W. Lezniak and J. R. Winckler, "Structure of the Magnetopause at 6.6 $R_E$ in Terms of 50- to 150-keV Electrons," J. Geophys. Res., *73*, September 1, 1968, pp. 5733–5742.
13. G. A. Paulikas, J. B. Blake, S. C. Freden, and S. S. Imamoto, "Boundary of Energetic Electrons during the January 13–14, 1967, Magnetic Storm," J. Geophys. Res., *73*, September 1, 1968, pp. 5743–5750.

14. D. J. Williams, private communications, 1973, 1974; L. J. Cahill, private communications, 1973, 1974.
15. T. Rikitake, *Electromagnetism and the Earth's Interior*, Amsterdam: Elsevier Publishing Company, 1966.
16. J. R. Wait, *Electromagnetic Waves in Stratified Media*, New York: Pergamon Press, 1962.
17. W. J. Peeples, *Magneto-telluric Profiling over a Deep Structure*, Ph.D. Thesis, University of Alberta, Edmonton, Alberta, 1969.
18. H. F. Morrison, *Magneto-telluric Profile Across the State of California*, Ph.D. Thesis, University of California, Berkeley, 1967.
19. T. Madden and P. Nelson, "A Defense of Cagnaird's Magneto-telluric Method," Geophysics Laboratory, MIT, Project NR-371-401, 1964.
20. Illinois Geological Survey, private communication.
21. C. W. Anderson, III, "Effects of the August 4, 1972, Magnetic Storm on Bell System Long-Haul Communication Routes," *Corrosion '74*, National Association of Corrosion Engineers, March 1974.
22. E. H. Angell and K. D. Tentarelli, private communications, 1973.
23. K. D. Tentarelli, private communication, 1974.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 9, November 1974
*Printed in U.S.A.*

# New Frontiers of Varactor Harmonic Power Generation in the C-Band

By S. V. AHAMED and J. C. IRVIN

(Manuscript received June 24, 1974)

*Gallium-arsenide "stacked" varactors employed in a frequency doubler and tripler have given 8- to 10-watt output at 4 and 6 GHz with efficiencies of 70 to 80 percent.*

## I. INTRODUCTION

We have designed and developed a new class of varactors that have been noteworthy beneficiaries of the techniques evolved for GaAs IMPATTs. In this paper we report the experimental results obtained from double-stacked diodes for frequency doubling (2 to 4 GHz), and double- and triple-stacked diodes for frequency tripling (2 to 6 GHz).

## II. VARACTOR CONSTRUCTION

The power and efficiency sought in the present application (8- to 10-W output at 4 and 6 GHz with a minimum efficiency of 70 percent) dictate a low-loss, high-voltage varactor. These conflicting goals are best achieved by a series connection of two or more chips in a single package.[1] Using a Schottky-barrier GaAs chip, the power dissipation per chip is so small that a simple series stacking of one chip upon another is adequate, without heat sinks for any chip except the bottom one. This scheme is facilitated by using dice of typical C-band IMPATT design,[2] i.e., squat cylinders or truncated cones (0.2 mm in diameter by 0.08 mm tall, for example) in which the active area is that of the cylinder itself. The dice are thermocompression-bounded in "flip-chip" position, one on another, in a package with a diamond heat sink. The series combination not only increases power-handling capacity considerably, but it also alters the input and output impedances favorably.

The chips employed here were originally tested as IMPATTs, where they gave about 3 W at 10 to 12 percent efficiency in C-band.

As a double stack, these chips formed a varactor with typical values of 7.5-pF zero-bias capacitance, 174-V breakdown, and 0.65-ohm zero-bias resistance. In a triple stack, the varactor's values were typically 5.0 pF, 265 V, and 0.98 ohm. The cutoff frequencies at breakdown are 300 GHz or greater.

### III. THE 2- TO 4-GHz FREQUENCY DOUBLER

A coaxial embodiment of the doubler (see Fig. 1) has been chosen to yield experimental flexibility. The diode is mounted at the center of a 30-cm, 50-ohm coaxial airline. A 4-GHz reentrant choke to block the output frequency from the input and a 2-GHz quarter-wave-long transformer to match the impedance constitute the input side. A quarter-wave 4-GHz transformer and a twin capacitor 2-GHz filter make the output side. The axis of the inner conductor is positioned at the axis of the outer conductor by radial pressure between the diode and a long, thin, spring-loaded polystyrene pin. The frequency doubling by the diode has been simulated on a digital computer, and the efficiencies and bandwidth data are obtained together with the impedance for best results. The correlation between the simulated results and experimental data is shown in Fig. 2. The half-dB bandwidth at 4 GHz is approximately 300 MHz.

### IV. THE 2.115- TO 6.345-GHz TRIPLER

A coaxial embodiment with three limbs in a T configuration (see Fig. 3) for the input, output, and idler circuits has been designed and constructed for the tripler circuit. The diode is mounted along the axis

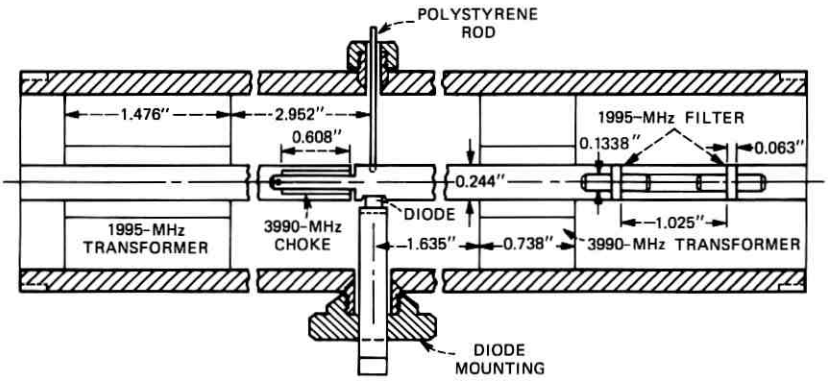



Fig. 1—Cross section of the 1995- to 3990-MHz doubler.

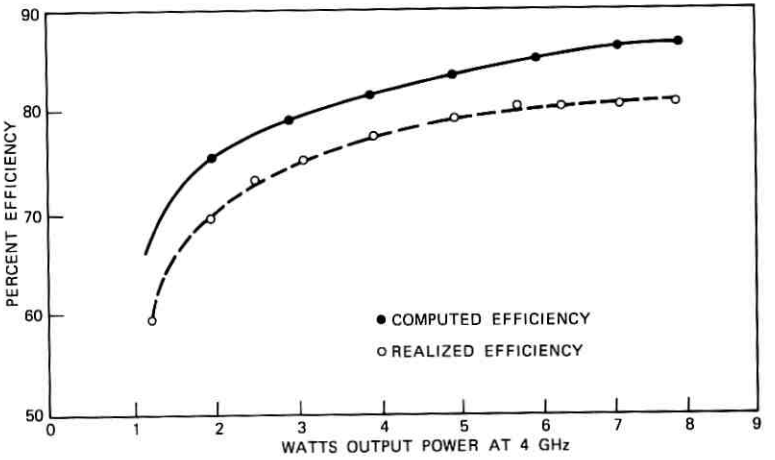



Fig. 2—Efficiency characteristics of the doubler, double stack 7.5 pF, 174V, 0.65 ohm.

of the T opposite the idler stub with the remaining two limbs for input and output sides. A 2.115-GHz quarter-wave transformer, a 4.23-GHz (idler frequency) choke, and a 6.345-GHz choke suitably located with respect to one another and the diode constitute the input side. A 2.115-GHz choke, a 6.345-GHz transformer, and twin-capacitor 4-GHz filter (also serving as a dc block) constitute the output. The idler circuit consists of a 0.53-pF capacitor (which also blocks the dc voltage because of average charge on the diode), a suitable length of the inner conductor, and a sliding short between the inside and outside conductors. The axis of the central conductor between the input and output sides is aligned with the axis of the outer conductor by opposing radial pressures between the stub and the diode. The positions of the various components in the tripler that can be computed agree well with the experimentally determined locations. Fine tuning is accomplished by four pin tuners, two on the idler stub and one each on the input and output limbs. The correlation between the simulated results obtained from a digital computer and the experimental data is shown in Fig. 4. The bandwidth that has been computed to be about 70 MHz at 6.345 GHz is experimentally measured at 0.5-dB points as 61.2 MHz at 8 W, 50.4 MHz at 7 W, 58.8 MHz at 6 W, and 66 MHz at 5 W for a *triple*-stacked varactor. This varactor is extremely resilient to the extent that no damage occurs by gross mistuning at 10 W. The *double*-stacked varactor has yielded 10 W but has been occasionally damaged while tuning, probably because of excessive avalanching.

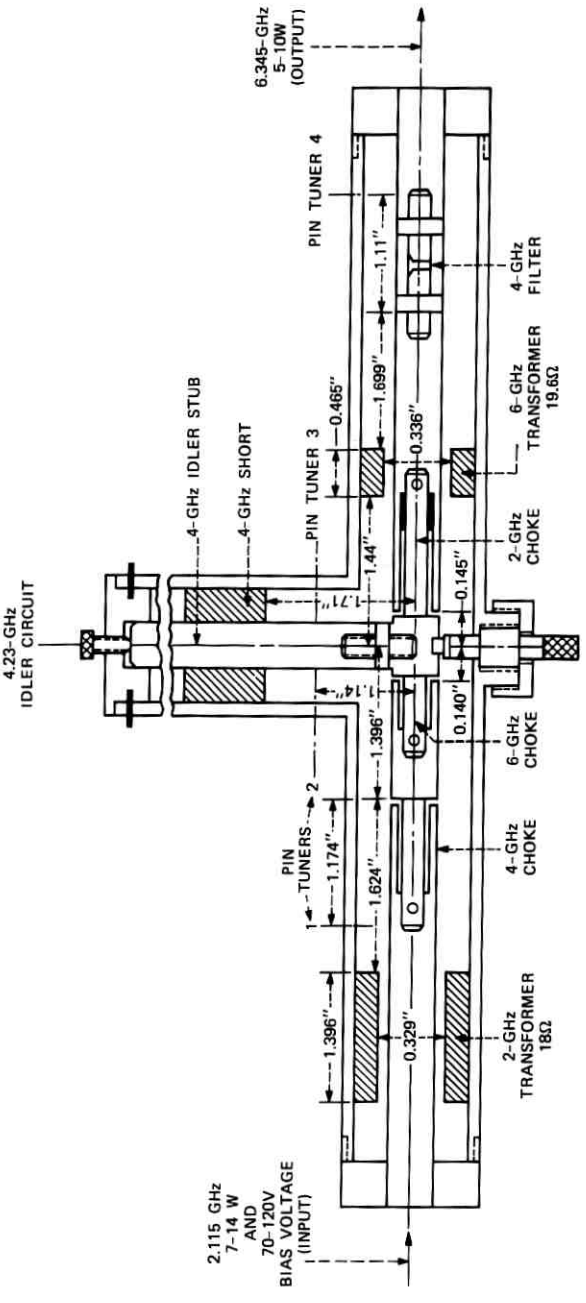



Fig. 3—Assembly of tripler.

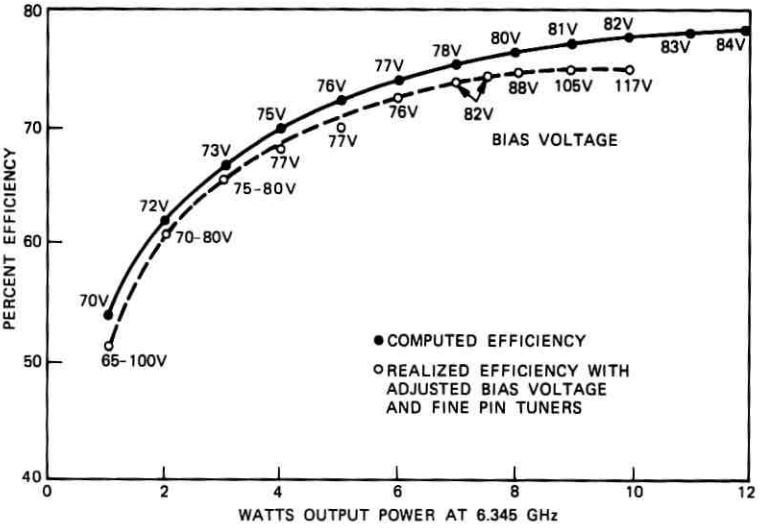



Fig. 4—Efficiency characteristics of the tripler, triple stack 5 pF, 262V, 0.97 ohm.

## V. CONCLUSIONS

Both high power and high efficiency that approach the theoretical limit for GaAs diodes of this design may be simultaneously achieved by proper design of the varactors. The good control in doping density, the use of nonohmic back contacts, and a diamond-based heat sink that permits a tolerable temperature rise of the most thermally isolated chip in a stacked varactor, make these new varactors good contenders for microwave power by frequency multiplication.

## REFERENCES

1. J. C. Irvin and C. B. Swan, "A Composite Varactor for Simultaneous High Power and High Efficiency Microwave Generation," IEEE Trans on Elec. Devices, *ED-13*, May 1966, pp. 466–471.
2. J. C. Irvin, D. J. Coleman, Jr., W. A. Johnson, I. Tatsuguchi, D. R. Decker, and C. N. Dunn, "Fabrication and Noise Performance of High Power GaAs Impatts," Proc. IEEE, *59*, August 1971, pp. 1212–1215.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 9, November 1974
*Printed in U.S.A.*

# The Use of Negative-Impedance Units Inserted Uniformly Into a Transmission Line to Reduce Attenuation

By J. M. MANLEY

(Manuscript received October 29, 1973)

*This paper describes the general properties of a negative-impedance-boosted (NIB) transmission line, an otherwise uniform line to which lumps of negative impedance have been introduced periodically along its length. The purpose of these lumps is to reduce the attenuation of the line, but they also make possible self-oscillation and cause reflections. However, these reflections are cancelled up to a cutoff frequency that has an inverse relationship to the spacing of NIB units. Summaries are included of the latest NIB project; of previous investigations into NIB lines, many of which were not published; and of the first work in the Bell System to reduce transmission-line attenuation through the introduction of loading coils and then unilateral amplifiers. It is shown that many of the differences between properties of NIB lines and properties of a line fitted with unilateral amplifiers come from the different ways the source of energy replenishment is coupled into the line.*

*It is shown that, under suitable conditions, the addition of small simple NIB units to a transmission line can reduce its loss to a low level in a stable manner and also equalize it. Because of some properties of NIB lines, they cannot be used successfully in all situations; in particular, an environment in which large longitudinal currents are induced in the lines is unfavorable.*

## I. INTRODUCTION

The use of negative impedances to reduce the attenuation of telephone transmission lines is the subject of this paper. The principal material is in Section V and deals with the general properties of transmission lines to which negative impedances have been added periodically. These properties are compared with those of lines with loading

coils and lines having conventional unilateral amplifiers. This use of negative impedances has been investigated a number of times in several ways over the past forty years, but very little of this work has been published. A brief history of this effort is given in Section VI. A short section on the properties of negative-resistance elements is also included.

For perspective in viewing the use of negative impedances in transmission lines, a few highlights are provided from the early development of means to reduce attenuation in the lines of the Bell System. This development from early Bell System work makes a very interesting story in itself, parts of which may be found in Refs. 1 through 8—the account in Ref. 1 being especially good. However, this background information is given here not only for perspective, but because the introduction first of loading coils and then of gain by means of unilateral amplifiers furnishes two important points of reference in discussing the properties of attenuation reduction by means of negative impedance.

The most recent study of the use of negative impedances in transmission lines, begun several years ago by L. A. Meacham, is summarized briefly in Section VII.

## II. LOSS REDUCTION BY LOADING COILS AND AMPLIFIERS

The first major advance in pushing back the barrier of distance in telephone transmission came around 1900 with the introduction of loading coils. Through the work of Heaviside and others, it had been known for some years that if the series inductance of a line could be increased, its attenuation would be decreased. But there was no satisfactory way to increase continuously a line's inductance, although this was done to a limited extent in submarine cables later on. Soon after joining the AT&T Laboratories in Boston, Dr. George A. Campbell in 1899 developed a theory of loading for adding inductance in finite lumps to a line.[3] In an unusual coincidence, a similar theory was worked out by Professor Michael Pupin of Columbia University at about the same time. In a patent-interference case, it was decided that Professor Pupin preceded Dr. Campbell by a few days. Campbell's analysis, which was more detailed than Pupin's, gave the relation between kind of line, size of coils, and their spacing and cutoff frequency. These relations were used to begin the introduction of the loading process and are still widely used today.

Campbell's invention of the wave filter came from the results of his analysis of the loading process, which was seen to make a low-pass filter of the line. The making of suitable coils was quite a problem

itself in those days, but within about a year, installations of coils were beginning and attenuations were being reduced to about one-half their former values. This was a great advance and made possible a large expansion of the telephone network in distance and use. The theoretical cutoff frequency of these loaded lines was about 2300 Hz, but because of the core material in the coils, the actual line losses began to increase considerably before this cutoff. With later improvements in coils and technique, line losses were often reduced to one-third or one-fourth the nonloaded values.[2] By 1905, with the use of these coils and #8 or #10 wire, the long-distance open-wire network in the East had been extended as far west as St. Louis and Kansas City without benefit of repeaters and by 1911, Denver had been reached.

An important goal at the time was to link the east and west coasts by telephone. To do this, either #5 wire (0.18-inch diameter) had to be used in the line from Denver to San Francisco, or a suitable form of repeater had to be found. The repeater path appeared to be the more economical and so the work already under way on repeaters and their use in transmission lines was expanded.

In fact, between the invention of the telephone and the appearance of the high-vacuum-tube amplifier in 1913, there was a large amount of effort expended, both inside and outside the Bell System, to devise a workable amplifier for telephone signals. During this time, many inventions were submitted to the AT&T Company for possible use as telephone repeaters. Some of these involved rotating electrical machinery or other ponderous mechanical devices that were not suited to high-frequency telephone signals. Others, on analysis, turned out to be simply transformers with no energy source.[4] An amplifier is really a modulator in which an incoming signal, with a small expenditure of energy, modulates a larger local energy source to produce a gain as it repeats the signal. In a vacuum-tube amplifier, the dc plate current supplied by a battery is modulated by the grid voltage. In a transistor amplifier, the collector current is, in effect, modulated by the much smaller base current. In a carbon-button amplifier, the varying resistance of the carbon granules in response to the motions of the receiver armature modulates the dc flowing through the button with a power gain of several hundred. In a parametric amplifier, the source of energy is a local oscillator supplying pump current which is modulated by the signal. And so on.

During the time when a suitable repeater was most actively sought, only two devices showed enough promise to be seriously considered. One of these was a combination of carbon-button transmitter and

receiver which, in its best version, had the transmitter and receiver coupled mechanically instead of acoustically. A few commercial installations of this device were made although it had a number of problems, one of which was variability of gain. The other device was a gaseous discharge that fundamentally had the characteristics of a negative resistance rather than a unilateral amplifier.

At this point, the three-electrode vacuum tube appeared on the scene and showed such great promise that the first two devices were soon put aside. The first high-vacuum-type repeater was installed in late 1913, a little less than a year after H. D. Arnold had been shown DeForest's three-electrode tube. By the middle of 1914, three improved repeaters had been successfully installed and operated in the new transcontinental line. The net loss of this line was 20 dB and the 10-dB bandwidth was from 350 Hz to 1250 Hz.[1,4]

## III. PROBLEMS AND PROPERTIES OF REPEATERED LINES USING UNILATERAL AMPLIFIERS

The introduction of repeaters into the telephone plant showed up transmission problems that had been obscured before. At any point in the line where its uniformity is disturbed (called an irregularity), part of a wave traveling down the line will be reflected back toward the wave's source. Such a reflected wave is like an echo. When the net loss of the line is 20 dB or more, the returning echoes are weak and hardly noticeable. But when gain was added to such a circuit, the echoes could be very objectionable. In fact, they could limit the amount of usable gain. For example, the net loss in the first transcontinental line in 1914 was limited to 20 dB.

Another factor in the effect of echoes is the transmission time of the line. Because of the inductive loading, the transmission time of the transcontinental line was 70 ms, which is high enough so that if strong echoes occur they will be objectionable. As understanding of these effects grew, the loading was removed, reducing the transmission time to 20 ms. With this change, the loss increased about 100 dB but more gain could be added, reducing the net loss to 12 dB. An extra benefit that came from the removal of the loading was higher-quality speech transmission because of the increased bandwidth.[5] Since the attenuation vs frequency characteristic of open-wire lines is fairly flat, the use of loading restricted the bandwidth. Gradually all loading coils were removed from open-wire lines. The situation is different with cable pairs, which have a considerably higher capacitance. Here the coil loading tended to flatten the attenuation vs frequency curves

so that only a little equalization was required at the repeaters. Hence, loading has generally been kept in voice-frequency cable circuits, although the amount of loading has been reduced in long circuits to reduce transmission time.

Another effect of the reflections when repeaters are in the line is the possibility of free oscillation in the line, or "singing." If an amplifier is inserted into a perfectly uniform transmission line in such a way that it does not cause reflections, the gain of the amplifier may be as large as desired. But if there are reflections, the amount of gain that can be inserted will be very definitely limited by either echoes or singing.[5,8–11] Reflections can be caused by an improper termination of the line, by a missing or misplaced loading coil, or by the way amplifiers are coupled into the line.

To obtain bilateral transmission with a unilateral amplifier, a special kind of circuit is required to couple the amplifier into the line. Some form of bridge circuit is the usual means—a Varley loop was first used for telegraph repeaters. Coil arrangements for doing this are called hybrid coils, one of which plus one amplifier make a 21-type repeater that can amplify in both directions. But its losses are high and it sends amplified signals in both directions with the effect of an echo even though the lines in both directions are well-balanced against each other. Hence, they were not used very much. The 22-type repeater, shown in Fig. 1, uses two hybrid coils, two amplifiers, and two line-balancing networks, and gives much better performance. Losses are minimal and reflections come only from imperfections in balance between the line in each direction and its balancing network. This means for coupling amplifiers into a line was invented in 1895 (U. S.

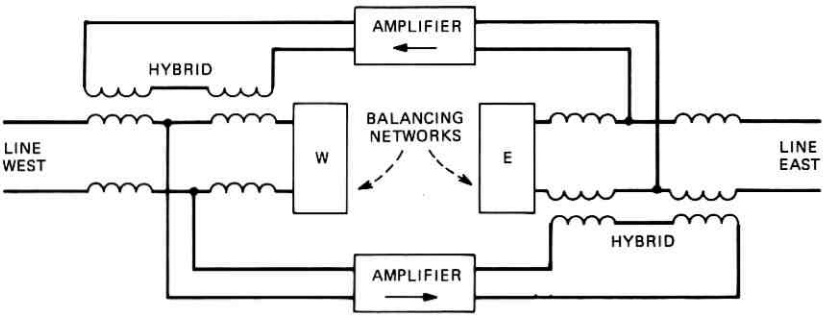



Fig. 1—Diagram for connecting unilateral amplifiers and hybrid coils to make a 22-type repeater.

Patent 542,657) by a Bell System engineer, W. L. Richards, long before there were good amplifiers with which to use it.[4]

After the successful introduction of the 22-type repeater into the system, the AT&T Company continued to have submitted to it inventions that purported to reduce in varying degrees the complexity of the means for coupling unilateral amplifiers into the line. They were mainly variations of the basic bridge circuit or were completely unworkable. No satisfactory substitute for the hybrid coil arrangement of Fig. 1 has appeared except for the very different approach to overcoming line loss by the introduction of negative resistances, which will be discussed below.

As may be seen from Fig. 1, the amount of gain that can be obtained from this repeater configuration before local singing occurs depends on the loss across the hybrids (vertically in the diagram) and is higher for higher losses. This loss is infinite if the network exactly balances its corresponding line and is zero if the line presents a short or open impedance. Thus, to operate these repeaters with considerable gain in two-wire lines requires a fairly high degree of balance between networks and lines. When a number of repeaters is used in tandem in a line, additional possible singing paths are introduced as well as additional echo sources at each departure from uniformity, which from a practical viewpoint is at each repeater. To eliminate these multiple singing paths and echo sources, four-wire transmission was introduced in many toll circuits and is universally used now in all carrier circuits. In this arrangement only two hybrid coils are used, one at each terminal. The upper path in Fig. 1 becomes a two-wire line with a sufficient number of east-west amplifiers approximately uniformly spaced. The lower path is similar but transmits only west to east. The singing problem is greatly reduced with this arrangement because both horizontal transmission paths between the terminal hybrids have very little if any gain since each includes the line loss as well as the amplifier gain. In some respects, the four-wire configuration of repeaters is similar to the use of series negative-impedance boosters, as will be seen below.

If the two wires of the pairs in the two-wire and the four-wire circuits, and the repeater circuits connected to them, are well-balanced against each other, interference currents will largely be drained off to ground away from the repeaters and terminals.

The 22-type repeaters were economically as well as technically successful in toll transmission, but were deemed too expensive and complicated for exchange trunks where gain was needed in the general

upgrading of telephone quality in the nineteen twenties. Partly because of continuing study of repeatered systems and partly because of the need for a cheaper repeater, as indicated above, different ways of reducing line attenuation were sought.

One way, quite different from the use of unilateral amplifiers described above, and which appeared to hold considerable promise, was that of introducing negative resistances into the line to cancel, or partly so, the positive line resistance, which, because of good dielectrics, is the principal source of loss in lines. While some thought had been given to such use of negative resistances at about the time the vacuum-tube repeater appeared, as indicated above, a considerable effort in this direction began a little before 1930 and has continued off and on until the present. A review of this work is given below in Section VI.

Now, with our summary of the processes by which gain may be obtained, we are able to see a certain equivalence between an amplifier and a negative resistance. Yet the difference in the ways they must be used in a transmission line, or in the particular form one or the other takes, may make it more desirable to introduce the gain in one of the ways instead of in the other.

The next section will describe some of the properties of negative resistances and devices that exhibit this characteristic.

## IV. NEGATIVE RESISTANCES

A positive resistance absorbs energy from connected circuits and dissipates it. A negative resistance delivers energy to connected circuits and so must have within it a source of energy on which it may draw. More precisely, a negative resistance as a device is able to convert the energy of a local source, to which it is connected, into a form suitable for passing on to other circuits. In other words, it must be able to modulate the energy from the local source as an amplifier does. An amplifier, though, is usually a three-terminal device in which the energy stream is controlled externally. A negative resistance, on the other hand, is a two-terminal device utilizing some internal phenomenon that depends on the voltage across it or the current through it to control the energy stream. An early paper by Crisson[12] describes two kinds of negative resistance. He shows also in this paper that either form of negative resistance can be generated by connecting the output of an amplifier to its input either in a series way or a shunt way. This suggests that in general a negative resistance is the result of some kind of feedback process, that is, one in which current or voltage depends on itself in some way, either externally or internally. A good general

paper on negative resistance and devices which exhibit it is the one by E. W. Herold.[13]

Some parasitic reactance may be associated with a negative-resistance device and/or some may be intentionally added to it so that, in general, an impedance with a negative resistance component is what is used. The general term negative impedance is applied to such a device. Because of the reactance, the impedance usually does not have a negative resistance component above a certain frequency. It is possible also for negative impedances to be generated in a largely reactive device, such as a varactor, by an internal modulation process. Here the energy source is not dc but the external pump oscillator.

When a negative resistance is generated by connecting the output of an amplifier to its input, the voltage-current characteristic is a line with negative slope extending from the origin to where the amplifier begins to overload. Here, in effect, the true energy source is concealed within the device. However, in many negative-resistance devices, the energy comes to the device from the source through a direct current that also flows through the two terminals of the device. Over a certain range of values of this current, the device exhibits a negative resistance and so is able to convert energy from the source to a form suitable for use in the connected circuits. All known negative resistance devices of this type have either one or the other of two shapes of voltage-current characteristic, as illustrated in Fig. 2. The series type is shown in Fig. 2a, and the shunt type in Fig. 2b. Between points $A$ and $B$, the slope $dV/dI$ is negative and, for current variations in this range, the device presents a negative resistance to external circuits. The magnitude of the negative resistance is this slope expressed in ohms.

To get an understanding of the effect of a negative resistance in a circuit, consider the diagrams of Fig. 2c, where a positive resistance $R_o$ has been added to the negative resistance $R$ and the resulting overall voltage-current curve drawn. When the added positive resistance just balances the negative component, the V-I curve of the combination has a slope of zero between $A$ and $B$ (Fig. 2c). That is, if, as indicated in the figure, a varying signal current is superimposed on the bias of direct current supplied by the energy source, the voltage drop across the combination of positive resistance and negative resistance device is zero as long as the variations remain within the range $A$ to $B$. This means that all the positive signal energy dissipated by the positive resistance $R_o$ comes from the negative resistance $R$ and none from the signal generator $e$. This is an indication of how the negative resistance

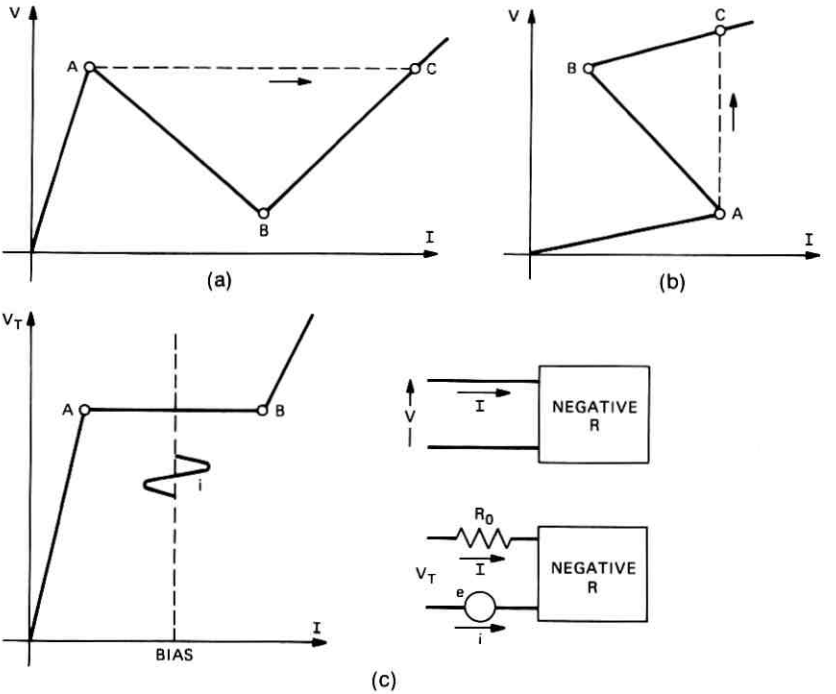



Fig. 2—Voltage-current curves for negative resistance.

supplies energy to the connected circuits by appearing to cancel or nullify the positive resistance, or a part of it.

Of course the borderline case just discussed could not really exist because the signal generator $e$, being zero, is not in control. To understand this incipient instability, let $R_o$ be less than $R$ so that the slope of the V-I curve between $A$ and $B$ is now negative as in Fig. 2a, and consider also the small parasitic inductance $l$ which would be in series with the resistances. Then the signal current $i$ would be given by the expression

$$i = e/(sl - R + R_o),$$

where $s$ is complex frequency $\sigma + j\omega$. If the magnitude of $R$ is greater than that of $R_o$, the impedance of the circuit has a root in which $s$ has a positive real part and so is unstable. If, in addition, there is a resonance of reactances associated with the circuit, sustained oscillation will occur at the frequency of resonance. Thus, there is an

instability barrier that prevents the negative resistance from supplying any more energy to connected circuits than just enough to cancel the losses in the external positive resistance.

Because of the instability conditions just described, the V-I characteristics of Fig. 2 cannot be obtained directly from measurements on the device alone. If the voltage in Fig. 2a or the current in Fig. 2b is increased from 0 until point $A$ is reached, the current in Fig. 2a or the voltage in Fig. 2b will appear to jump to point $C$, skipping over the intermediate part of the characteristic. Actually this is a very fast transient with a positive real exponent as indicated above. This instability will not occur if a positive resistance a little larger than the magnitude of negative resistance is connected in series with the series device, nor will it occur if a positive conductance a little larger than the negative conductance is shunted across the shunt device. Hence, the names, series and shunt, are given to the two types of negative resistance. The characteristics, Figs. 2a and 2b, are deduced by measurements of the combination circuit, as shown in Fig. 2c. The slope of the negative-resistance region can be measured quite accurately in such a circuit by adjusting $R_o$ to obtain a null of signal voltages across the series combination of $R$ and $R_o$.

One of the earliest negative-resistance devices is the carbon arc (or other gaseous discharge). This is the series type and is generated when the ionization in the discharge begins to increase by means of the current of the discharge itself. The avalanche process in semiconductors is the modern counterpart of this. The dynatron invented by A. W. Hull[14] is of the shunt type and depends on secondary emission, which becomes appreciable when a voltage reaches a certain value. Other devices that can exhibit negative resistances are thermistors, avalanche diodes and transistors, tunnel diodes, IMPATT diodes,[15] and Gunn-effect diodes.[16] The paper on IMPATT diodes by K. D. Smith[15] is a particularly good description of this device and of the general properties of such negative resistances. As the physical dimensions are reduced, the parasitic reactances are reduced and the device can show a negative resistance at higher frequencies. The IMPATT and Gunn-effect diodes differ from the others in that they show a negative resistance only in a narrow frequency band, depending on their sizes.

Other sources of negative resistance are made by combining several devices such as two vacuum tubes or two transistors and, in this form, they are often called negative-impedance converters.[17] Vacuum tubes were used in the first E-type negative-impedance repeaters and transistors in the latest ones and also in the work described below.

When negative resistances are made in this way, the opportunity exists of modifying the variations of the negative resistance with frequency and/or current by adding passive circuit elements to the configuration in order to serve particular needs. They also can be made in integrated-circuit form for small size.

## V. GENERAL PROPERTIES OF TRANSMISSION LINES WITH NEGATIVE-RESISTANCE LUMPS ADDED PERIODICALLY

It was shown in Section IV how a negative-resistance device supplies energy at signal frequencies to circuits in which it is connected, acting as if it cancelled the effect of positive resistance in the same circuit. Hence, it is natural to think of this as a way to reduce the attenuation of a transmission line. This is not a new idea. It was mentioned as a possibility in the 1919 repeater paper[4] referred to above. The term booster was applied to negative-impedance devices used in this way in that paper and also in another.[12] This name was revived by Meacham in his negative-impedance work.[18] In quite a bit of the literature, the process of adding negative impedances to a transmission line is called negative-impedance loading, drawing a parallel to the familiar practice of coil loading. The term boosting seems to the author to be a better description of the process and will be used here.

This section deals with some of the general properties of negative-impedance-boosted (NIB) lines. This will be done without reference to any particular form of negative impedance, as far as is possible. However, to discuss the particular shapes of the resulting transmission curves or the particular details of stability, it is necessary to completely specify the properties of the negative impedance used. It is a fact also that whether or not the general idea of boosting is attractive may depend on the existence of a booster with certain properties. Thus, while the possibility of using boosters to overcome attenuation was recognized in the Jewett paper[4] of 1919, as mentioned above, no very suitable devices were available then.

### *5.1 Coupling negative impedance into the line*

The simplest method of utilizing negative-impedance boosters (NIB) in a line is to connect the series-type units in series with both wires of the line and uniformly spaced along it, as shown in Fig. 3. In this form, the boosters are simply two terminal devices through which the line currents flow. A direct current, which flows through boosters and line along with signal currents, supplies the necessary energy to the boosters. The unit may consist of a single special element or a combination of several circuit elements. Experimental versions of

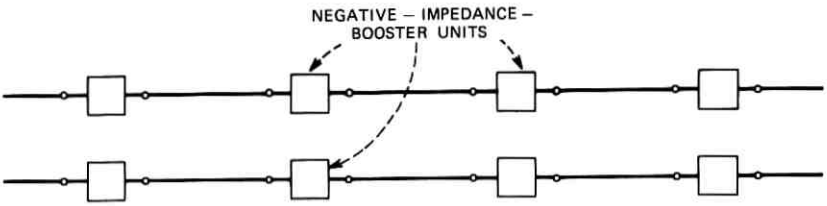



Fig. 3—A NIB line made by the periodic series insertion of negative-impedance units into a transmission line.

the latter arrangement have been made very small in size using hybrid integrated-circuit techniques. In the E-type negative-impedance repeaters,[19,20] transformers are used to couple the boosters into the line, and these repeaters are used only at central offices. Thus, the manner of using negative-impedance boosters is quite different from the manner of using unilateral amplifiers.

While most of the descriptions in this paper are in terms of boosters added in series with the line, reduction of attenuation may be obtained also by adding suitable negative resistances in shunt to the line. Or a combined series and shunt connection may be used. In the latter circumstance, the combination booster may be made to present to the line an impedance that approximates the characteristic impedance of the line. This has been done in some of the E-type negative-impedance repeaters[21] discussed in Section 6.3. Reflections in this case are small and the behavior of the line in this respect is more like one equipped with conventional repeaters.

#### 5.2 Bilateral transmission

As may be seen from the discussion about Fig. 2c in the previous section, the performance of the negative resistance is just the same whether generator $e$ is on the right or the left. And so another general property of boosted-transmission lines is that transmission over the line is bilateral and symmetrical, as it is for the uniform nonboosted line. However, this is true only over a certain range of current through the line, i.e., the range within which the booster presents a negative resistance, as shown in Fig. 2. This property exhibits a fundamental difference between the negative-resistance energy-conversion device and the unilateral-amplifier energy-conversion device.

#### 5.3 Necessity of using lumps of negative resistance

If the negative resistance could be added in infinitesimal bits of ohms, the line resistance would be neutralized continuously and an

ideal line with propagation constant $j\omega\sqrt{LC}$ and image impedance $\sqrt{L/C}$ would result. But, for the present at least, negative resistance is available only in sizable finite lumps and so this is what must be used. There are several consequences. First, adding lumps of negative resistance does not yield the same propagation constants and image impedance that adding the same total amount of negative resistance continuously does. Formulas are given in Section 5.5 below. Second, the added lumps cause reflections, the effects of which are described in Section 5.4. Third, the consideration of even very small lumps of pure negative resistance is nonphysical, since this implies an infinite energy source to supply the infinite band of frequencies. Of course, parasitic reactances in real devices prevent this conceptual difficulty from arising in practice. But there is a further restriction. In Section 5.9, it may be seen that, because of transmission-line properties, the negative resistance must be reduced as frequency increases to avoid instability.

#### 5.4 Reflection effects and cutoff frequency

The addition of these lumps of negative resistance to the line introduces reflections, because each negative resistance presents a discontinuity to traveling waves and so each is a source of reflections. But, as shown by G. A. Campbell in the case of added series inductance,[3] if the added series impedances are spaced uniformly, the sum of all the reflections is zero up to a certain cutoff frequency. This cutoff occurs at the frequency for which the spacing of the added impedances is one-half wavelength of the boosted line. Thus, an infinitely long boosted line behaves like a smooth line up to the cutoff frequency. Another consequence of the process is that the cutoff frequency is related in an inverse way to the spacing of the added impedances, as shown in Fig. 4. More detailed relations are given in Ref. 22.

Another general property, then, is that a boosted line has a cutoff frequency because of the reflections introduced by the boosters and that this cutoff is related to the spacing between boosters in a more or less inverse way, and that up to the cutoff, all the reflections cancel so that the line looks like a smooth line for these frequencies.

#### 5.5 Propagation formulas for NIB lines

Adding circuit elements in finite lumps to an otherwise uniform transmission line results in a line with different properties than one to which the same total amount of elements has been added continuously.

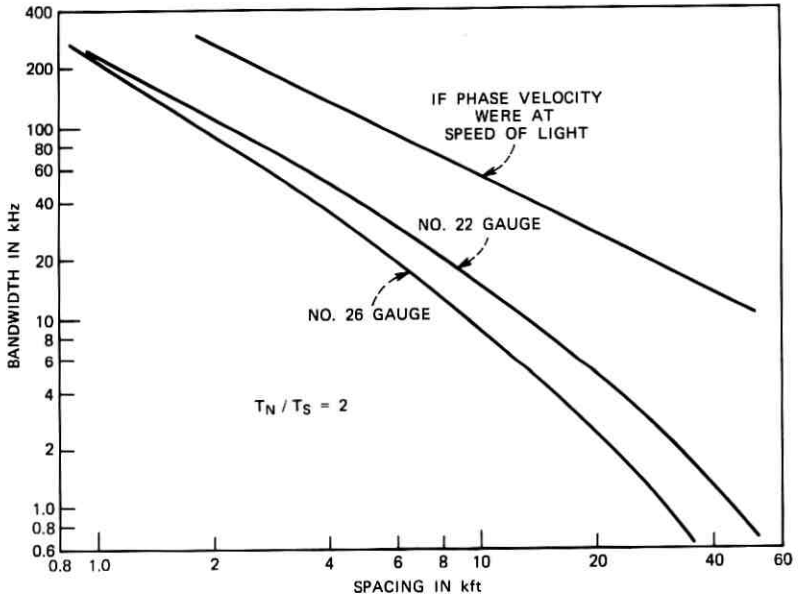



Fig. 4—Relationship between bandwidth and spacing in NIB lines.

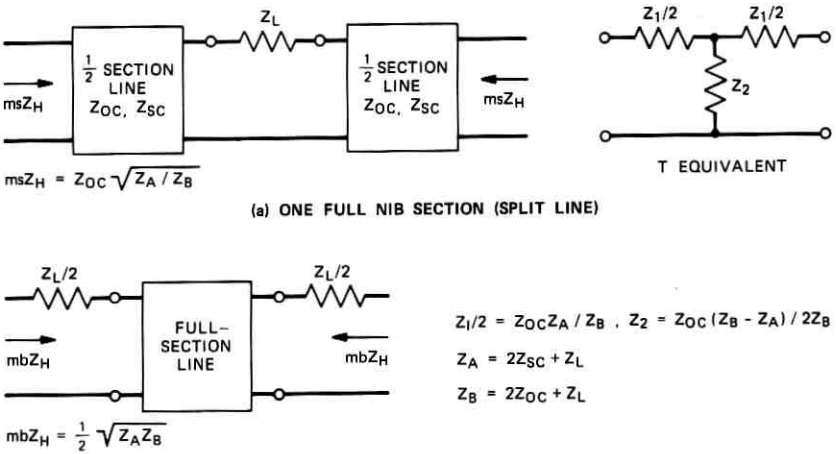



Fig. 5—Two forms of NIB line sections with equivalent circuits and image impedances. $Z_L$ is the total impedance of a pair of NIB units.

The formulas for propagation constant and characteristic impedance of a uniform line cannot be modified by replacing the old element value with a new average value. The method used to derive formulas for the boosted line was first suggested by Campbell,[3] in his calculations on the addition of loading coils. Many of the formulas appear in unpublished notes written around 1940 by Bullington[23] and around 1950 by Van Wynen,[24] both of Bell Laboratories. To obtain these formulas, the appropriate line sections are combined with the negative-impedance unit $Z_L$ in either of two ways shown in Fig. 5 and an equivalent $T$ or $\pi$ network derived. These complete NIB sections, in either of the two forms, may then be cascaded to form a NIB line of the desired length. The midsection and midbooster image impedances of the boosted line are

$$msZ_H = Z_{oc}\sqrt{Z_A/Z_B}$$

and

$$mbZ_H = \tfrac{1}{2}\sqrt{Z_A Z_B},$$

where

$$Z_A = 2Z_{sc} + Z_L$$
$$Z_B = 2Z_{oc} + Z_L,$$

and $Z_{sc}$ and $Z_{oc}$ are the short- and open-circuit impedances of one-half a line section and $Z_L$ is the booster impedance, as indicated on Fig. 5. The propagation factor per section of the boosted line, $P$, is given by

$$\sinh (P/2) = \sqrt{Z_A/(Z_B - Z_A)}.$$

Another method of analysis, used in the work of A. L. Hopper,[22] is to represent each of the three cascaded parts of the NIB section by a matrix. The $A$ form transmission matrix for the half-section of line is

$$M \text{ line} = \begin{pmatrix} \cosh p & Z_o \sinh p \\ \dfrac{\sinh p}{Z_o} & \cosh p \end{pmatrix},$$

where $p$ is the propagation factor for the half-section of line and $Z_o$ is its characteristic impedance. The matrix for the series booster is

$$M \text{ booster} = \begin{pmatrix} 1 & Z_L \\ 0 & 1 \end{pmatrix}.$$

By multiplication of the three appropriate matrices, a new matrix for the full NIB section of Fig. 5a is found. The elements of this new $A$

matrix are

$$A_{11} = \cosh 2p + \frac{1}{2}\frac{Z_L}{Z_o}\sinh 2p = \cosh P$$

$$A_{12} = Z_o \sinh 2p + Z_L \cosh^2 p = Z_H \sinh P$$

$$A_{21} = \frac{\sinh 2p}{Z_o} + \frac{Z_L}{Z_0^2}\sinh^2 p = \frac{\sinh P}{Z_H}$$

$$A_{22} = A_{11}.$$

Here, $2p$ is the propagation factor for the full section of line only. In the final column above, the matrix elements are expressed in terms of the parameters of the full boosted section, $P$ being the propagation factor and $Z_H$ the image impedance. These may be calculated from the matrix coefficients. By multiplying in sequence $n$ matrices of the last form, the equivalent transmission matrix of a cascade of $n$ NIB sections may be obtained and its properties determined.

### 5.6 Image impedance

Another general property of boosted lines is that the image impedance is fairly close to being resistive and fairly close to being flat with frequency. This property is described best by the curves of Fig. 6 computed for the particular realization of negative impedance devised by L. A. Meacham and described in more detail below in Section VII. The parameter $R_{NET}$ or $\Delta r$ is introduced here. It is the sum of line resistance plus booster resistance, at very low frequencies, per section or per mile or for a whole line, as designated. The figure shows the image impedance $Z_H$ for a particular NIB line configuration along with the $Z_o$ of the nonboosted line. While $Z_H$ is smaller than $Z_o$ over much of the frequency range, it is not as small as $\sqrt{L/C}$, which it would be if the negative resistance had been added continuously. However, its property of being fairly close to resistive and flat with frequency is a great improvement over the $Z_o$ of nonboosted lines. Its smaller magnitude reduces crosstalk considerably.

As will be seen in Section 5.10, the image impedance at very low frequencies may not be just as shown in Fig. 6. The frequency scale would have to be considerably expanded to show this.

### 5.7 Transmission properties

#### 5.7.1 Infinite line

The next general property (related to the previous one, of course) is that the transmission characteristics of the boosted line are much better than those of the nonboosted line. As mentioned above, the shape

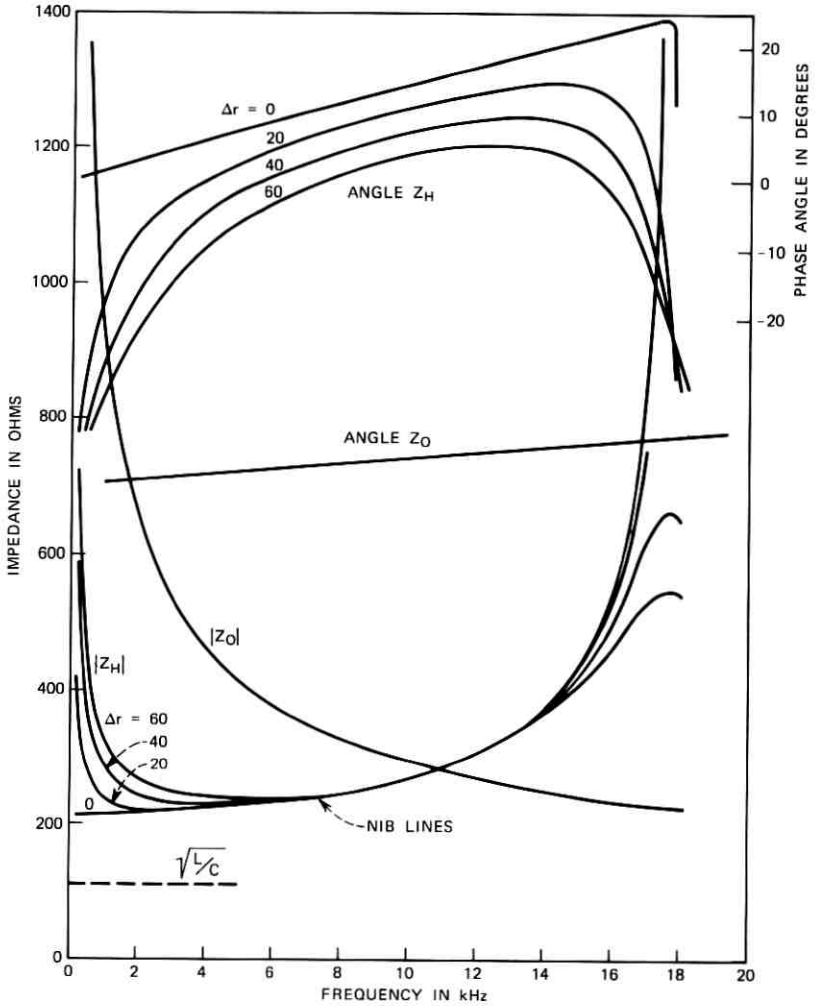



Fig. 6—Image impedance vs. frequency for a 26-gauge NIB line with 6000-foot booster spacing. Parameter $\Delta r$ is the net resistance at very low frequencies per section of NIB line. The characteristic impedance $Z_o$ of the line before the insertion of NIB units is also shown.

of these characteristics may be chosen by adjustment of the booster parameters. When the choice is for as flat an attenuation curve as possible, the two parts of the propagation constant are plotted in Figs. 7, 8, and 9 for the same particular line configuration used to describe $Z_H$, along with propagation of the nonboosted line.

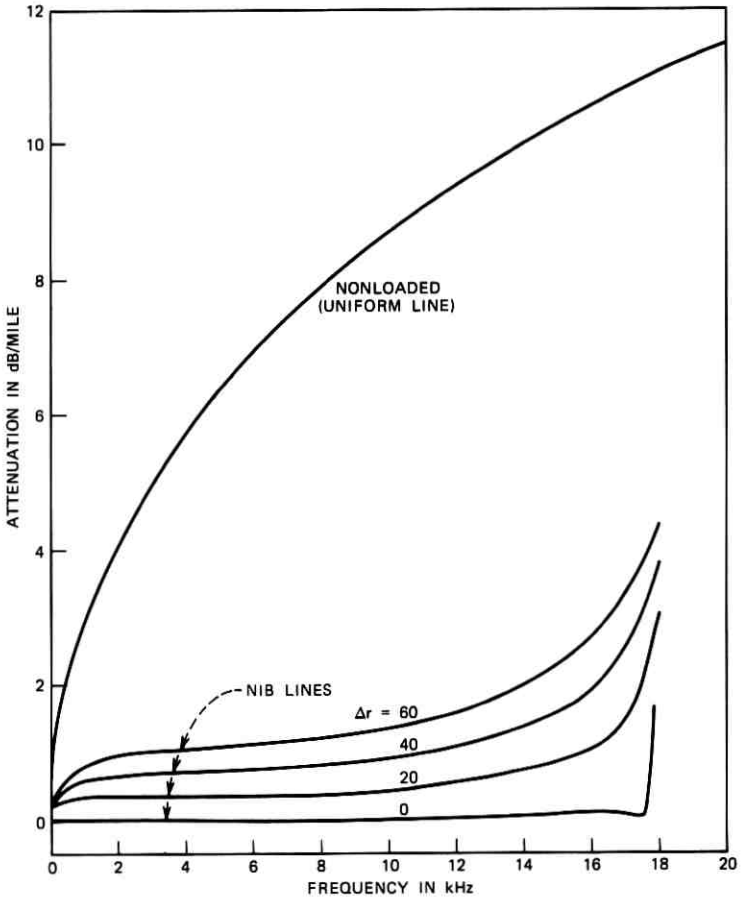



Fig. 7—Attenuation vs frequency for a NIB line and the uniform nonboosted line.

To further highlight the transmission properties of boosted lines, a comparison with coil-loaded lines is made. The boosting process is a much more powerful one because it adds energy at signal frequencies. In a coil-loaded line, the reduction in attenuation is directly related to the amount of inductance added. Hence, reduction in loss is related to cutoff frequency and spacing because of the relation between these two and inductance. In the boosted line, the attenuation reduction can be varied by means of the parameter $\Delta r$, independently of the booster spacing and cutoff frequency. The relation between these two latter parameters is shown in Fig. 4.

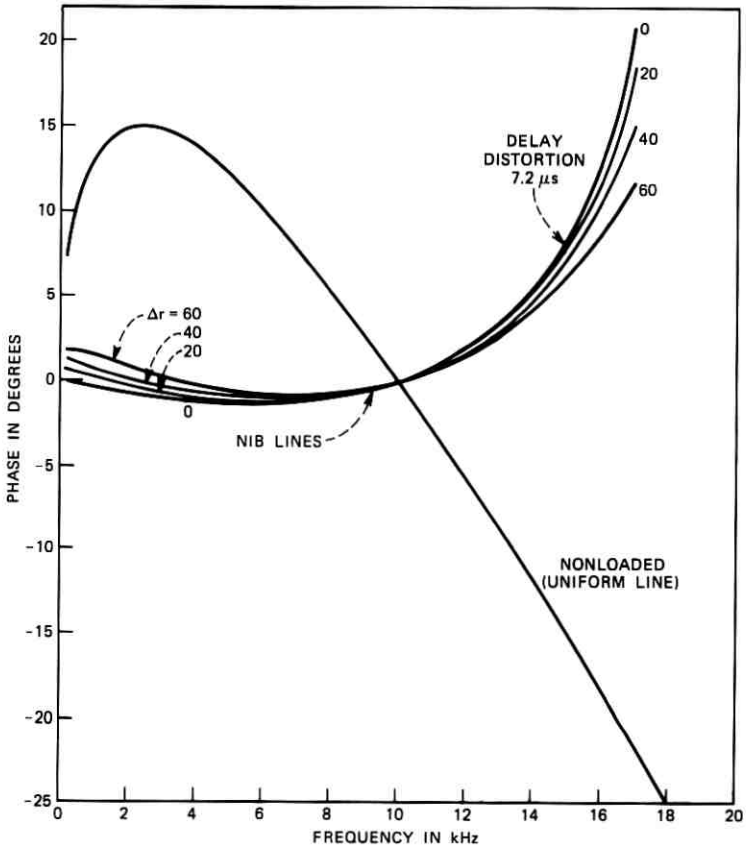



Fig. 8—Deviation of phase from linearity for a NIB line (26-gauge wire and 6000-foot booster spacing) and the uniform line.

The comparatively sharp cutoff of the coil-loaded line is accompanied by large phase distortion, roughly 10 or more times that of the boosted line. The curves of Fig. 9 show that the phase velocity for coil-loaded lines is very low while that of NIB lines is high. However, the introduction of coil loading has been very beneficial in improving and extending the telephone plant.

#### *5.7.2 Finite lines*

All the discussion so far has been in terms of the properties of the infinite line. These can be approached in the finite line if the terminations are close to the image impedance, otherwise there are differences.

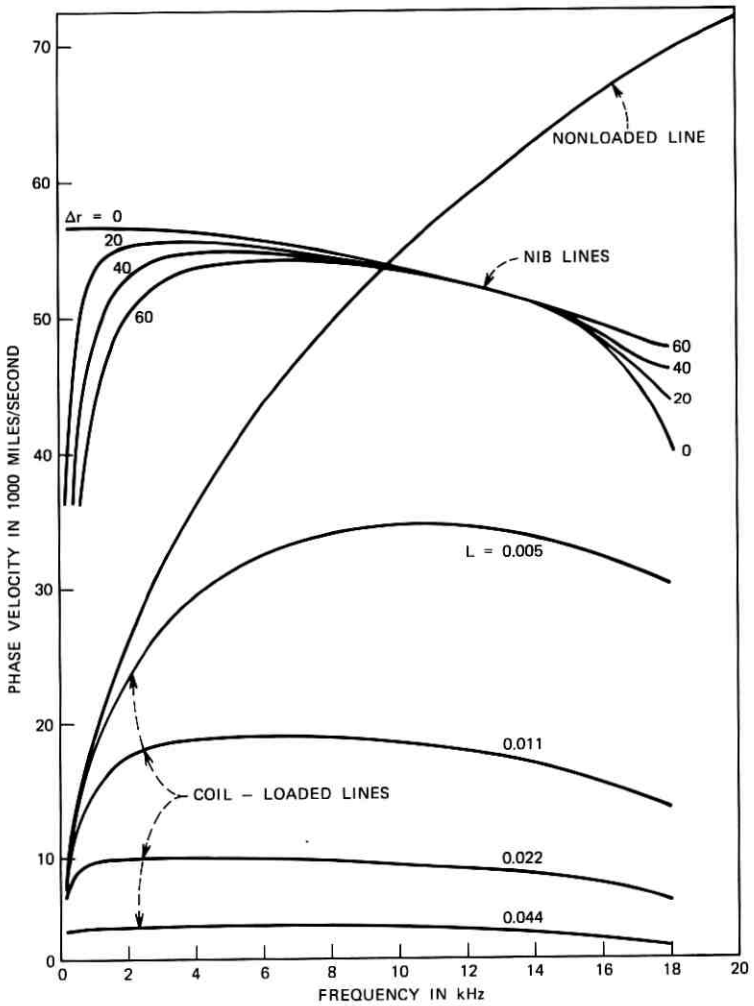



Fig. 9—Phase velocity vs frequency for a NIB line and for several coil-loaded lines.

We saw above that all the reflections from the NIB units cancel in the infinite line. But when the terminations of a finite line differ from the image impedance, all these reflections no longer cancel and the input impedance of such a line is not smooth like the $Z_H$ curves. There will be one ripple for each boosting point and the magnitude of the ripples will be directly related to the magnitude of the difference between the actual termination and the image impedance. The same applies to the insertion-loss curves.

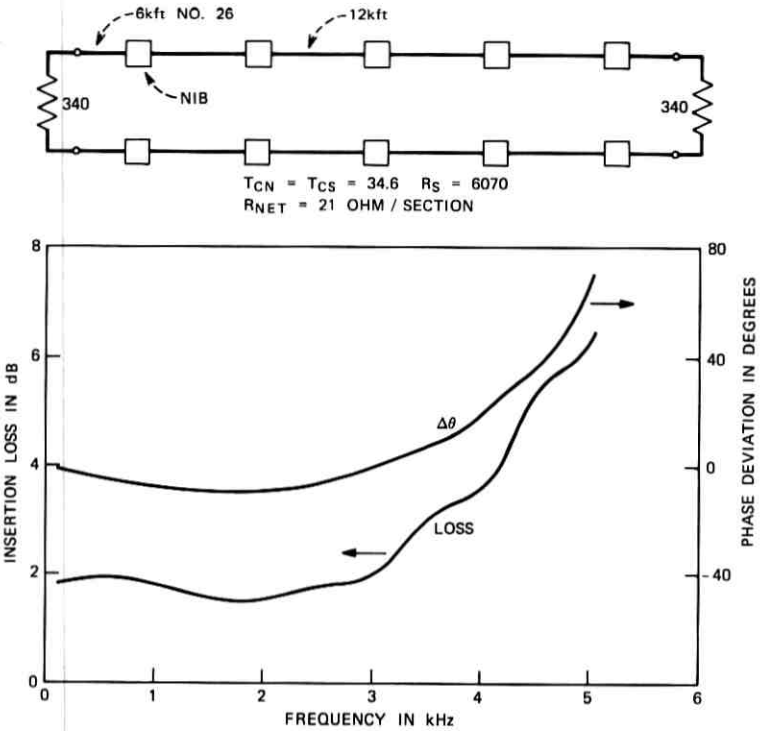



Fig. 10—Insertion loss (in dB and deviation of phase angle from linearity) of a five-section NIB line terminated with resistances.

An example of this effect and of the performance of a NIB line of finite length having resistance terminations, which are simple approximations to the image impedance but not the best, is given in Fig. 10. It should be noticed that the spacing of NIB units and, hence, the frequency band for this line are not the same as in Figs. 6, 7, 8, and 9. Where it is desired to have the return loss as high as possible, a better approximation to the image impedance should be made. See Section 6.2 for a possible method.

The ripples in the transmission curves are considerably larger when the line ends with half-NIB units rather than half-line sections as in Fig. 10.

### *5.8 Limit of gain*

Another general property of boosted lines, in common with all circuits to which negative resistance or gain is added, is the possibility

of self-oscillation, or instability. While, under certain conditions of termination, a small amount of insertion gain may be introduced by a boosted line, in most applications, the boosters should go no further than to cancel, or nearly cancel, the line losses in order to maintain stability, as suggested in Section IV. How closely this limit can be approached depends largely on the line terminations and on the margin against instability in the infinite line used in the choice of booster parameters. A similar limitation exists for a uniform line equipped with unilateral amplifiers when provision is made for two-way transmission. A net loss of zero can be approached as all reflections from imperfect terminations approach zero. Of course, a one-way transmission line can, in principle, have as high a gain as desired if terminations are perfect and there are no return paths for the signal. In some ways the NIB line is like the transmission portion of a four-wire repeatered line in which the overall net loss is near zero. The complex subject of stability of NIB lines will be dealt with in more detail in Section 5.10, below.

#### 5.9 Restriction on the shape of negative resistance vs frequency

There is a further restriction on finite lumps of negative resistance used as boosters at finite spacings in a transmission line. If the total amount of negative resistance added to a line approximately cancels the series resistance of the line wires at low frequencies, then the magnitudes of the negative resistances must be reduced at higher frequencies to avoid self-oscillation. This is a consequence of the wide range of impedances that a finite length of transmission line can present for various frequencies and terminations. How the restriction arises will be seen in the next part. If the reduction of negative resistance at higher frequencies is made with careful consideration, the transmission characteristics can be shaped in a number of desirable ways, and stability is achieved. This will be discussed below in the sections on particular realization. Incidently, another consequence of the wide range of line impedances is that a series negative resistance, often referred to as open-circuit stable, can cause instability in an open-circuited transmission line.

#### 5.10 Stability in NIB lines

When lumped negative impedances are inserted into a transmission line periodically, self-oscillation can occur; that is, the line can be unstable. While it is the presence of the negative impedances that brings about the possibility of instability, it is pointless to try to talk

about the stability of the negative impedances. What needs to be investigated is the whole circuit since, with the same negative impedance unit, stability will exist for some values of connected passive elements and instability for others.

#### 5.10.1 *Infinite line*

First consider the infinite line with uniformly spaced NIB units since this is simpler to deal with than the finite terminated line and also is the initial condition from which the latter is derived. The following material is presented as reasonable description rather than precise, rigorous analysis.

The boosted line is characterized by a propagation constant and image impedances. The two principal image impedances, as with coil-loaded lines, are the midsection one, $msZ_H$, and the midbooster one, $mbZ_H$, as indicated in Fig. 11. Calculation and measurement of boosted lines show that when the negative resistances nearly cancel the line resistance, $msZ_H$ becomes very large and $mbZ_H$ becomes very small near the frequency for which the spacing between NIB units is one-half wavelength. This is illustrated in Fig. 12. The line impedance will have values intermediate between these extremes at other points of the section. Also, all sections will be alike in this because of the assumed uniformity. Since the midbooster image impedance can become very small, it appears to be a likely parameter to consider in investigating the stability of the line. To carry this out, imagine that, as indicated in Fig. 11, a booster has been divided into two equal parts and separated so that a generator $e$ can be inserted into the line there. The current that flows at this point will be

$$i = e/(2mbZ_H).$$

This would be the same at any one of the boosters because of the

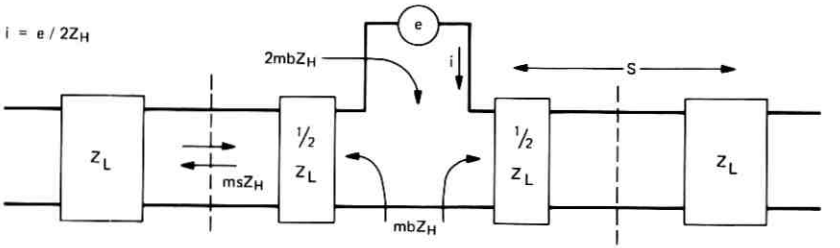



Fig. 11—Representation of NIB line for studying stability. $S$ is space between NIB units of impedance $Z_L$ as in Fig. 5.

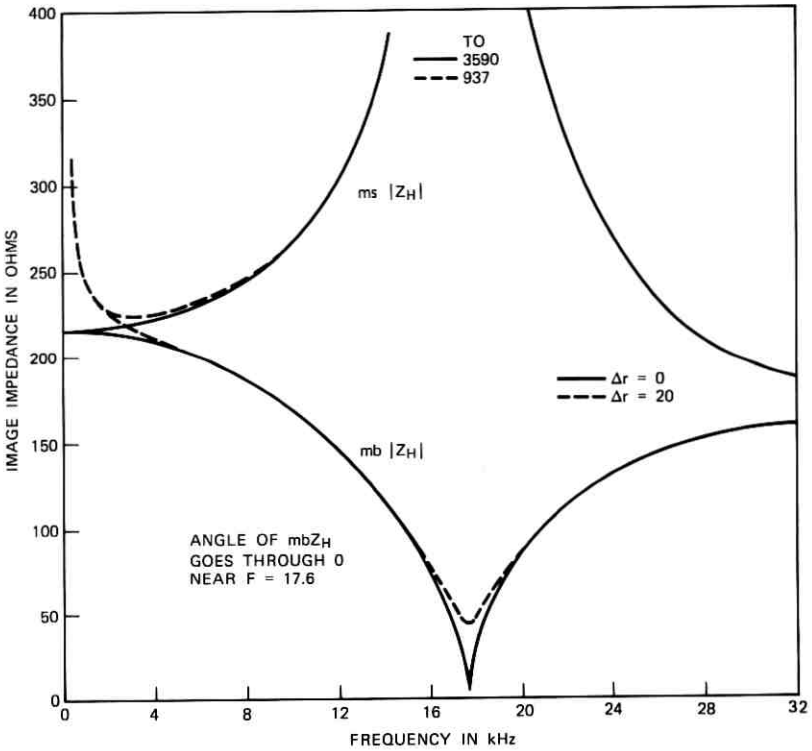



Fig. 12—The two image impedances $Z_H$ of a NIB line.

uniformity. As shown in Section 5.5 above, the midbooster image impedance may be calculated from the formula

$$mbZ_H = \tfrac{1}{2}\sqrt{Z_A Z_B},$$

where

$$Z_A = 2Z_{sc} + Z_L$$
$$Z_B = 2Z_{oc} + Z_L,$$

and where $Z_{sc}$ and $Z_{oc}$ are the impedances of one-half section of line that has been short-circuited then open-circuited at the far end, and $Z_L$ is the impedance of the NIB unit. Thus, $Z_A$ is the impedance seen looking in at the test point of Fig. 11 when the two adjacent midsection points are shorted and $Z_B$ is the impedance seen when they are opened.

The irrational nature of $Z_H$ makes it difficult to analyze in the usual way. However, if $Z_H$ becomes zero under certain circumstances, we can be fairly sure that instability will accompany those circumstances. First consider separately $Z_A$ and $Z_B$ as plotted in Figs. 13 and 14 from calculations of a particular NIB line. In each of these is a separate enlarged plot of the region near $Z_A = 0$ and $Z_B = 0$. Only in these enlarged plots do the differences caused by giving $\Delta r$ the three values, $-1.0$, 0.0, and $+1.0$, show up. But these differences are very significant for the question of stability. Extensions of these plots for negative frequencies would yield mirror images of the curves actually plotted reflected about the resistance axis. In the case of $Z_A$, it is seen that the curve extends into quadrants II and III to the left of the origin, that is, into the region of negative resistance, at very low frequencies when $\Delta r < 0$, and otherwise stays in the region of positive resistance. The curve of $Z_B$ extends into quadrants II and III to the left of the origin for frequencies near 17.6 kHz when $\Delta r < 0$.

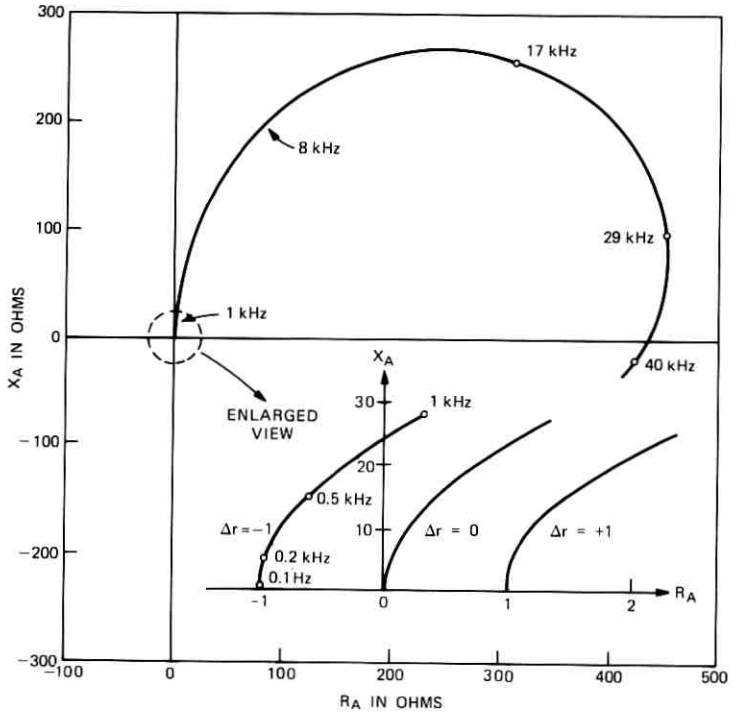



Fig. 13—Reactance vs resistance plot of impedance $Z_A$ as frequency varies. $Z_L$ is impedance of a pair of NIB units and $Z_{sc}$ is the open circuit impedance of a half section of uniform line.

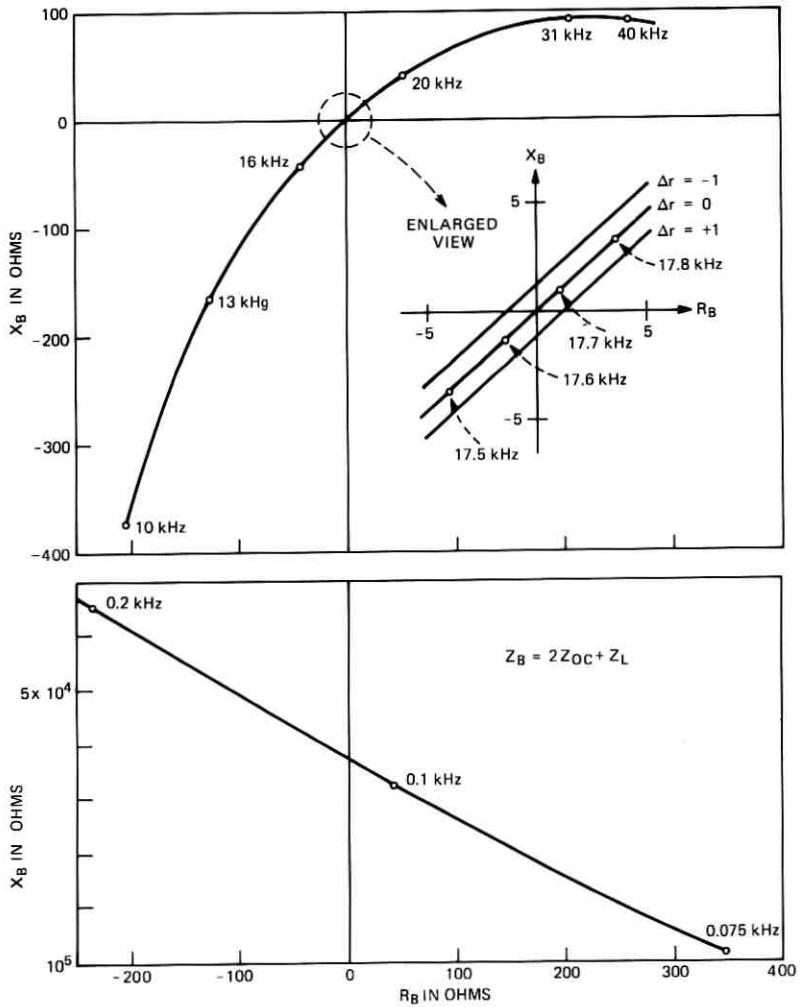



Fig. 14—Reactance vs resistance plot of impedance $Z_B$ in which $Z_{oc}$ is the open-circuit impedance of a half section of uniform line.

Thus, when $\Delta r < 0$, the product $Z_A Z_B$ can be zero at very low frequencies by means of $Z_A$ and at a considerably higher frequency by means of $Z_B$. The latter is responsible for the zero at 17.6 kHz in $mbZ_H$, shown in Fig. 12, and arises through a resonance of the reactance of the booster and that of the open-circuit section of line. It is impossible to show the effects of the zero in $Z_A$ on $Z_A Z_B$ in Fig. 12 because

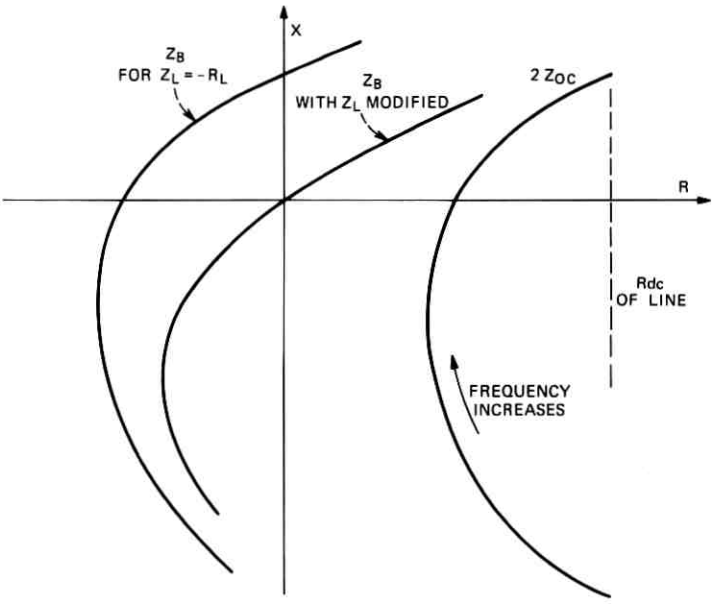



Fig. 15—Reactance vs resistance plots of $Z_B$ and $Z_{oc}$.

of the very low frequencies (see Fig. 13) at which they occur. Plots of the impedance $Z_H = \frac{1}{2}\sqrt{Z_A Z_B}$ similar to those of Figs. 13 and 14 touch the origin at a very low frequency or loop around it at the zero in $Z_B$ when $\Delta r < 0$. A lumped network, whose impedance may be expressed as a rational function of complex frequency $s$, would have a similar plot looping around or touching the origin if one of its roots occurred at a complex frequency with positive real part. In these circumstances the network is unstable. From this we may infer that the NIB line can be unstable in the two ways shown when $\Delta r < 0$. The way involving $Z_B$ depends on other circumstances as well. The booster parameters for Figs. 13 and 14 were chosen with respect to the line parameters so that a just-stable condition exists for $\Delta r = 0$. If these parameters were different, e.g., for a larger time constant in the booster, $\Delta r$ could be negative by a limited amount without instability at the higher frequency. There is no qualification about the zero through $Z_A$. This always arises in the infinite line for $\Delta r < 0$.

Now we may see why the lumps of negative resistance must be reduced with increasing frequency to avoid instability. Consider the $X$ vs $R$ diagram of Fig. 15 in which $Z_B$ and $2Z_{oc}$ are plotted. If the booster impedance $Z_L$ is a pure resistance, $-R_L$, which cancels

all the line resistance, the diagram for $Z_B$ would be that at the left which loops around the origin and would be unstable as we have seen. But if the negative resistance component of $Z_L$ is reduced in magnitude as frequency increases, we may obtain the $Z_B$ curve shown in the middle, which is just stable. $R_B$ must be positive where $X_B = 0$. Reducing $|R_L|$ for this purpose may be done in such a way as to also shape the transmission properties of the NIB section in some desirable way, as indicated above.

These conditions may be summarized by the diagrams of Fig. 16, which are sketches made from calculated values of $Z_A Z_B$ but are not plots of actual data. In the top diagram, two conditions, $\Delta r < 0$ and $\Delta r > 0$, are shown, but the booster capacitance is sufficiently large that a zero in $Z_B$ does not occur. In the bottom left diagram, the same two conditions are shown, but here the booster capacitance is sufficiently small that the zero in $Z_B$ occurs even when $\Delta r > 0$. In the bottom right diagram, no zero occurs even though $R_B$ may be negative at some frequencies. These diagrams may be interpreted as Nyquist diagrams for testing the stability of impedances,[23,25] as the encirclements of the

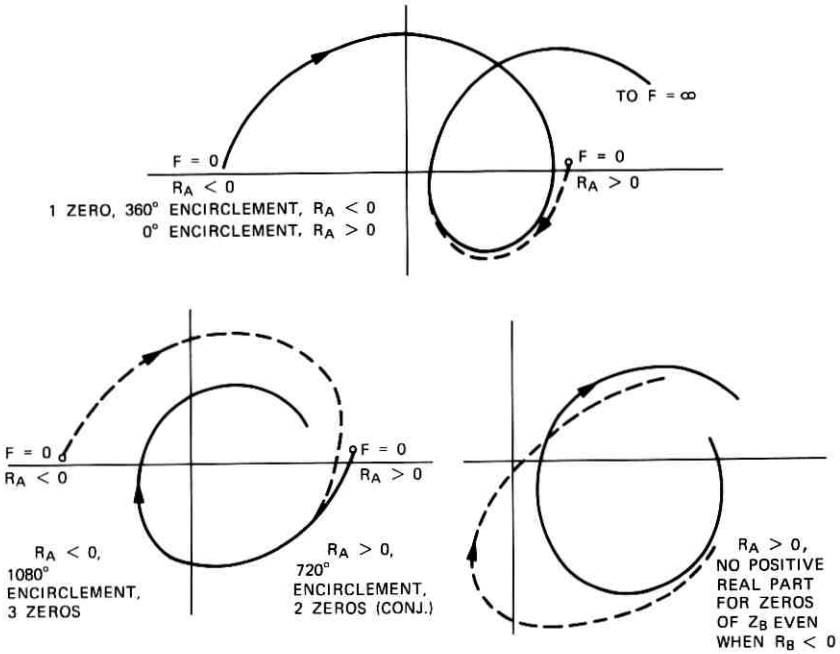



Fig. 16—Nyquist plots of $Z_H^{\lambda}$.

origin in the various cases agree or not with the rules for stability, as noted. Only the positive frequency parts of the plots are shown in the diagrams for simplicity.

It can be realized from Figs. 13 and 14 that whether or not a zero occurs in $Z_B$ depends on the details of the booster involved and cannot be determined in general circumstances.

While the above discussion of stability in the infinite NIB line is not a rigorous mathematical one, it is the result of much study, calculation, and experiment and is believed to be a good description of the situation. These conclusions are also in agreement with a computer investigation of the nature of the roots of the input impedance of a NIB section in which the line was approximated by a rational network and the section terminated in various ways.

#### *5.10.2 Finite line*

The stability of a NIB line consisting of a finite number of sections is considerably more complicated than the stability of the infinite line because the terminations (if different from $Z_H$) must be considered also. Because details of the configuration of line and booster, as well as the terminations, are important in these problems, they are more appropriately discussed in connection with particular arrangements. Two particular situations will be described briefly here.

The extreme condition for line stability with any passive terminations whatsoever requires that

$$(R_{11}^2 - R_{12}^2) > 0 \text{ for all frequencies.}$$

In this, $R_{11}$ and $R_{12}$ are the real parts of $Z_{11}$ and $Z_{12}$, the impedance parameters of the matrix for the cascade of NIB sections under consideration. This condition, first derived by Gewertz,[26] has also been derived by Llewellyn[27] and others. A simple neat derivation is given by Walter H. Ku.[28] To satisfy this condition, the net resistance, $\Delta r$, of line and booster must be fairly positive. Thus, what has sometimes been a goal in NIB work, namely, a line with zero net loss and stable for *all* terminations, is an impossibility. Of course, we cannot expect a line equipped with 22-type repeaters to be stable at zero net loss for all terminations either.

On the other hand, when resistance terminations near the low-frequency magnitude of $Z_H$ are used, the line can be stable even when the net resistance, $\Delta r$, is negative. This is illustrated in Fig. 17, which shows some stability relations for a four-section NIB line. A number of other configurations and numbers of sections have been studied, but

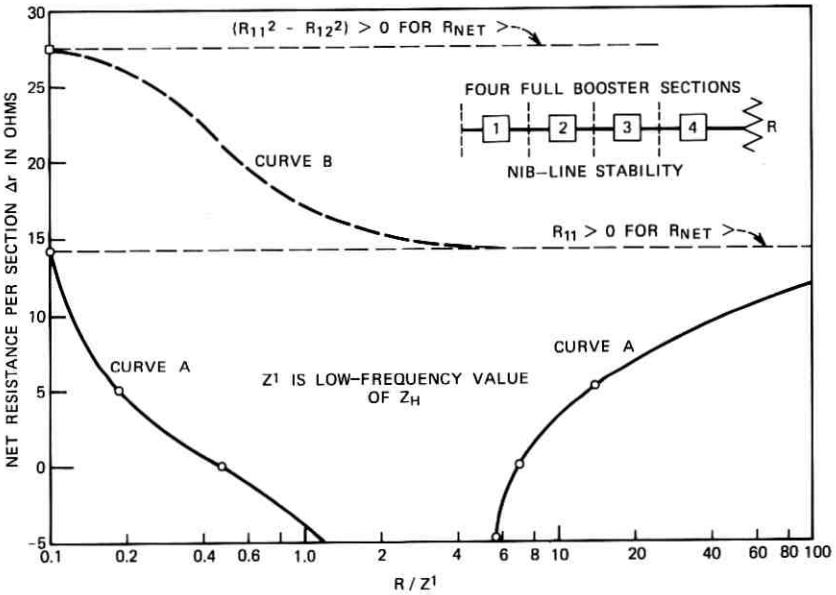



Fig. 17—Boundary curves between stability and potential instability in finite NIB lines terminated with resistance.

only this one is presented here. In this figure, the ordinate is the net resistance $\Delta r$ per section of line, as already discussed, and the abscissa is the resistance $R$ (normalized) which terminates the line. The two parts of the solid curve $A$ would meet if more negative values of $\Delta r$ had been plotted. For values of $\Delta r$ and $R$ that lie above this curve, the input resistance of the line is positive at all frequencies; below it the input resistance is negative at some frequency. This curve is, thus, a boundary between stability above and potential instability below. Instability will exist in fact if, in addition, the input reactance is zero, or if it is cancelled by external reactance at the frequency for which the input resistance is negative. The dashed line at the top of the figure indicates the value of $\Delta r$ necessary for the line to be stable for any termination. The worst termination is a reactance. If some resistance $R$ is added to this worst reactance, the dashed boundary curve $B$ is obtained, separating stability (above) from potential instability (below). Above the dashed line near the middle of the figure, the NIB line is stable for any resistance termination.

While the curves of Fig. 17 were obtained by means of a rather elaborate computing process, they were verified on a laboratory setup.

Additional verification was obtained from a different computing process. The two-line sections of a single-section NIB line were approximated by lumped-element networks. Then the roots of the characteristic equation of the whole NIB line section, including termination, were found. The conditions of instability or stability, as determined from the nature of these roots, was in agreement with those corresponding to Fig. 17.

It is seen from this figure that, for a small range of terminating resistance, stability prevails even for negative values of the net resistance $\Delta r$. However, as the number of sections in the line increases, the amount of negative $\Delta r$ that can be tolerated becomes smaller, approaching none for the infinite line, as already determined. Actually, for a line of 16 sections, $\Delta r$ cannot be less than zero.

The boundary curve A of Fig. 17 is one section of a boundary surface over the half-plane of a general terminating impedance.

It should be noted that the $\Delta r = 0$ dividing line discussed here is a result of other NIB parameters having been chosen so that line attenuation is zero and as flat as possible over the appropriate frequency band. If these other NIB parameters had been chosen in other ways, the dividing line could have been at some positive or negative value of $\Delta r$.

The NIB configuration in which the line ends with half-NIB units instead of half-line sections requires a considerably higher value of $\Delta r$ for the same degree of stability.

#### *5.11 Determining the booster impedance $Z_L$*

Two methods that have been used in determining the required booster impedance $Z_L$ are described briefly.

The first method was originally described by K. Bullington,[23] and is a calculation made to determine the booster impedance $Z_L$ that will give some desired propagation factor $P$. This may be done using the $A_{11}$ matrix coefficient for one NIB section given in Section 5.5, above. If the two forms of this coefficient are equated and solved for $Z_L$, we get

$$Z_L = 2(Z_o/\sinh 2p)(\cosh P - \cosh 2p),$$

where $2p$ and $P$ are the propagation factors for one section of plain line and boosted line, respectively.

Then a network which approximates the computed impedance $Z_L$ as closely as possible is devised. This may be done formally, by using a "negative-impedance converter," as indicated by John Linvill[17] and others, to convert a positive impedance, $-Z_L$, into the negative im-

pedance, $Z_L$. Or it may happen that a large part of $Z_L$ is supplied by some special device and only an additional "trimming" network is needed. Small differences between the real network and the calculated $Z_L$ can result in stability problems. These should be investigated with the real network.

The second method is to take a particular network configuration having a negative resistance and vary its parameters and/or add some trimming elements to get desired transmission properties and stability for the resulting NIB line.

The calculated performance curves shown in Figs. 6, 7, 8, and 9 were obtained from NIB sections derived by the second method using the booster circuit (Fig. 23) devised by L. A. Meacham.[18] Two of these curves are replotted in Fig. 18.

To compare with these, one example using the first method is given also. For this, the third-degree maximally flat characteristic was taken as the desired transmission curve. If this is expressed in complex frequency form, i.e.,

$$Y = 1 + 2s + 2s^2 + s^3,$$

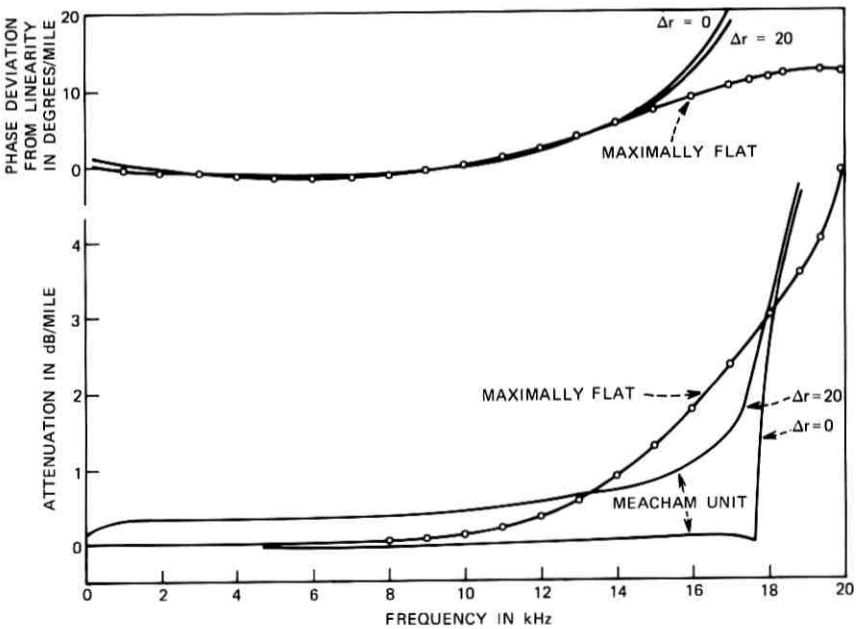



Fig. 18—Attenuation and phase curves resulting from two forms of NIB units in NIB lines.

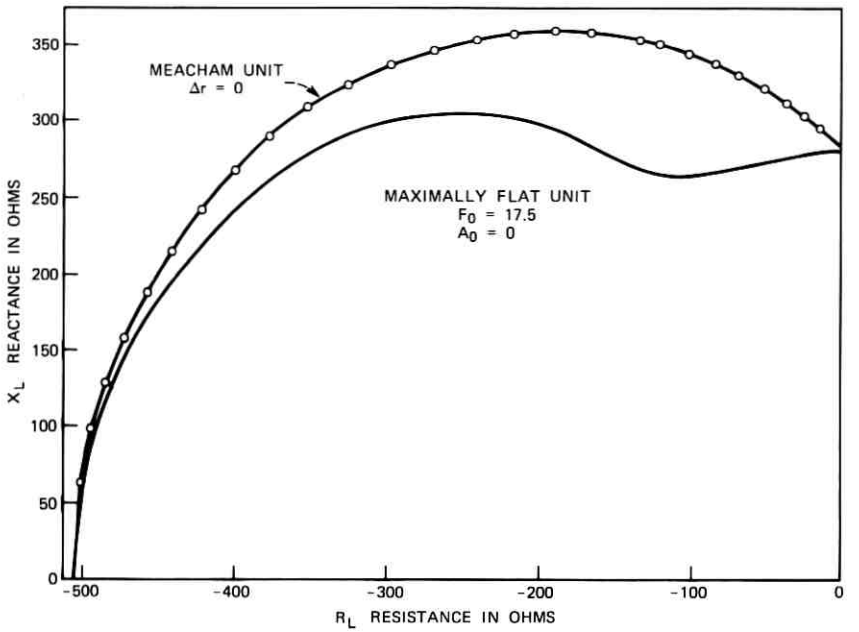



Fig. 19—Reactance vs resistance plots of the two forms of NIB units used in Fig. 18.

then compatible amplitude and phase curves are realized. The transmission curves for the NIB section designed in this way are shown in Fig. 18, designated maximally flat.

Curves, $X_L$ vs $R_L$, describing the negative impedances $Z_L$, derived in the two ways, are plotted in Fig. 19.

A good computer program is essential to the successful use of either method. How these two methods have been used in actual work will be mentioned in the section on history, which follows.

#### *5.12 Pulse transmission*

In the transmission curves of Fig. 18 for the Meacham unit, the booster parameters were chosen so that the attenuation curve of the NIB line would be as flat as possible. The parameters can also be chosen so that the phase curve of the line is as linear as possible. This makes for a minimum of distortion in pulse transmission as indicated in oscillograms shown in Ref. 18. Photographs of pulse waveforms received through NIB lines formed in both of these ways are shown

in Fig. 24. NIB parameters were chosen for flat attenuation for Fig. 24a and b and for the linear phase in Fig. 24c and d.

### 5.13 Temperature compensation

In most transmission lines, it is desirable that compensation be made for the variation in attenuation caused by temperature variations. Several methods have been tried. When the NIB unit is the Meacham one, the compensation may be accomplished quite readily in many situations, as mentioned in Ref. 18, by using a temperature-sensitive resistor as a part of one of the NIB resistors. This method is particularly simple and effective for underground cable when the temperature-sensitive resistor can be underground also at essentially the same temperature as the cable.

### 5.14 Interference susceptibility

One problem which may arise with the use of NIB lines, especially when series units are used, is that of interference currents. Even if the line is well balanced so that longitudinal-induced interfering currents do not appear in the metallic path, they must pass through the negative-impedance units. If these currents are large compared with the signal current, they may carry the total current beyond the negative-impedance region of the unit and so cause distortion in the signal (see Fig. 2). To avoid this, a much larger current capacity in the NIB unit than would be necessary for the signal alone must be provided, or operation in a region of high interference must be avoided.

## VI. HISTORY OF NEGATIVE-IMPEDANCE BOOSTING

### 6.1 Crisson

It was mentioned above that the possibility of using negative-resistance boosters to reduce the attenuation of transmission lines was recognized before 1919.[4] But the first theoretical and experimental consideration of the use of such boosters in telephone lines was described by Crisson in his 1931 paper in *The Bell System Technical Journal*.[12] However, it is believed that this consideration was of the occasional or single addition of a negative-resistance unit to a line rather than of their periodic addition.[23]

### 6.2 Bullington and Edwards

The first appearance of the concept of introducing lumps of negative resistance periodically into a line to obtain a new "uniform" line with

desirable transmission properties seems to be in the work of K. Bullington[23] in 1940 and 1941. This effort was stimulated partly by the availability of a quite small thermistor unit which had a negative-resistance region in its voltage-current curve in the audio-frequency range. At that time, it was recognized that the use of a net negative resistance at all frequencies would cause instability. However, it was also thought that if each negative-impedance section could meet the Nyquist stability requirement for impedances (see Refs. 23 and 25, and Fig. 19), any desired number of such sections could be connected in tandem with overall stability. While these point contact units were mechanically fragile, a stable net gain of 2 to 3 dB over the frequency range of 200 to over 3000 Hz was demonstrated on a 54,000-foot section of 19-gauge cable pair. Five negative-impedance series-type units were inserted at 12,000-foot intervals with 3000-foot end sections. The application intended was for voice-frequency toll circuits. Considerable analyses were made[23] that are quite complete and very readable. The discussion covers several of the general properties of Section V, particularly Sections 5.4, 5.5, 5.6, 5.7, and 5.10, that is, reflection effects and cutoff, propagation formulas, image impedance, and stability. The first method (see Section 5.11) of choosing the booster impedance $Z_L$ was used, assuming zero loss and linear phase in the transmission band for a start. The thermistor impedance provided a negative resistance that decreased with frequency and also a positive reactance. Both of these were a major part of the desired $Z_L$. The remainder was made up by a simple added network.

An interesting suggestion, which does not seem to have been attempted, was made by Bullington for the termination of finite NIB lines. His calculations indicated that if the line were ended with 0.7-line sections instead of 0.5-line sections, the correct terminating impedance would be a very nearly constant resistance plus an inductance, which would be simpler to construct than the image impedance, particularly near the cutoff frequency. This particular project was ended by the more pressing demands of World War II.

#### *6.3 Merrill and the E repeater*

The next stage of effort produced the first E-type negative impedance repeaters, in the development of which the work of J. L. Merrill, Jr.[19,20] occupies a large part. Merrill had done some work before the war in pursuing the work begun by Crisson. As a part of this effort, the analysis of K. G. Van Wynen, referred to above,[24] dealt with the impedance and transmission properties of networks equivalent to the

boosted sections. While the use of an E repeater introduces a negative resistance in series with the line to reduce its loss, the concept of use is quite different from that of a line with a considerable number of negative-impedance units added periodically along its length. The E repeaters are used only in central offices and usually with only one or two in a given line. Hence, they disturb the uniformity of the line and cause reflections. However, the latest versions of these repeaters have been installed at a great rate and are filling a real need in making gain available for the trunks of the exchange plant. One of the important factors in the simplicity of the use of E repeaters is that the dc continuity of the circuit is maintained and so dc and very-low-frequency signaling currents go through without rearrangement or new apparatus. Also, if the repeater fails, the unrepeatered line is still available. How this comes about may be seen from Fig. 20 which shows a block diagram of the E1 series repeater and how it is connected. The negative impedance is coupled into both conductors of the line by means of the transformer. The negative impedance is generated by a device called a negative-impedance converter which, when a passive impedance $Z_n$ is connected to two of its terminals, presents an impedance of approximately $-Z_n$ at its other two terminals. The "simplicity" factor mentioned above emphasizes the fact that one of the principal differences between conventional amplifying and the use of negative resistance lies in the manner in which the source of energy replenishment for the signal is coupled into the line.

The first model of the E repeaters, the E1, was introduced in 1948.[19] The converter used a twin triode vacuum tube with positive feedback coupling, and the $Z_n$ network could be adjusted for various amounts of negative impedance, or loss reduction. The repeater can be used in coil-loaded lines or nonloaded lines.

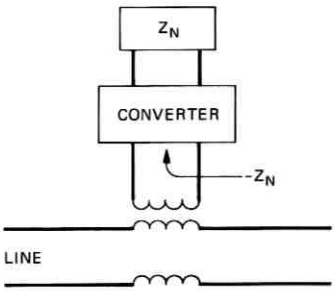



Fig. 20—Method of coupling E-type negative-impedance repeater into transmission line.

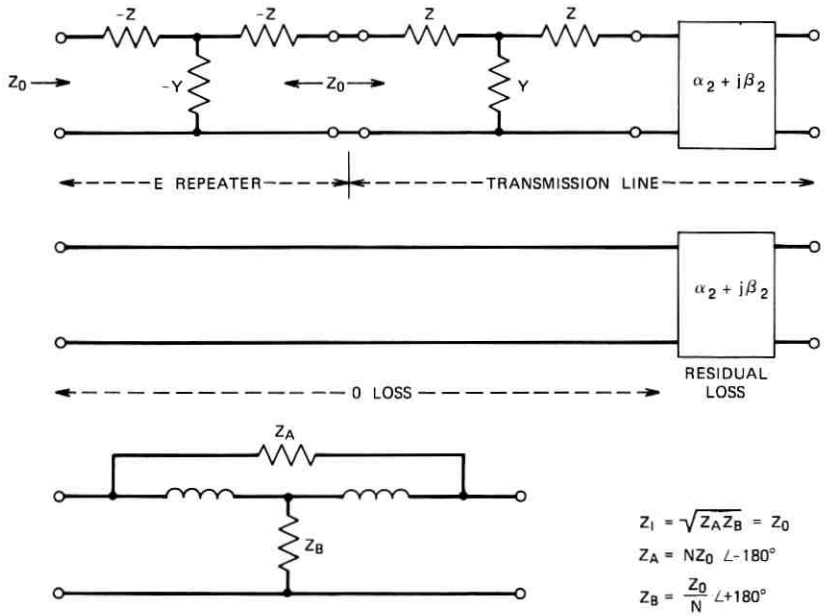



Fig. 21—Equivalent circuits for E-type negative-impedance repeaters.

As discussed above, the introduction of these impedance lumps into a line causes reflections or echoes when the same line $Z_o$ termination is retained. Hence, the principal use of E1 repeaters has been in interoffice trunks rather than in toll trunks, where, because of the longer delays involved, echoes can be much more objectionable. Where there is only one repeater in the line, as in the majority of these cases, the amount of gain and the amount of reflection are directly proportional to the magnitude of the negative resistance used.

As mentioned above, if series and shunt negative impedances of appropriate magnitude are combined properly, not only will losses be reduced, but the characteristic impedance of the line may be matched approximately. This was done in the E23 repeater[20] to obtain much smaller reflections than those that accompany the E1 repeater, and so be usable in toll-connecting trunks. This process is indicated in the diagrams of Fig. 21 taken from Ref. 20. It shows how the negative-impedance repeater cancels a part of the line loss leaving a residual propagation factor of $\alpha_2 + j\beta_2$, while also matching the characteristic impedance $Z_o$ of the line. It also shows that by means of the bridged-T structure, only two instead of three negative-impedance units need be

built, and how the two are proportioned. The coils provide coupling into the line as well as being part of the bridged-T structure.

A recent version of the E repeaters, the E6, is like the E23 in that it matches the line, but it is different in some other respects.[21] It was introduced about 1960. It uses transistors instead of vacuum tubes in the basic converter circuit devised by R. L. Wallace and J. G. Linvill.[17] In the physical embodiment, the amount of negative resistance, or gain, adjustment is separated from the line impedance-matching function to simplify the operation of the repeaters.

The question may be asked, since two amplifiers are used to generate the two negative impedances, why not use them to make a 22-type repeater? There seems to be no simple direct answer to this question. In some situations or at some times, one of the methods may provide an advantage over the other and vice versa. For example, when a simple way of providing continuity for dc signaling along with gain is needed, there is an advantage in using a negative-impedance repeater. The advantage may lie with the 22-type repeater in other circumstances.

The E repeaters provide a simple way to add small amounts of gain to exchange trunks. But in their present conception, they do not provide the broad possibilities that can be envisioned with the use of periodically added negative impedances to obtain various desirable transmission characteristics.

#### *6.4 Schott and Wallace*

During the period 1953 to 1956, L. O. Schott did some design and experimental work on NIB lines. Two particular situations were considered. In the first, six negative-impedance units were spaced one-third mile apart in 22-gauge cable to give a signal band from about 7 kHz to 100 kHz for possible carrier applications. The negative-impedance unit shown in Fig. 22a was similar in form to the E1 repeater, but with the converter using the transistor circuit of Linvill.[17]

A considerably different negative-impedance-unit circuit, Fig. 22b, suggested by R. L. Wallace, was used in the second situation. Instead of the two like transistors used earlier, one npn, and one pnp were used in the new arrangement. With this change, the coupling transformer was eliminated and the current for powering the unit allowed to flow through the line and units in series. While this requires two units, one for each conductor, it brings a great simplification to the circuit with the opportunity for miniaturization. These units were applied to six 3900-foot sections of 32-gauge cable, providing a transmitting band

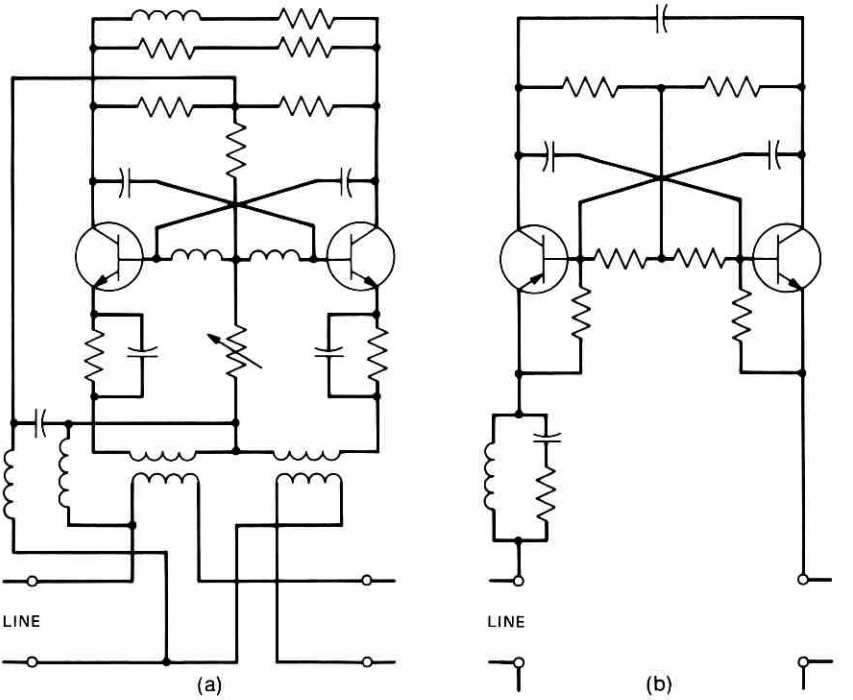



Fig. 22—Circuit diagrams for two negative-impedance converters: (a) Linvill circuit, and (b) Wallace-Schott circuit.

from 0.1 kHz to 7.5 kHz. The *RLC* network in series with the unit was added to provide a margin against instability.

#### *6.5 Hoth and Ross*

During the period from about 1954 to 1957, a study in considerable detail was made by D. F. Hoth and W. L. Ross concerning the possibility of using negative-impedance boosting to provide gain in exchange trunks. Their main conclusion was that, though their complete experimental NIB system operated successfully and appeared to offer small cost savings, the E repeaters for voice and the T1 system for carrier seemed to hold more promise in the exchange trunk area. This conclusion appears to have been influenced, in part at least, by some unsatisfactory aspects of the negative-impedance unit used.

This unit was an avalanche transistor with one resistor and one capacitor connected externally. The negative resistance is generated by the avalanche discharge process within the transistor and is the

series type. Some of the unsatisfactory properties were lack of uniformity of the transistors, nonlinearity of the negative resistance developed, and comparatively high voltage required at each unit.

A real system was set up consisting of three 9000-foot sections of 26-gauge cable, with provision being made for signaling, powering, and matching of the line to office impedance. The line had an image impedance of about 300 ohms and a sharp cutoff near 8 kHz. The low image impedance is a distinct advantage for reduced crosstalk. Regulation of net loss was by means of a 40-Hz pilot tone. The change of amplitude of this tone in response to cable-temperature change was used to shift the dc line current, which in turn would change the amount of negative resistance because of the nonlinearity of the NIB units.

The shape of the booster impedance $Z_L$ was chosen by the first method described in Section 5.11 and the actual unit made to approximate this fairly well over the transmission band. Instead of giving each NIB section sufficient loss so that the system would be stable with any passive termination, they designed the sections for zero net loss and then put terminating networks at each end so that the system was stable for any passive termination beyond these networks. This is equivalent to restricting the termination of the actual NIB sections to lie in a certain impedance range.

Considerable analysis was done on the problem of stability, particularly on the conditions required for stability with any passive termination.

#### *6.6 Other projects and general comments*

While there have been quite a number of projects, beyond these listed above, in which the use of negative impedances in transmission lines has been studied, the list is representative except for the work of L. A. Meacham dealt with separately in Section VII.

A handicap in many of the projects was the goal of a line with zero net loss which at the same time could tolerate terminations of any passive impedance whatsoever. We have seen that this is an impossible goal. The closer we approach zero net loss, the more restricted the range of terminating impedance becomes until, at zero, termination must be perfect except in the few special situations indicated in Section 5.10.2. But similar restrictions apply to other ways of reducing line loss also. It was shown clearly in the early years of repeatered lines[8–10] that irregularities in the line and imperfect terminations severely limited the amount of gain that could be used in a line. A paper, "Some Fundamental Properties of Transmission Systems,"[27] was written by F. B. Llewellyn in part to show that the restrictions that are necessary for

transmission circuits to be stable are essentially the same whether unilateral amplifiers or negative impedances are used to reduce the losses.

## VII. MEACHAM'S WORK

In 1963, a new look at the use of negative-resistance lumps placed uniformly in a line was begun.

### *7.1 Meacham NIB unit*

The point of departure here was a fairly simple realization of a negative-resistance circuit. The generation of negative resistance was not accomplished in a single device, as in some previous situations, but by the cross coupling of two transistors in a simple way. Though derived independently, it is similar to the Wallace circuit of Fig. 22b in that it uses an npn and a pnp transistor. But the Meacham circuit does not use coupling capacitors. Two advantages come from this: the negative resistance is usable down to dc, and the physical bulk of the capacitors is eliminated. The circuit could be realized also by a single pnpn device. Later in the investigation, the negative resistance was made highly linear by the addition of two diodes, and the transmission shaping possibilities were enlarged by adding a resistance and capacitance in shunt.[18] The Meacham negative-impedance circuit and its voltage-current characteristic are shown in Fig. 23. These negative impedances were used for the transmission curves of Figs. 6, 7, 8, and 9. And while these were plotted from calculated data, measured performances of the real physical lines are always in very good agreement with those calculated. Some work was done to demonstrate that the NIB units could be built in very small integrated-circuit form.

### *7.2 Field tests*

Two field tests were successfully completed and numerous laboratory demonstrations made. The lines constructed for these tests provided excellent transmission with low phase and nonlinear distortion for a variety of signals. Lines in both tests were about 32 miles long. One was a single circuit in a suburban exchange area, part aerial cable and part underground cable.[18] The other test involved three circuits using three order wire pairs in the L4 spur cable from Netcong to Newark, New Jersey. Extensive measurements of these were made over a period of about a year. They were operated at a net loss of about 3 dB and had very low noise and a satisfactory return loss. The longitudinal interfering currents were negligible. Signals from a 202D Data Set system (frequency-shift keying at 1200 bits/second) with its

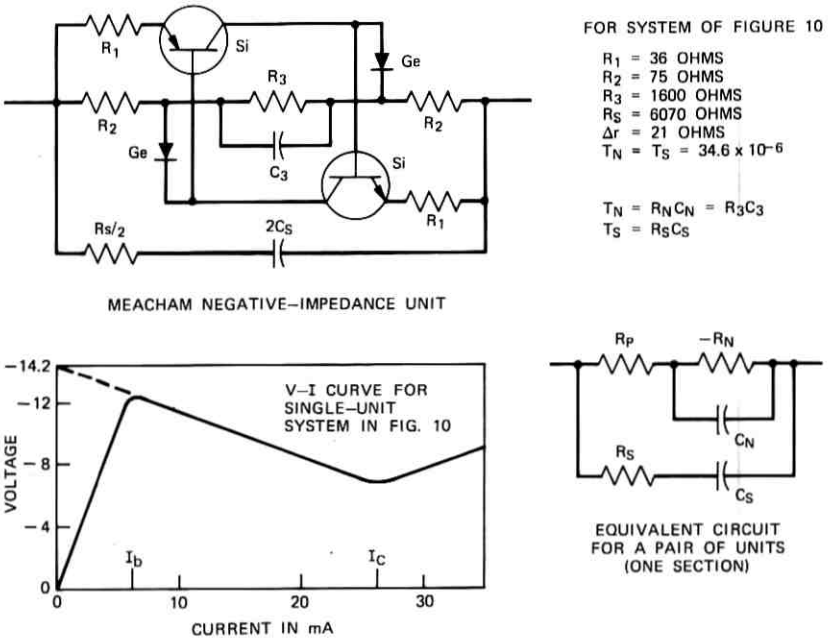



Fig. 23—Circuit diagram for Meacham negative-impedance unit with equivalent circuit and voltage-current curve.

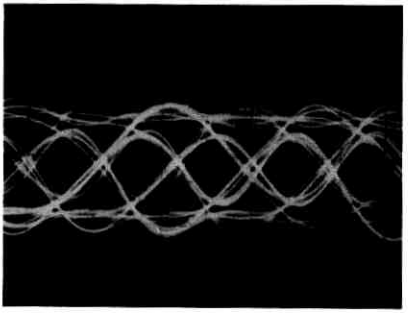
equalizers removed were sent through one of these lines having a bandwidth of about 5 kHz with no noticeable distortion (see Fig. 24a). With a slight rearrangement at the terminals, baseband rectangular binary pulse signals at a 12-kHz rate were easily transmitted, as may be seen from the eye diagram of Fig. 24b. These lines were designed for essentially flat attenuation curves. If the NIB unit parameters are chosen to have the most linear phase curve instead, pulse transmission is even better. This is shown in Figs. 24c and 24d for such a line consisting of nine 6000-foot sections. Figure 24c is the response to a single 50-$\mu$s rectangular pulse and Fig. 24d is the response to a pseudo-random-pulse train at 14 Kbauds and 16 levels.

Suitable means were worked out for precise measuring of the NIB units and for measuring the overall net resistance of the operating line.

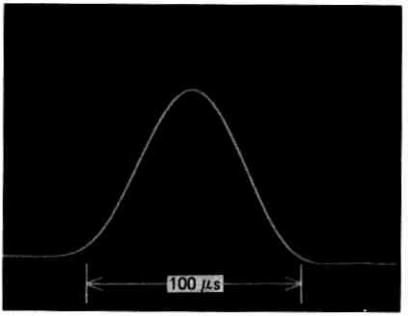
### *7.3 Papers published*

Two papers have been published on parts of the work, one by L. A. Meacham[18] and the other by A. L. Hopper.[22] The purposes of the present paper are: (*i*) to present a comprehensive general view of the

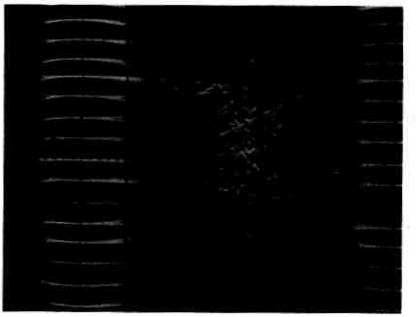
(a) 202D DATA SET SOURCE
EQUALIZERS REMOVED 1.2 kBIT/s



(b) BASEBAND TRANSMISSION -
12 kHz BINARY




(c) RESPONSE TO SINGLE 50 μs
RECTANGULAR PULSE



(d) RESPONSE TO PSEUDO-RANDOM
TRAIN — 14 kBAUD,16 LEVELS

Fig. 24—Pulse transmission in NIB lines. Parameters chosen for flat attenuation in (a) and (b) and for linear phase in (c) and (d).

properties of NIB lines (Section V), (*ii*) to describe the earlier work on negative-impedance boosting, and (*iii*) to present additional results not given in the two previous papers, notably the work on stability outlined in Section 5.10. The author of the present paper began work with this project late in 1968.

The paper by Meacham is a general one describing the negative-impedance unit, how it is used in the line, and the improved transmission which results in various situations. The one by Hopper is on the design of NIB lines following the viewpoint mentioned above, namely, varying the parameters of the NIB unit to achieve desired transmission properties in a NIB line made up of these units added periodically to a particular uniform line.

In all the work on this project, the approach was not to try to build a NIB unit to have a certain impedance $Z_L$, but to vary the parameters of this particular NIB circuit configuration to obtain certain desirable

transmission properties of the combination of transmission line plus booster. This provides advantages as well as restrictions.

### 7.4 Design of NIB lines

The use of the high-speed digital computer was essential in carrying out this process. Details of this method and design results for a number of uniform lines and various bandwidths are given in Hopper's paper. The curves of Fig. 4 on the relation between spacing and bandwidth are one result of this work. These designs are for zero net loss and are just stable. Some backing away from this condition is necessary for practical operation. The transmission objective in most of these examples is as flat an attenuation curve as possible. However, it is possible to make a somewhat different choice of parameters to get as nearly linear phase curve as possible. Some work, not published, was also done in choosing parameters so that some measure of distortion in pulse transmission would be a minimum. In this work of designing NIB lines, the two NIB units in the two wires of one section of line are represented by the equivalent circuit shown in Fig. 23.

### 7.5 Choosing element values for the NIB unit

The designs for the lines are expressed in terms of the parameters of the equivalent circuit. How these are translated into the element values of the real physical circuit (see Fig. 23) is described in unpublished notes by D. M. Chapin and is outlined below.

An important parameter of the Meacham NIB unit is the voltage $V_K$, which is the difference between the emitter-base voltage for a transistor and the voltage drop across its associated diode. The diodes were added to counteract the nonlinearity of the transistor emitter-base junction. Most of the units were built with silicon transistors and germanium diodes that provide a value of about 0.5 V for $V_K$. For smaller values of $V_K$, the current range over which the negative resistance appears is less, and disappears for $V_K = 0$. The circuit will operate as a negative resistance without the diodes, but will not be as linear.

The first element to be chosen is the resistor $R_2$. There are three ways this may be chosen, but the most direct is the following:

$$R_2 = \frac{\alpha}{2\alpha - 1}\frac{R_p}{2} + \frac{V_K}{(2\alpha - 1)\mathrm{D}},$$

where $\alpha$ is the $\alpha$ of the transistor, $V_K$ the parameter mentioned above, $R_p$ is the positive resistance in the equivalent circuit, and $D$ is the

current range of the negative resistance $I_c - I_b$ in the V-I curve of Fig. 23. The parameter $R_p$ is

$$R_p = \frac{4R_1R_2}{R_1 + R_2},$$

and, as mentioned above, applies to the two units, whereas the element values are being derived for a single unit. From studies of the NIB line process, a very good value for $R_p$ which is used in most designs is

$$R_p = 100 \text{ ohms}.$$

The above choice for $R_2$ is nearly the same as that which would have been obtained to minimize the total dc voltage drop across units and line per section. Or it could have been done by

$$R_2 = V_K/I_b,$$

where $I_b$ is the value of current at which the negative resistance begins in the V-I curve.

Then, $R_1$ is determined by

$$R_1 = \frac{R_2R_p}{4R_2 - R_p}.$$

It may be shown that

$$\frac{R_n}{2} = R_3 \frac{(2\alpha - 1)R_2 - R_1}{R_1 + R_2}$$

and

$$C_n = \frac{C_3}{2} \frac{R_1 + R_2}{[(2\alpha - 1)R_2 - R_1]}.$$

If $R_L$ is the total (both conductors) resistance of the nonboosted line per section, we have $\Delta r = R_p - R_n + R_L$ with which to determine the required $R_n$, where $\Delta r$ is the net low-frequency resistance per section as used in Section 5.10. From these we have for $R_3$

$$R_3 = \frac{(R_p + R_L - \Delta r)(R_1 + R_2)}{2[(2\alpha - 1)R_2 - R_1]}.$$

If the circuit is operated without the diodes, $V_K$ is replaced by $V_{eb}$.

The choice of $R_s$ and $C_s$ is discussed in Hopper's paper, Ref. 22. This completes the translation of the equivalent-circuit parameters into the elements of the real NIB circuit.

To keep the net resistance $\Delta r$ close to its desired value, it is necessary that negative resistance $(R_n - R_p)/2$ be adjusted quite precisely.

The simplest way to do this is to build the units with elements of reasonable tolerances except for $R_3$. Then, with a standard variable resistor connected in place of $R_3$, this variable resistor is adjusted until the measured value of $(R_n - R_p)/2$ reaches its desired value. The variable resistor is then replaced by a fixed one of this value.

The measurement is simply made by superimposing a small low-frequency current on the bias current, which is $(I_c + I_b)/2$. A standard variable positive resistance in series with the NIB unit is then adjusted for a null in voltage across the combination which indicates equality of positive and negative resistance.

### *7.6 Temperature compensation*

Compensation for temperature variations in the lines of the two field tests was made by replacing part of the fixed resistor $R_3$ in Fig. 23 by a resistor wound with a nickel alloy having a suitable positive temperature coefficient. Thus, the negative resistance of the NIB unit, $R_n$, increases as the cable resistance $R_L$ increases. This method is particularly effective and simple when the temperature-sensitive resistor can be in the same temperature environment as the line.

## VIII. CONCLUSIONS

Knowledge about the properties of NIB transmission lines, which are lines into which negative-impedance units have been inserted periodically, has been advanced considerably by this latest investigation. The distinctive properties of these lines stem largely from the way the sources of energy replacement, that is, the negative-impedance units, are coupled into the line. The cost and bulk of coil or resistance hybrids, necessary when unilateral amplifiers are used, are avoided and bilateral transmission is provided directly. The reflections caused by the insertion of these units into the line cancel each other up to a certain cutoff frequency when the units are uniformly spaced and the line is terminated approximately in its image impedance. There is an approximately inverse relationship between this cutoff frequency and the spacing of the units.

Two separate field test lines, each about 32 miles long, were constructed and operated for some time. One of these was in a suburban exchange area and used partly aerial and partly underground cable. The other used order wires in the L4 spur cable (underground) between Netcong and Newark, New Jersey. Both of these demonstrated, in agreement with analysis, that these NIB lines can be operated stably, with a margin, at a net loss of about 2 dB and that they can provide

high-quality transmission with low-phase and nonlinear distortion for speech or pulse signals. Because the NIB units are effective down to dc, baseband transmission of pulse signals is possible and was demonstrated on the test lines.

Two problems have held back the use of these lines. The first has to do with the effect of temperature variations of the transmission lines. A simple method for the correction of these variations which has been effective in the test lines is not universally applicable. It requires the temperature of the booster to be simply related to that of the line. The second is the susceptibility of the line to certain kinds of interfering currents and is more fundamental, as no satisfactory way to overcome it has been found. In well-balanced conventional transmission lines, the effect of induced longitudinal interfering currents is negligible. In NIB lines, these currents must flow through the NIB units and if the interfering currents are large compared with the signal currents, they may cause the total dynamic current to go outside the region of negative resistance and, hence, cause large distortion. This effect limits the use of NIB lines to environments where interfering currents are small. The interfering currents were negligible in the two test lines, and there are probably many such situations. Nevertheless, it is apparent that the use of the negative-impedance method to reduce attenuation is worthy of further consideration in certain situations, as for example in those which are similar to the Netcong-Newark field test, even though because of the above difficulties the method may not be suitable for use in all situations.

## IX. ACKNOWLEDGMENT

The writer is glad to acknowledge the encouragement given by J. A. Young and B. G. King, and the important contributions of A. L. Hopper and D. M. Chapin in the work of this project begun by L. A. Meacham. He is grateful for the suggestions and criticisms of K. Bullington who kindly read through the paper.

## REFERENCES

1. T. Shaw, "The Conquest of Distance by Wire Telephony; A Story of Transmission Development From the Early Days of Loading to the Wide Use of Thermionic Repeaters," B.S.T.J., *23*, No. 4 (October 1944), pp. 337–421.
2. T. Shaw, "Evolution of Inductive Loading for Bell System Telephone Facilities," B.S.T.J., *30*, Nos. 1, 2, 3, and 4 (January, April, July, and October 1951).
3. G. A. Campbell, "On Loaded Lines in Telephonic Transmission," Phil. Mag., Sixth Series, *5* (January-June 1903), pp. 313–330. Also in *Collected Papers of George A. Campbell*, a commemorative volume published by AT&T in 1937.
4. B. Gherardi and F. B. Jewett, "Telephone Repeaters," Trans. AIEE, *38* (1919), pp. 1287–1345.

5. A. B. Clark, "Telephone Transmission Over Long Cable Circuits," B.S.T.J., *2*, No. 1 (January 1923), pp. 67–94.
6. A. B. Clark, "The Development of Telephony in the United States," Trans. AIEE, Part I, Com. and Elec., *71* (November 1952), pp. 348–363.
7. F. B. Jewett, "Dr. George A. Campbell," and "Dr. Campbell's Memoranda of 1907 and 1912," B.S.T.J., *14*, No. 4 (October 1935), pp. 553–572.
8. G. Crisson, "The Limitation of the Gain of Two-Way Telephone Repeaters by Impedance Irregularities," B.S.T.J., *4*, No. 1 (January 1925), pp. 15–25.
9. G. Crisson, "Irregularities in Loaded Telephone Circuits," B.S.T.J., *4*, No. 8 (October 1925), pp. 561–585.
10. L. G. Abraham, "Certain Factors Limiting the Volume Efficiency of Repeatered Telephone Circuits," B.S.T.J., *12*, No. 4 (October 1933), pp. 517–532.
11. L. G. Abraham, "Circulating Currents and Singing on Two-Wire Cable Circuits," B.S.T.J., *14*, No. 4 (October 1935), pp. 600–631.
12. G. Crisson, "Negative Impedances and the Twin 21-Type Repeater," B.S.T.J., *10*, No. 3 (July 1931), pp. 485–513.
13. E. W. Herold, "Negative Resistance and Devices for Obtaining It," Proc. IRE, *23*, No. 10 (October 1935), pp. 1201–1223.
14. A. W. Hull, "Description of the Dynatron," Proc. IRE, *6*, No. 1 (February 1918), pp. 5–35.
15. K. D. Smith, "The IMPATT Diode—A Solid-State Microwave Generator," Bell Laboratories Record, *45*, No. 5 (May 1967), pp. 144–148.
16. R. S. Englebrecht, "Bulk Effect Devices for Future Transmission Systems," Bell Laboratories Record, *45*, No. 6 (June 1967), pp. 192–198.
17. J. G. Linvill, "Transistor Negative Impedance Converters," Proc. IRE, *41*, No. 6 (June 1953), pp. 725–739.
18. L. A. Meacham, "Negative Impedance Boosting," B.S.T.J., *47*, No. 6 (July-August 1968), pp. 1019–1041.
19. J. L. Merrill, Jr., "Theory of the Negative Impedance Converter," B.S.T.J., *30*, No. 1 (January 1951), pp. 88–109.
20. J. L. Merrill, Jr., A. F. Rose, and J. O. Smethurst, "Negative Impedance Telephone Repeaters," B.S.T.J., *33*, No. 5 (September 1954), pp. 1055–1092. J. Gammie and J. L. Merrill, Jr., "Stability of Negative Impedance Elements in Short Transmission Lines," B.S.T.J., *34*, No. 2 (March 1955), pp. 333–360.
21. A. L. Bonner, J. L. Garrison, and W. J. Kopp, "The E6 Negative Impedance Repeater," B.S.T.J., *39*, No. 6 (November 1960), pp. 1445–1504.
22. A. L. Hopper, "Computer-Aided Analysis and Design of Negative Impedance Boosted Transmission Lines," IEEE Trans. on Comm. Technology, *19*, No. 4 (August 1971), pp. 501–516.
23. K. Bullington, "U. S. Patent 2,360,932, Negative Resistance Loading," April 25, 1942.
24. K. G. Van Wynen, unpublished notes, 1950.
25. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Chapter 8.
26. C. Gewertz, "Synthesis of Networks," J. Math. and Phys., M.I.T., January 1933.
27. F. B. Llewellyn, "Some Fundamental Properties of Transmission Systems," Proc. IRE, *40*, No. 3 (March 1952), pp. 271–283.
28. W. H. Ku, "A Simple Derivation for the Stability Criterion of Linear Active Two-Ports," Proc. IEEE, *53*, No. 3 (March 1965), pp. 310–311.

# Contributors to This Issue

**Syed V. Ahamed,** B.E., 1957, University of Mysore, India; M.E., 1958, Indian Institute of Science; Ph.D., 1962, University of Manchester, U. K.; Post Doctoral Research Fellow, 1963, University of Delaware; Assistant Professor, 1964, University of Colorado; Bell Laboratories, 1966—. Mr. Ahamed has worked in computer-aided engineering analysis and software design. He has applied algebraic analysis to the design of domain circuits and investigated computer aids to the design of bubble circuits. Since 1972 he has been investigating microwave devices.

**Chester W. Anderson III,** B.A. (Physics), 1964, University of Wichita; M.Sc. (Physics), 1966, Wichita State University; Ph.D. (Physics), 1970, University of Alberta, Edmonton, Canada; Captain, U. S. Army, Electronic Technology and Devices Laboratory, USECOM, Fort Monmouth, N. J., 1970–1972; Bell Laboratories, 1972—. Mr. Anderson is a member of the Fundamental Interference Studies Group of the Interference and Protection Laboratory at Bell Laboratories. Member, American Geophysical Union, American Association for the Advancement of Science, Institute of Navigation.

**David McE. Boulin,** B.S. (Physics) 1965, Bethany College; M.S. (Physics), 1971, Newark College of Engineering; Bell Laboratories 1966—. Mr. Boulin was initially involved in crystal growth studies. He is presently engaged in the fabrication and study of dual-dielectric charge-storage devices.

**J. E. Goell,** B.E.E., 1962, M.S., 1963, and Ph.D. (E.E.), 1965, Cornell University; Bell Laboratories, 1965–1974. While at Cornell, Mr. Goell was a teaching assistant and held the Sloan Fellowship and the National Science Cooperative Fellowship. At Bell Laboratories, he worked on solid-state repeaters for millimeter wave communication systems and optical integrated circuits, and repeaters for optical fiber communication systems. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, Phi Kappa Phi, IEEE.

**Robert M. Gray**, B.S. and M.S., 1966, Massachusetts Institute of Technology; Ph.D., 1969, University of Southern California. Mr. Gray is an Assistant Professor of Electrical Engineering at Stanford University engaged in research and teaching in random processes communication and information theory, and data compression. He is past chairman of the San Francisco Chapter of the IEEE Information Theory Group. Member, IEEE, SIAM, MAA, IMS, AAAS, Sigma Xi, Eta Kappa Nu.

**John C. Irvin**, B.A. (physics), 1949, Miami University, M.A., 1953, and Ph.D., 1957 (both physics), University of Colorado; Bell Laboratories, 1957—. Mr. Irvin was initially engaged in the investigation of bulk silicon and of diffused layers of silicon. He later became involved in the development of microwave diodes, especially gallium arsenide varactor, mixer, Gunn, and IMPATT diodes. After an interlude in surface-state physics, he returned to the microwave device field and is currently concerned with gallium arsenide IMPATT diodes and field-effect transistors. Member, IEEE, American Physical Society, Sigma Xi, Phi Beta Kappa.

**Dawon Kahng**, B.Sc., 1955, Seoul University; M.Sc., 1956, Ph.D., 1959, The Ohio State University; Instructor, The Ohio State University, 1959; Bell Laboratories, 1959—. Mr. Kahng has worked on studies of impurity diffusion into silicon, feasibility studies of MOS transistors, hot electron devices, and on silicon epitaxial film doping profile studies. Since 1964, he has supervised a group concerned with development of surface barrier diodes, large gap and ferroelectric semiconductors, and luminescence in thin-film devices, charge-coupled devices, and MIS charge-storage memory devices. Fellow, IEEE; Life Member, Korean Physical Society; Member, Sigma Xi, Pi Mu Epsilon.

**Louis J. Lanzerotti**, B.S. (Engineering Physics), 1960, University of Illinois; M.A., 1963, and Ph.D., 1965, Harvard University; Bell Laboratories, 1965—. In 1972, Mr. Lanzerotti served as Visiting Scientist in the Department of Physics, University of Calgary. At Bell Laboratories, he is a member of the Radiation Physics Department. Member, American Physical Society, American Geophysical Union, Society of Terrestrial Magnetism and Electricity of Japan.

**Joseph R. Ligenza,** B.S. (Chemistry), 1951, Illinois Institute of Technology; M.A., 1952, and Ph.D. (Physical Chemistry), 1954, Columbia University; Bell Laboratories, 1954—. Mr. Ligenza has been engaged in studies on aspects of silicon oxidation and chemical reactions in plasmas. Member, Sigma Xi, Phi Lambda Epsilon.

**Carol G. Maclennan,** A.B. (Mathematics), Pembroke College in Brown University, 1960; Bell Laboratories, 1960—. In 1963–1964, Ms. Maclennan was on the staff of the Cornell University Computation Center. Since 1964, she has been a member of the Radiation Physics Research Department at Bell Laboratories.

**Jack M. Manley,** B.S. (Electrical Engineering), 1930, University of Missouri; Bell Laboratories, 1930–1974. He was first concerned with theoretical and experimental studies of nonlinear electric circuits. The Manley-Rowe relations are a result of this work. A few years were spent in adapting nonlinear coils to generate very high voltage pulses for use in radar transmitters. He later worked with new multiplex methods for communication systems, including early research work on PCM. Afterward, he was engaged in transmission-line research, then in study of noise problems in digital transmission systems. During the past several years he has been studying the use of negative impedances in transmission lines. Upon retirement, he was appointed Visiting Professor in the Department of Electrical and Computer Engineering of the University of Wisconsin at Madison. Fellow, IEEE; member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

**Dietrich Marcuse,** Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–1957; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research and studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah. He is presently working on the transmission aspect of a light communications system. Mr. Marcuse is the author of three books. Fellow, IEEE; member, Optical Society of America.

**Charles T. Neppell,** B.A. (Physics), 1965, Queens College; Bell Laboratories, 1968—. Mr. Neppell's early work at Bell Laboratories was in semiconductor device physics and included spectroscopic investigations of semiconductor properties. As a member of the Unipolar Integrated Circuit Laboratory, he is presently involved in designing and testing semiconductor memory circuits.

**William J. Sundburg,** B.S.E.E., 1968, Newark College of Engineering; M.S.E.E., 1971, Rutgers; Bell Laboratories, 1959—. Mr. Sundburg initially worked in the Surface Physics Research Department on semiconductor surface phenomena. In 1965 he transferred to the Physical Chemistry Research Department to study oxide film growth on single crystal metals in ultrahigh vacuum. From 1969 to present, Mr. Sundburg has been working in the Unipolar Integrated Circuit Laboratory designing and testing integrated circuits.

**K. K. Thornber,** B.S., 1963, M.S. (E.E.), 1964, Ph.D. (E.E.), 1966, California Institute of Technology; Research Associate, Stanford Electronics Laboratories, 1966–68; Research Assistant, Physics Department, University of Bristol, 1968–69; Bell Laboratories, 1969—. Mr. Thornber is a member of the Unipolar Integrated Circuit Laboratory.

**Aaron D. Wyner,** B.S. 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, Columbia University; Bell Laboratories, 1963—. Mr. Wyner has been doing research in various aspects of information and communication theory and related mathematical problems. He spent the year 1969–1970 visiting the Department of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, and the Faculty of Electrical Engineering, the Technion, Haifa, Israel on a Guggenheim Foundation Fellowship. He has also been a full- and part-time faculty member at Columbia University and the Polytechnic Institute of Brooklyn. He has been chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group, and has served as an associate editor of the Group's *Transactions* and as cochairman of two international symposia. He is presently second vice-president of the IEEE Information Theory Group. Member, IEEE, AAAS, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.